

Review Article

An update on chloroplast genomes

V. Ravi, J. P. Khurana, A. K. Tyagi, P. Khurana

Interdisciplinary Centre for Plant Genomics and Department of Plant Molecular Biology, University of Delhi South Campus, New Delhi, India

Received 9 July 2007; Accepted 4 September 2007; Published online 28 November 2007
© Springer-Verlag 2007

Abstract. Plant cells possess two more genomes besides the central nuclear genome: the mitochondrial genome and the chloroplast genome (or plastome). Compared to the gigantic nuclear genome, these organelle genomes are tiny and are present in high copy number. These genomes are less prone to recombination and, therefore, retain signatures of their age to a much better extent than their nuclear counterparts. Thus, they are valuable phylogenetic tools, giving useful information about the relative age and relatedness of the organisms possessing them. Unlike animal cells, mitochondrial genomes of plant cells are characterized by large size, extensive intramolecular recombination and low nucleotide substitution rates and are of limited phylogenetic utility. Chloroplast genomes, on the other hand, show resemblance to animal mitochondrial genomes in terms of phylogenetic utility and are more relevant and useful in case of plants. Conservation in gene order, content and lack of recombination make the plastome an attractive tool for plant phylogenetic studies. Their importance is reflected in the rapid increase in the availability of complete chloroplast genomes in the public databases. This review aims to summarize the progress in chloroplast genome research since its inception and tries to encompass all related aspects. Starting with a brief historical account, it gives

a detailed account of the current status of chloroplast genome sequencing and touches upon RNA editing, *ycfs*, molecular phylogeny, DNA barcoding as well as gene transfer to the nucleus.

Keywords: Chloroplast genome; History; Inverted repeat; AT/GC content; *ycf*; RNA editing; Phylogeny; DNA barcoding; Gene transfer to the nucleus

Introduction and brief history

The origin of chloroplasts dates back to more than a billion years (Brocks et al. 1999, Kostianovsky 2000, Embley and Martin 2006). The identification of *Bangiomorpha*, a fossil red alga aged 1.2 billion years, provides strong support to this point (Butterfield 2000). Schimper (1883) is regarded as the first to have considered that plastids arose from endosymbiotic photosynthetic bacteria. He noted that plastids divide by binary fission, independently of the surrounding cell, and showed a remarkable resemblance with free-living cyanobacteria. Mereschowsky (1905), however, was the one who elaborated about the

concept of a bacterial cell living within a nucleated host cell. He termed the plastids as “little green slaves” working for their host cells to produce food from sunshine. The origin of oxygenic photosynthesis and eukaryotic acquisition of plastids are regarded as landmark events in the history of life. After an “eclipse period” of around 50 years (Taylor 1987), the discovery of DNA inside the plastids gave a big boost to the endosymbiont theory. In 1951, Chiba, using cytological studies, suggested that chloroplasts from the moss fern *Selaginella* and two flowering plants contained DNA. Stocking and Gifford (1959) detected DNA within plastids of the green alga, *Spirogyra* and thus, in the late 1960s, the endosymbiont theory was proposed by Lynn Margulis. She spent much of the 1960s emphasizing that symbiosis was a major force in the evolution of cells. In 1970, she published her argument in *The Origin of Eukaryotic Cells*. Figure 1 (adapted from Archibald 2005) shows the widely accepted scheme for the origin and spread of plastids via endosymbiosis.

Chloroplasts are semi-autonomous organelles possessing their own genetic material – the chloroplast genome or plastome. The existence of a unique DNA species in plastids was verified more than forty years ago by Sager and Ishida in 1963. Since then, studies on plastid genome organization, gene expression and sequence-based phylogeny have come up in a big way. It is now indisputable that chloroplasts, as well as mitochondria, are of prokaryotic origin. Free-living, oxygen producing cyanobacteria were the progenitors of present-day chloroplasts and oxygen-consuming proteobacteria gave rise to mitochondria. It is also widely accepted that the endosymbiont genomes exported their genes to the nucleus through endosymbiotic gene transfer which resulted in genome size reduction and a compartmented, integrated eukaryotic genetic system (Brennicke et al. 1993, Martin and Herrmann 1998, Race et al. 1999, Timmis et al. 2004). Today, most of the proteins present in chloroplasts are encoded by the nuclear genome and then imported into the organelles (Jarvis and Soll 2001, Leister 2003). The present day chloroplast genomes code for roughly less than 200

proteins ranging from 26 in *Toxoplasma gondii* (an apicomplexan) to 209 in *Porphyra purpurea* (a red alga). The average is around 90 proteins for the green plants.

General structure of chloroplast DNA

Chloroplast DNAs (cpDNAs) of higher plants are double-stranded molecules of relatively small size, ranging from 35 to 217 kilobases (kb) with most of the photosynthetic organisms in the range of 115 to 165 kb (Table 1). Only a proportion of the total cpDNA is circular as opposed to earlier views (Bendich 2004). They are present in 1,000–10,000 copies per cell. The first plant genome to be completely sequenced was a chloroplast genome (Shinozaki et al. 1986) owing to the relatively small size. The cpDNA usually consists of two copies of inverted repeats (IR_A and IR_B), separated by a large and a small single-copy region (LSC and SSC, respectively). The genes found in cpDNAs are generally conserved with respect to content and order especially within a particular group of organisms. These can be divided into three broad categories: The first category comprises of genes for the photosynthetic apparatus. This category includes photosystem I (*psaA*, *psaB*, etc.), photosystem II (*psbA*, *psbB*, etc.), cytochrome b6f (*petA*, *petB*, etc.), ATP synthase (*atpA*, *atpB*, etc.), RuBisCo (*rbcL*) and NAD(P)H dehydrogenase genes (*ndhA*, *ndhB*, etc.). The second category comprises RNA genes and genes for the genetic apparatus. This includes transfer RNA (*trnH*, *trnK*, etc.), ribosomal RNA (*rrn16*, *rrn5*, etc.), RNA polymerase (*rpoA*, *rpoB*, etc.) and ribosomal subunit genes (*rps2*, *rps3*, *rpl2*, *rpl16*, etc.). The third category comprises of conserved ORFs called *ycfs* (see section: *Hypothetical chloroplast open reading frames*) and potential protein-coding genes like *matK* and *cemA*.

In 1986, the chloroplast genomes of tobacco and the liverwort *Marchantia polymorpha* were sequenced (Ohshima et al. 1986, Shinozaki et al. 1986), making these two the first complete plastome sequences. During the last two decades, plastome sequences have increased at a rapid rate and, currently, 82 complete plastid genome

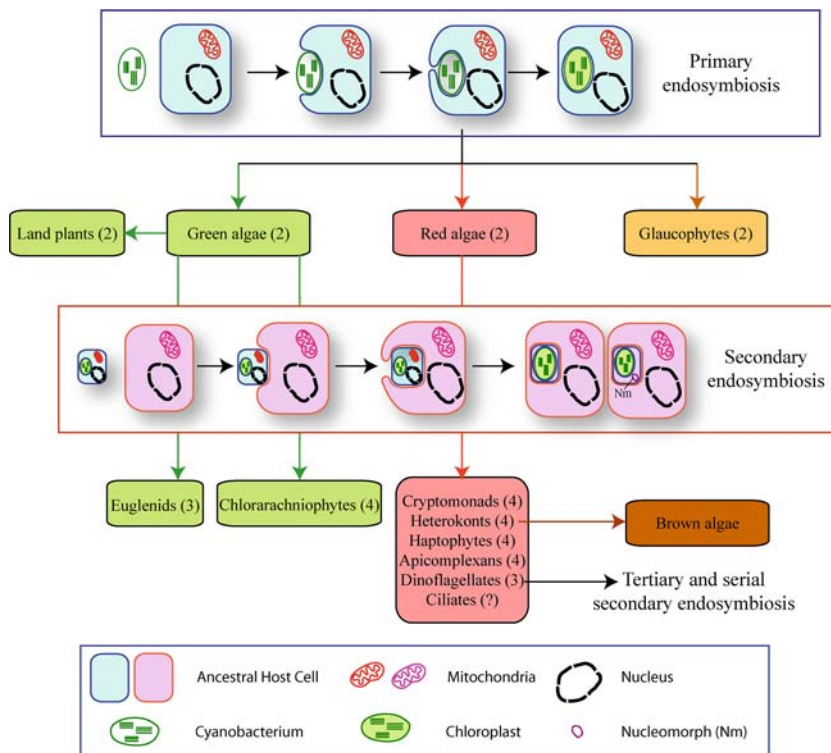


Fig. 1. Schematic representation of primary and secondary endosymbiotic events giving rise to different green and red-algal lineages (colour-coded, adapted from Archibald 2005). A single endosymbiotic event gave rise to three primary endosymbiotic groups – green algae, red algae and glaucophytes, the green algae ultimately giving rise to the present day land plants. Euglenids and chlorarachniophytes were derived from a green algal ancestor by separate secondary endosymbiosis events. On the other hand, the secondary red algal lineages were derived from a single event. Tertiary and serial secondary endosymbiosis has been observed in some species of dinoflagellates. Brown algae are believed to be derived from the heterokonts. Figures within parentheses represent the number of membranes surrounding the plastids in those lineages. The identity and nature of the plastid within ciliates is not known (?)

sequences are available in public databases (see Table 1 and <http://www.ncbi.nlm.nih.gov/genomes/ORGANELLES/plastids.html>; http://megsun.bch.umontreal.ca/ogmp/projects/other/cp_list.html). These include plastomes from various groups of organisms ranging from green and red algae to land plants (Fig. 2). Almost all sequences are available from the National Center for Biotechnology Information (NCBI) website, except for *Populus* which is available from the US Department of Energy-Joint Genome Initiative webpage (also see footnote below Table 1).

Out of the 82 taxa with sequenced chloroplast genomes, 67 (81.7%) belong to the kingdom Viridiplantae (green plants) which also includes green algae. *Mesostigma viride*, a green alga,

occupies a basal position with respect to divisions Streptophyta (57 taxa) and Chlorophyta (nine taxa) and is considered as one of the most primitive organisms with a completely sequenced plastome. It is now included under the division Streptophyta. The remaining 57 streptophytes comprise 53 land plants and 4 charophytes (the closest living relatives of land plants). The majority of the land plants are of the vascular type (50). The three non-vascular plants with complete chloroplast genomes are the liverwort *Marchantia polymorpha*, the hornwort *Anthoceros formosae*, and the moss *Physcomitrella patens*. Seed plants (spermatophytes) dominate the vascular plant category with 45 angiosperm and two gymnosperm chloroplast genome sequences.

Table 1. Alphabetical list of completely sequenced chloroplast genomes (as per NCBI, dated: 16.11.2006). Sizes of the plastome, large single copy (LSC) region, small single copy (SSC) region and inverted repeats (IR) were obtained from the respective references. AT/GC content was calculated using BioEdit. Italicized figures represent those, which were calculated manually by BLAST search of the plastome against itself (due to the non-availability of sizes in the respective reference or due to the unpublished nature of the work). Dashes represent cases where the typical quadripartite structure is not present

S. No.	Species	Size (bp)				Composition % AT/GC	Accession No.	Reference
		Total	LSC	IR	SSC			
1	<i>Acorus calamus</i>	153821	84149	25697	18278	61.40/38.60	NC_007407	Goremykin et al. 2005
2	<i>Adiantum capillus-veneris</i>	150568	82282	23447	21392	57.99/42.01	NC_004766	Wolf et al. 2003
3	<i>Amborella trichopoda</i>	162686	90970	26651	18414	61.66/38.34	NC_005086	Goremykin et al. 2003b
4	<i>Anthoceros formosae</i>	161162	107503	15744	22171	67.10/32.90	NC_004543	Kugita et al. 2003a
5	<i>Arabidopsis thaliana</i>	154478	84170	26264	17780	63.71/36.29	NC_000932	Sato et al. 1999
6	<i>Atropa belladonna</i>	156687	86868	25906	18008	62.44/37.56	NC_004561	Schmitz-Linneweber et al. 2002
7	<i>Bigeloviella natans</i>	69166	46282	9380	4124	69.69/30.17	NC_008408	Rogers et al. 2007
8	<i>Calycanthus floridus</i> var. <i>glaucus</i> (syn. <i>Calycanthus fertilis</i>)	153337	86948	23295	19799	60.73/39.27	NC_004993	Goremykin et al. 2003a
9	<i>Chaetosphaeridium globosum</i>	131183	88682	12431	17639	70.38/29.62	NC_004115	Turmel et al. 2002
10	<i>Chlamydomonas reinhardtii</i>	203828	80873	22211	78100	65.54/34.46	NC_005353	Maul et al. 2002
11	<i>Chlorella vulgaris</i>	150613	–	–	–	68.44/31.56	NC_001865	Wakasugi et al. 1997
12	<i>Citrus sinensis</i>	160129	87744	26996	18393	61.52/38.48	NC_008334	Bausher et al. 2006
13	<i>Coffea arabica</i>	155189	85166	25943	18137	62.57/37.43	NC_008535	Samson et al. 2007
14	<i>Cucumis sativus</i>	155527	86879	25187	18274	63.05/36.95	DQ119058	Kim et al. 2006
15	<i>Cyanidioschyzon merolae</i>	149987	–	–	–	62.37/37.63	NC_004799	Ohta et al. 2003
16	<i>Cyanidium caldarium</i>	164921	–	–	–	67.27/32.73	NC_001840	Glöckner et al. 2000
17	<i>Cyanophora paradoxa cyanelle</i>	135599	94946	11285	18083	69.53/30.47	NC_001675	Stirewalt et al. 1995 ^U
18	<i>Daucus carota</i>	155911	84242	27051	17567	62.34/37.66	NC_008325	Ruhlman et al. 2006
19	<i>Drimys granadensis</i>	160604	88685	26649	18621	61.21/38.79	NC_008456	Cai et al. 2006
20	<i>Eimeria tenella</i>	34750	23999	5361	29	79.36/20.63	NC_004823	Cai et al. 2003
21	<i>Emiliania huxleyi</i>	105309	84444	4841	11183	63.19/36.81	NC_007288	Sánchez Puerta et al. 2005
22	<i>Epifagus virginiana</i>	70028	19799	22735	4759	64.00/36.00	NC_001568	Wolfe et al. 1992
23	<i>Eucalyptus globulus</i>	160286	89012	26393	18488	63.14/36.85	AY780259	Steane 2005

Table 1. (Continued)

S. No.	Species	Size (bp)				Composition % AT/GC	Accession No.	Reference
		Total	LSC	IR	SSC			
24	<i>Euglena gracilis</i>	143171	–	–	–	73.87/26.13	NC_001603	Hallick et al. 1993
25	<i>Euglena longa</i> (syn. <i>Astasia longa</i>)	73345	–	–	–	77.59/22.41	NC_002652	Gockel and Hachtel 2000
26	<i>Glycine max</i>	152218	83175	25574	17895	64.63/35.37	DQ317523	Saski et al. 2005
27	<i>Gossypium hirsutum</i>	160301	88816	25608	20269	62.76/37.24	NC_007944	Lee et al. 2006
28	<i>Gracilaria tenuistipitata</i> var. <i>liui</i>	183883	–	–	–	70.85/29.15	NC_006137	Hagopian et al. 2004
29	<i>Guillardia theta</i>	121524	96308	4780/ 4904	15532	67.03/32.97	NC_000926	Douglas and Penny 1999
30	<i>Helianthus annuus</i>	151104	83530	24633	18308	62.38/37.62	NC_007977	Timme et al. 2007
31	<i>Helicosporidium</i> sp.	37454	–	–	–	73.08/26.92	NC_008100	de Koning and Keeling 2006
32	<i>Huperzia lucidula</i>	154373	104088	15314	19657	63.75/36.25	NC_006861	Wolf et al. 2005
33	<i>Jasminum nudiflorum</i>	165121	92877	29486	13272	62.02/37.98	NC_008407	Lee et al. 2006 ^U
34	<i>Lactuca sativa</i>	152765	84103	25033	18596	62.45/37.55	NC_007578	Timme et al. 2007
35	<i>Liriodendron tulipifera</i>	159886	88150	26386	18964	60.84/39.16	NC_008326	Cai et al. 2006
36	<i>Lotus corniculatus</i> var. <i>japonicus</i>	150519	81936	25156	18271	63.97/36.03	NC_002694	Kato et al. 2000
37	<i>Lycopersicon esculentum</i>	155461	85882	25608	18363	62.14/37.86	NC_007898	Kahlau et al. 2006
38	<i>Marchantia polymorpha</i>	121024	81095	10058	19813	71.19/28.81	NC_001319	Ohyama et al. 1986
39	<i>Medicago truncatula</i>	124033	–	–	–	66.03/33.97	NC_003119	Lin et al. 2003 ^U
40	<i>Mesostigma viride</i>	118360	83627	6057	22619	69.85/30.15	NC_002186	Lemieux et al. 2000
41	<i>Morus indica</i>	158484	87386	25678	19742	63.63/36.37	NC_008359	Ravi et al. 2006
42	<i>Nandina domestica</i>	156599	85473	26062	19002	61.68/38.32	NC_008336	Moore et al. 2006
43	<i>Nephroselmis olivacea</i>	200799	92126	46137	16399	57.86/42.14	NC_000927	Turmel et al. 1999
44	<i>Nicotiana sylvestris</i>	155941	86684	25342	18573	62.15/37.85	NC_007500	Yukawa et al. 2006
45	<i>Nicotiana tabacum</i>	155943	86686	25342	18573	62.15/37.85	NC_001879	Shinozaki et al. 1986; Yukawa et al. 2006
46	<i>Nicotiana tomentosiformis</i>	155745	86392	25429	18495	62.21/37.79	NC_007602	Yukawa et al. 2006
47	<i>Nymphaea alba</i>	159930	90014	25177	19562	60.85/39.15	NC_006050	Goremykin et al. 2004
48	<i>Odontella sinensis</i>	119704	65346	7725	38908	68.18/31.82	NC_001713	Kowallik et al. 1995
49	<i>Oenothera elata</i> subsp. <i>Hookeri</i>	163935	89393	27807	14436	60.89/39.11	NC_002693	Hupfer et al. 2000
50	<i>Oltmannsiellopsis viridis</i>	151933	33610	18510	81303	59.53/40.47	NC_008099	Pombert et al. 2006
51	<i>Oryza nivara</i>	134494	80544	20802	12346	60.99/39.01	NC_005973	Shahid Masood et al. 2004

Table 1. (Continued)

S. No.	Species	Size (bp)				Composition % AT/GC	Accession No.	Reference
		Total	LSC	IR	SSC			
52	<i>Oryza sativa (indica</i> cultivar-group)	134496	80553	20798	12347	61.00/39.00	NC_008155	Tang et al. 2004
53	<i>Oryza sativa</i> (<i>japonica</i> cultivar-group)	134525	80592	20799	12335	61.01/38.99	NC_001320	Hiratsuka et al. 1989
54	<i>Ostreococcus tauri</i>	71666	35684	6824/ 6825	22333	60.11/39.89	NC_008289	Robbens et al. 2007
55	<i>Panax ginseng</i>	156318	86106	26071	18070	61.89/38.11	NC_006290	Kim and Lee 2004
56	<i>Pelargonium x</i> <i>hortorum</i>	217942	59710	75741	6750	60.39/39.61	NC_008454	Chumley et al. 2006
57	<i>Phalaenopsis</i> <i>aphrodite</i>	148964	85957	25732	11543	63.35/36.65	NC_007499	Chang et al. 2006
58	<i>Physcomitrella patens</i> subsp. <i>formosana</i>	122890	85212	9589	18501	71.47/28.53	NC_005087	Sugiura et al. 2003
59	<i>Pinus koraiensis</i>	116866	63835	473/ 475	52083	61.20/38.80	NC_004677	Noh et al. 2003 ^U
60	<i>Pinus thunbergii</i>	119707	65696	495	53021	61.50/38.50	NC_001631	Wakasugi et al. 1994
61	<i>Piper cenocladum</i>	160624	87668	27039	18878	61.69/38.31	NC_008457	Cai et al. 2006
62	<i>Platanus occidentalis</i>	161791	92150	25066	19509	61.97/38.03	NC_008335	Moore et al. 2006
63	<i>Populus alba</i>	156505	84618	27660	16567	63.26/36.74	NC_008235	Okumura et al. 2006 ^U
64	<i>Populus trichocarpa</i>	157033	85129	27652	16600	63.32/36.68	**	DOE Joint Genome Initiative
65	<i>Porphyra purpurea</i>	191028	–	–	–	67.01/32.99	NC_000925	Reith and Munholland 1995
66	<i>Porphyra yezoensis</i>	191952	–	–	–	66.88/33.12	NC_007932	Kunimoto et al. 2006 ^U
67	<i>Pseudoclonium</i> <i>akinetum</i>	195867	140914	6039	42875	68.51/31.49	NC_008114	Pombert et al. 2005
68	<i>Psilotum nudum</i>	138829	84617	18954	16304	63.97/36.03	NC_003386	Wakasugi et al. 2002 ^U
69	<i>Saccharum hybrid</i> cultivar <i>SP-80-3280</i>	141182	83047	22796	12544	61.56/38.44	NC_005878	Calsa et al. 2004 ^U
70	<i>Saccharum officinarum</i>	141182	83048	22795	12544	61.56/38.44	NC_006084	Asano et al. 2004
71	<i>Scenedesmus obliquus</i>	161452	72440	12022	64968	73.11/26.89	NC_008101	de Cambiaire et al. 2006
72	<i>Solanum</i> <i>bulbocastanum</i>	155371	85814	25588	18381	62.12/37.88	NC_007943	Daniell et al. 2006
73	<i>Solanum tuberosum</i>	155298	85737	25594	18373	62.12/37.88	NC_008096	Gargano et al. 2005
74	<i>Spinacia oleracea</i>	150725	82719	25073	17860	63.18/36.82	NC_002202	Schmitz- Linneweber et al. 2001

Table 1. (Continued)

S. No.	Species	Size (bp)				Composition % AT/GC	Accession No.	Reference
		Total	LSC	IR	SSC			
75	<i>Staurostrum punctulatum</i>	157089	–	–	–	67.51/32.49	AY958085	Turmel et al. 2005
76	<i>Stigeoclonium helveticum</i>	223902	–	–	–	71.13/28.87	NC_008372	Belanger et al. 2006
77	<i>Theileria parva</i>	39579	–	–	–	80.51/19.48	NC_007758	Gardner et al. 2005
78	<i>Toxoplasma gondii</i>	34996	24363	5316/ 5317	–	78.57/21.43	NC_001799	Kissinger et al. 1999 ^U
79	<i>Triticum aestivum</i>	134545	80349	20703	12790	61.69/38.31	NC_002762	Ogihara et al. 2002
80	<i>Vitis vinifera</i>	160928	89147	26358	19065	62.60/37.40	NC_007957	Jansen et al. 2006
81	<i>Zea mays</i>	140384	82355	22748	12536	61.54/38.46	NC_001666	Maier et al. 1995
82	<i>Zygnema circumcarinatum</i>	165372	–	–	–	68.92/31.08	AY958086	Turmel et al. 2005

^U Unpublished; **Available at http://genome.ornl.gov/poplar_chloroplast

Two species of the gymnosperms, *Pinus* (*P. thunbergii* and *P. koraiensis*) have completely sequenced plastomes. The remaining seed plants comprise of a lycopod (*Huperzia*) and two ferns (*Adiantum* and *Psilotum*). The 45 angiosperms are further subdivided into monocots, eudicots, magnoliids (basal dicots) and basal angiosperms.

Eudicot plastomes are the highest in number (31), followed by monocots (eight), magnoliids (four) and two basal angiosperm plastomes (*Amborella* and *Nymphaea*). Figure 3 gives the hierarchical tree of all known green plant plastome sequences.

The circular plastome of most land plants and algae have a quadripartite structure com-

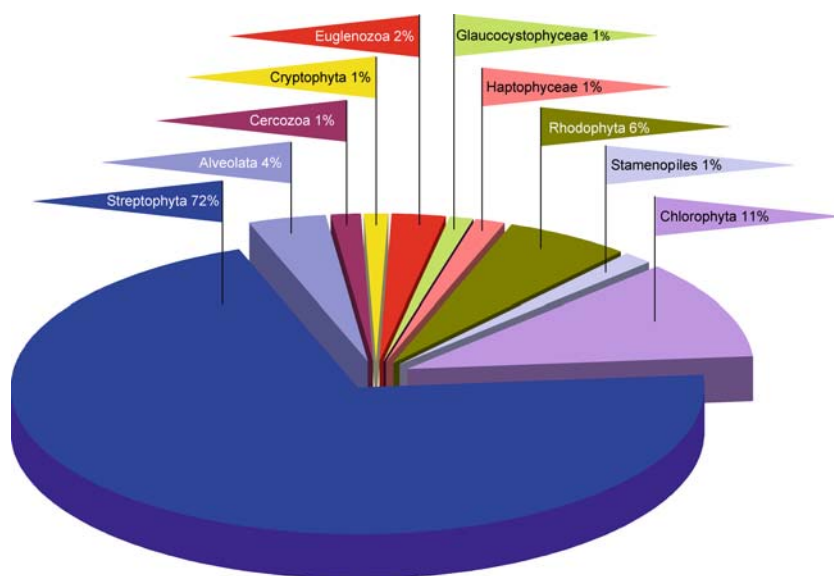


Fig. 2. Group-wise distribution of complete chloroplast genomes submitted in the GenBank (dated 16.11.2006). Classification is based on the NCBI scheme

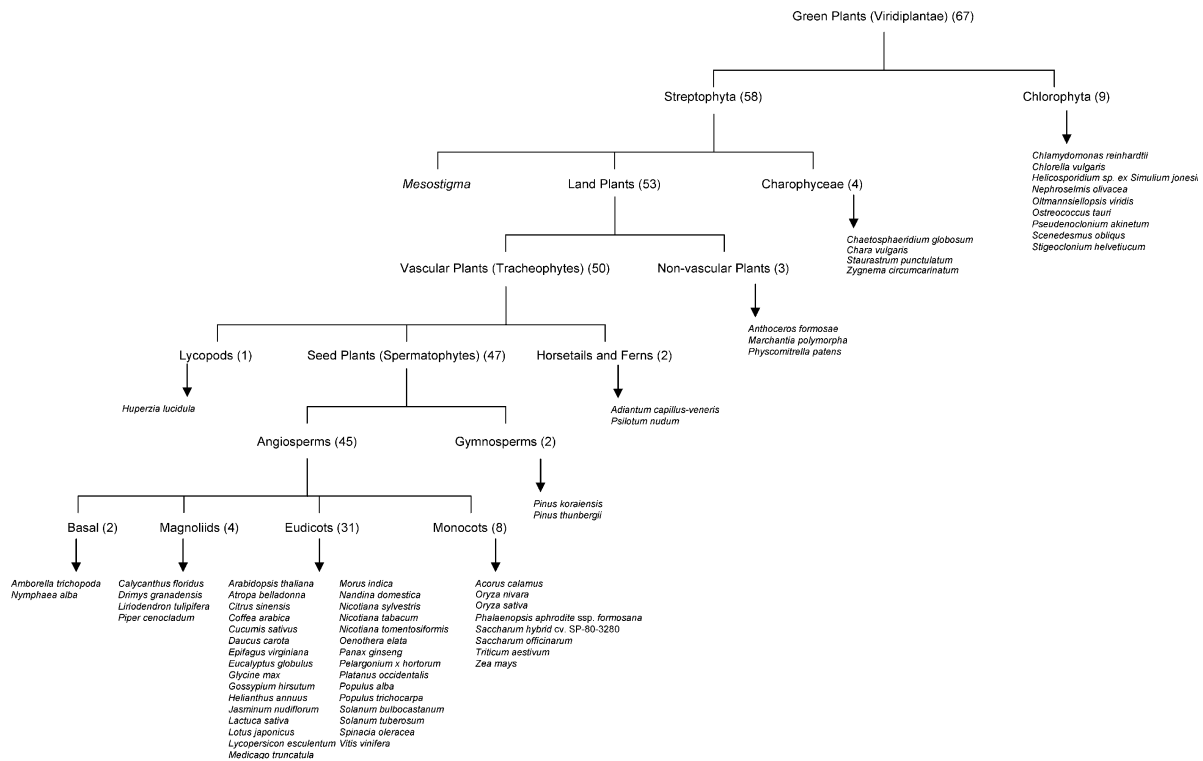


Fig. 3. Hierarchical arrangement of green plants (Viridiplantae) with completely sequenced chloroplast genomes

prising of two identical or non-identical inverted repeats (IRs) which separate two single copy (SC) regions, the large and small single copy regions, LSC and SSC, respectively. Size variation in the inverted repeats is the main cause for plastome size variation between different genera. IR size averages around 20–30 kb, although exceptions are known. Genes located in the IRs are, as expected, present as duplicates. The gymnosperm *Pinus thunbergii* has an extremely reduced pair of inverted repeats sizing a mere 495 bp – the smallest known among all plastomes – and in turn a small plastome of ~119 kb. This genus has lost all its genes coding for NADH-dehydrogenases (*ndh*). On the other extreme is *Pelargonium hortorum* which possesses massive IRs of ~76 kb giving its plastome a size of ~217 kb (Palmer et al. 1987, Chumley et al. 2006). Other exceptional plastomes, in terms of size are *Cyanidium caldarium* (~47 kb), *Euglena (Astasia) longa* (~73 kb), *Epifagus virginiana* (~70 kb), *Toxoplasma*

gondii (~35 kb), *Nephroselmis olivacea* (~200 kb), *Porphyra purpurea* (~191 kb) and *Chlamydomonas reinhardtii* (~203 kb).

IR-size variation is not the only cause for differences in the sizes of plastomes. Loss of genes, as in case of *Pinus* is another key reason. In *Epifagus*, a non-photosynthetic parasitic plant, all genes related to photosynthesis have been lost during evolution resulting in the reduced plastome size. In some species like *Medicago truncatula* one of the IRs has been lost and this has caused significant plastome size reduction. The lack of IR is thought to directly affect the nucleotide substitution rate in genes present in the remaining copy. For example, in legumes, Perry and Wolfe (2002) showed that the synonymous substitution rate in the IRs was 2.3 fold lower as compared to the SC region. In species lacking the IR (in other words having only one copy of the IR-genes), this rate was comparable to the SC regions, i.e. the substitution rate has gone up. IR presence or absence is thus a major

phylogenetic marker and within legumes this loss is common to *Medicago*, *Pisum* and *Vicia* but not to *Lotus* and *Glycine* (members of subfamily Papilionoideae). Differences in the intergenic regions are also universal players in plastome size variation.

The inverted repeats are thought to act as stabilizing regions and evolve two to three times more slowly as compared to the single copy regions. These are the most conserved regions of all plastomes and genes present in the IRs also show the same feature. Plastomes are able to retain signatures of evolutionary history much better than their nuclear counterparts due to lack of recombination which is, in turn, due to the presence of these IRs. The large single copy region is the least conserved when all parts of the plastome are compared. Although the organization and structure of plastomes is generally stable and conserved, it does not mean that rearrangements have not occurred. There have been group-specific inversions, translocations, insertions/deletions. A typical example is the inversion of a ~50 kb segment in the large single copy region of legume plastomes. In *Oenothera elata*, this inversion is in a region between the *accD* and *rps16* genes and sizes 54 kb. In *Lotus japonicus*, there is a 51 kb inversion between *rbcL* and *rps16*. Similarly, three inversions are distributed in different families of monocots – a 28 kb inversion common to families Restionaceae, Joinvilleaceae and Poaceae; a ~6 kb inversion common to only Joinvilleaceae and Poaceae; and the third inversion, which is the smallest (*trnT*), specific to grasses (Doyle et al. 1992).

Comparison of inverted repeat/single copy boundaries

Inverted repeats of many genera of higher plants are known to extend into neighbouring single copy regions. This causes differences at or near the border regions of the IR/SC. Large expansions and reductions in plastome sizes are attributed to this phenomenon. Genes lying next to the border regions in dicots are *tRNA-His* (IR_A/LSC), *rps19* (IR_B/LSC), *ycf1* (IR_A/SSC) and *ndhF* (IR_B/SSC). In case of grasses, these genes are *psbA* (IR_A/

LSC), *rpl22* (IR_B/LSC), *ndhH* (IR_A/SSC) and *ndhF* (IR_B/SSC). Expansion of the inverted repeats can lead to a situation where the IR/SC border lies within the coding region of a nearby gene. This is the case in vascular plants and is considered a common evolutionary event. Presence of the border within the coding region leads to the formation of a truncated counterpart at the other IR/SC border and hence a pseudogene. Thus, the presence of the IR_A/SSC border within the *ycf1* gene leads to the formation of a *ycf1* pseudogene (ψ *ycf1*) at the IR_B/SSC border. The lengths of these pseudogenes vary depending upon the extent to which the IR has extended into the SC region. Pseudogenes of *ycf1* sizing 996 bp (*Nicotiana*), 1438 bp (*Atropa*), 1030 bp (*Arabidopsis*), 266 bp (*Calycanthus*), 1100 bp (*Eucalyptus*), 1649 bp (*Panax*), 1444 bp (*Spinacia*) and 1001 bp (*Morus*) are known. A similar case is seen with *rps19* present at or near the IR_B/LSC border. The *rps19* pseudogenes are created at the IR_A/LSC border, e.g. *Atropa* (59 bp), *Arabidopsis* (113 bp), *Panax* (51 bp) and *Spinacia* (143 bp). Others have the *rps19* completely within their LSCs but at varying distances from the IR_B/LSC border; thus, a pseudogene is not created in these cases. In *Nicotiana*, the gene is located 4 bases upstream of the IR_B/LSC border. *Eucalyptus* has a 6 bp region in between, while *Morus* has its *rps19* gene situated sharply at the boundary. *Calycanthus* on the other extreme has 1552 bases between the border and its *rps19* gene. There is much less variation in case of the IR_A/LSC border region. The *tRNA-His* gene lies in the LSC region at varying distances from the border – *Nicotiana* (5 bp), *Atropa* (4 bp), *Arabidopsis* (3 bp), *Calycanthus* (0 bp), *Eucalyptus* (5 bp), *Panax* (5 bp), *Spinacia* (0 bp) and *Morus* (23 bp).

Similar to the *rps19* gene, the *ndhF* gene located at the IR_B/SSC border is also subjected to these variations. In case of *Nicotiana*, *Atropa*, *Calycanthus*, *Eucalyptus* and *Panax*, the gene lies completely within the SSC at distances 43 bp, 43 bp, 9 bp, 218 bp and 6 bp, respectively, from the border. In *Arabidopsis*, *Spinacia* and *Morus*, the border passes through the gene and thus there is an overlap with the neighbouring *ycf1* pseudogene (37 bp, 23 bp and 25 bp overlap, respec-

tively). The expansions/contractions of IR are probably mediated by intra-molecular recombination between two short direct repeat sequences that frequently occur within the genes located at the borders. Goulding et al. (1996) thus proposed two distinct mechanisms for IR junction evolution – (a) gene conversion for the small stretches and (b) recombinational repair of double strand breaks for incorporation of large chunks of single copy regions within the IR. The latter one operates very rarely while the former one is a continuous and random process maintaining the IR structure as a whole.

AT/GC content

Plastid genomes are generally highly AT-rich – the average being around 63%. A comparison between different plastomes in terms of total size, inverted repeats, single copy regions and AT/GC contents is presented in Table 1. The plastomes of the moss *Physcomitrella patens*, red alga *Gracilaria tenuistipitata* and the hornwort *Anthoceros formosae* have an AT-content toward the higher side (71.5%, 70.9% and 67.1%, respectively) while the green alga *Nephroselmis olivacea* has a low AT-content of 57.9%. AT-content of the single copy regions is generally much higher than the inverted repeat regions. The low AT-content of the IRs is attributed to the presence of ribosomal RNA genes which have the lowest AT-content. The strong AT-bias is also reflected in the codon usage. Roughly 70% of the codons of all plastomes end with the bases A or T. *Medicago*, which has lost one of its inverted repeats, has an AT-content of 66.03% which is comparable to other plastomes.

RNA editing

Another feature not uncommon among plastomes is the conversion of bases at the transcript level which can lead to a functional protein or a protein with a changed amino acid (Tillich et al. 2006). In many species, as in case of the hornwort *Anthoceros*, the fern *Adiantum* or the monocot *Saccharum*, there are stop codons present in the coding region of the plastomes, but it does not

lead to a truncated transcript or protein, instead RNA editing affects a single base change in the mRNA that alters the codon, changing it from a nonsense codon to a sense or from one sense codon to another, which might lead to change in the amino acid. The usual alterations are C to U ‘transitions’ (pyrimidine to pyrimidine) in the transcript molecule (Tillich et al. 2006). Reverse editing (U to C) is rare and was observed only in case of *Anthoceros*. This phenomenon was seen in case of two genes *rbcL* and *atpB* in the hornwort. Two stop codons present in the *rbcL* had no effect on the protein, and when analyzed it was seen that the two stop codons, UGA and UAA, had been converted to CGA (Arg) and CAA (Gln) in the transcript, respectively, thus restoring the reading frame for the large subunit of *rbcL* (Yoshinaga et al. 1996). In case of the *atpB* gene, an unusual initiation codon ACG along with three stop codons were converted to AUG and sense codons, respectively (Yoshinaga et al. 1997). RNA editing also has a role in speciation (Schmitz-Linneweber et al. 2002). *Atropa belladonna* and *Nicotiana tabacum* belong to the same family Solanaceae and have high similarity in their plastomes in regulatory regions and genes including introns. However, differences are present only in functionally insignificant regions but importantly in the editing sites. It can thus be hypothesized that this difference leads to reproductive isolation and in turn speciation, forming the two species. In tobacco, 69 potential editing sites were identified out of which 31 sites were actually edited (Hirose et al. 1999). All 31 were C to U transitions and out of these 29 resulted in a change in amino acid, 2 created start codons and one site was at the third position of a codon in *atpA* mRNA and did not result in amino acid change. Large scale editing has also been observed in the fern *Adiantum capillus-veneris* (Wolf et al. 2004) where 349 RNA editing sites were detected by comparison of cDNA sequence to the genomic sequence. It was seen that the level of RNA editing in this fern was more than ten times that of any other chloroplast genome examined across vascular plants. This number is even higher in case of the hornwort *Anthoceros* (942 sites). Of

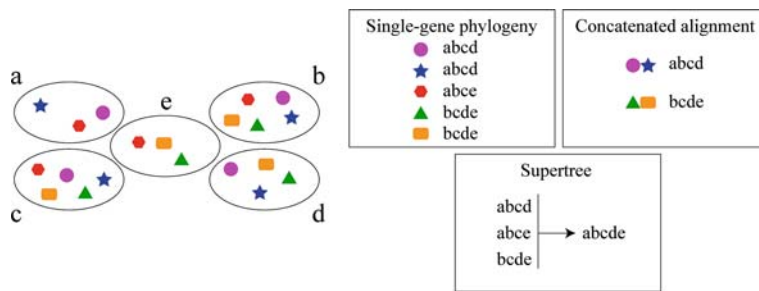


Fig. 4. Approaches for phylogenetic reconstruction. Taxa a–e contain some common and unique genes (coloured symbols) which can be used in different combinations to deduce relationships between a particular subset of the taxa. These individual trees can then be superimposed to form supertrees to obtain the relationship between all the taxa

these 509 were C to U transitions and 433 were U to C transitions (Kugita et al. 2003b). A total of 53 editing sites in *Adiantum* were found to be homologous to sites in *Anthoceros* and some land plants suggesting the possibility of conservation of a major component of RNA editing sites in these groups of organisms. Recent data supports the monophyletic origin of the RNA editing systems of seed plants, hornworts and ferns (Tillich et al. 2006).

Hypothetical chloroplast open reading frames (ycfs)

Plastid genomes of higher plants possess genes falling into three broad categories (Shimada and Sugiura 1991, Sugiura 1992). The first contains genetic system genes encoding for rRNAs, tRNAs, ribosomal proteins and RNA polymerase subunits. The second category contains genes for photosynthesis encoding subunits of the two photosystems, the cytochrome b6f complex, and the ATP synthase. In addition to these two, there are a number of open reading frames of unknown function and this constitutes the third category. These conserved *orfs* have been named *ycfs* (Hallick and Bairoch 1994, Maier et al. 1995). There has been a tremendous increase in the number of *ycfs* with accumulation of chloroplast and cyanobacterial genome sequence data and more than 80 are known presently (Stoebe et al. 1998, Glöckner et al. 2000). Some are specific to chloroplast genomes, some common to chloro-

plast and cyanobacteria, and some are found only in cyanobacteria and algal chloroplasts. The persistence of these sequences and their conservation within plastomes indicates their functional importance. Several *ycfs* have been worked upon in an effort to decipher their function and a few of them have been characterized in detail. In 1994, Monod and co-workers demonstrated that open reading frame 8 (*ycf8*) is expressed as part of the *psbB-psbH* operon and its product is associated specifically with the photosystem II complex. Xie and Merchant (1996) tested the function of *ycf5* which shows limited similarity to bacterial genes *ccl1/cycK* required for biogenesis of *c*-type cytochromes. The gene was renamed *ccsA* (*c*-type cytochrome synthesis). Takahashi et al. (1996) found that ORF43 (*ycf7*) was co-transcribed with the *psaC* gene and ORF58 in *Chlamydomonas*. They proposed the new name *petL* for *ycf7* after concluding that its gene product was an authentic subunit of the cytochrome *b6f* complex and was required for its stability, accumulation and efficiency. In 1997, Boudreau et al. characterized *ycf3* and *ycf4* from *Chlamydomonas*. Transformants lacking *Ycf3* and *Ycf4* showed lack of stable accumulation of photosystem I complex within the thylakoid membranes. In the same year, Rolland et al. (1997) concentrated their efforts on *ycf10* and localized it to the inner chloroplast envelope membrane. Their work suggested that there is a *ycf10*-dependent system promoting efficient inorganic carbon (Ci) uptake into chloroplasts. This was later named as *cemA*.

In 1998, Stoebe et al. published a tabulated listing of all protein-coding genes identified by BLAST searches from twelve sequenced chloroplast genomes and one cyanobacterial genome. Criterion for tabulation was presence of the gene in at least two chloroplast genomes or one chloroplast genome and the genome of *Synechocystis* sp. PCC6083. They updated the list of hypothetical chloroplast open reading frames and assigned tentative gene names based on sufficient similarity to prokaryotic sequences of known function. Hager et al. (1999) proposed to rename *ycf6*, the smallest conserved open reading frame in the plastid genome, to *petN*. Tobacco *ycf6* knock-out lines showed interruption in the electron transfer from photosystem II to photosystem I due to the complete absence of cytochrome *b₆f* complex, thus suggesting that Ycf6 was a genuine subunit of the complex playing an important role in its assembly and/or stability. Drescher et al. (2000) constructed several mutant alleles for targeted disruption and/or deletion of two giant *ycfs*, *ycf1* and *ycf2*. Although chloroplast transformants were obtained for all constructs, homoplasmic state could not be achieved even after repeated regeneration cycles. This indicated the indispensability of products encoded by *ycf1* and *ycf2* for cell survival. Almost simultaneously, Mäenpää et al. (2000) working on *ycf9* (*orf 62*) reached a similar conclusion when they could not obtain homoplasmicity in their transplastomic *Nicotiana* plants. They inferred that the *ycf9* gene product was essential for chloroplast function. This was later renamed to *psbZ* (Swiatek et al. 2001). Thus, it is clear that *ycfs* have been preserved and conserved because of their functions.

Out of the 80 odd *ycfs* known, more than 50% are exclusive to lower organisms like *Porphyra purpurea* (red alga), cyanelle *Cyanophora paradoxa*, the chlorophyll a + c containing alga *Odontella sinensis* and the cyanobacteria *Synechocystis* sp. PCC6803. The open reading frames (*ycfs*) 17, 21, 23, 27, 29, 34, 36–38 are exclusive to *Porphyra*, *Cyanophora* and *Synechocystis*; *ycfs* 32, 33, 35, 39 to *Odontella*, *Porphyra*, *Cyanophora* and *Synechocystis*; *ycfs* 40–47 are exclusive to *Porphyra*, *Odontella* and

Synechocystis; *ycfs* 48–51 to *Cyanophora* and *Synechocystis*; *ycfs* 19, 20, 22, 52–65 to *Porphyra* and *Synechocystis*. The least number of such open reading frames are in case of the apicomplexan *Plasmodium* and non-photosynthetic *Epi-fagus* owing to their parasitic nature.

Phylogeny

In molecular phylogeny as addressed with chloroplast genes, there is a trade-off between the number of taxa that can be conveniently sequenced and the number of genes (the amount of information) that can be conveniently obtained for each species to be sampled. There have been recent debates regarding the importance of number of taxa as compared to the size of the dataset and vice versa (Sanderson and Driskell 2003, Soltis et al. 2004, Stefanovic et al. 2004, Lockhart and Penny 2005, Martin et al. 2005).

A typical example of the ‘single (few) gene(s)/many taxa’ approach is the use of plastid gene *rbcL*. This gene has been sequenced from over 5000 species. In one of the largest phylogenetic analysis, *rbcL* gene sequences from 2538 species were used for parsimony jackknife analyses (Källersjö et al. 1998) in organisms ranging from cyanobacteria to flowering plants. On the other extreme are examples where concatenated data sets spanning few taxa have been used (‘many-genes/few taxa’) (Fig. 4). Among the first genome-scale phylogenetic studies were those by Goremykin et al. (1997) and Martin et al. (1998). One of the largest alignments till date was created in a study by Goremykin et al. (2005) who sequenced the plastome of *Acorus calamus* (an ancient monocot). The alignment of 89,436 bp comprising 15 taxa contains almost all sequences usable for phylogeny within the spermatophytes. The alignment contains sequences from 61 protein-coding genes, RNA genes (34 gene species), all introns and several conserved spacer sequences. The recent shift towards this trend is quite evident in several complete plastome sequencing articles (Turmel et al. 1999; Lemieux et al. 2000; Goremykin et al. 2003a, b; Kugita et al. 2003a; Goremykin et al. 2004;

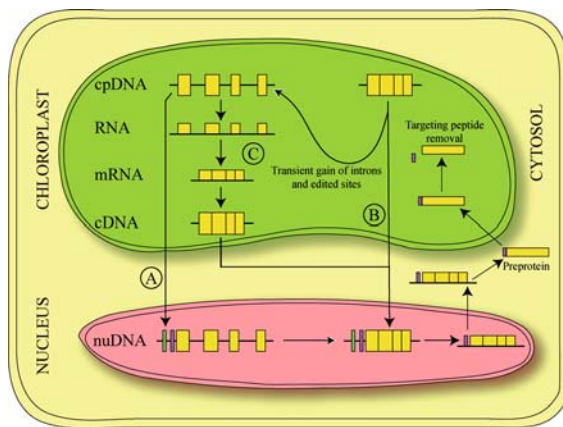


Fig. 5. Possible modes of transfer of chloroplast genes to the nucleus. **A** Direct transfer of chloroplast gene after origin of introns and/or editing, followed by gain of suitable regulatory elements and removal of introns in the nucleus. **B** Direct transfer of chloroplast gene prior to origin of introns and editing with or without transient intermediates. **C** Reverse transcription followed by integrative recombination of cDNA into the nuclear genome. Irrespective of the mode of transfer, the main requirement is integration and acquisition of suitable regulatory elements along with a transit peptide which will lead to the transfer of the protein back to the chloroplast

Hagopian et al. 2004; Goremykin et al. 2005; Wolf et al. 2005; Bausher et al. 2006; Jansen et al. 2006; Lee et al. 2006; Ravi et al. 2006;

Ruhlman et al. 2006). Although the ‘many genes/few taxa’ approach has taken over the phylogenetic centre stage, it still has the inherent problem of ‘few taxa’ because of the uneven distribution of sequences in the databases. To overcome this, an approach called the ‘supertree construction’ has been adopted which helps in combining smaller ‘subtrees’ from few organisms into a ‘consensus tree’ retaining all the organisms from the subtrees (Fig. 4).

Chloroplast genes and genomes have been major players in resolving portions of plant tree of life. Various genes and intergenic spacers have been identified as evolutionarily significant markers which have been widely used for phylogenetic analyses. These include the genes *rbcL*, *ndhF*, *matK*, *atpB*, *rpl16* and non-coding regions *trnL* intron and *trnL-trnF* intergenic region. Typical examples of the use of chloroplast genomes for resolving recalcitrant regions of the plant tree of life include the work by Jansen et al. (2006) who used the grape chloroplast genome sequence to establish its sister relationship with the remaining rosids. Another recent work used the mulberry plastome sequence to resolve the relationship between the orders of the nitrogen-fixing clade of eurosids I and established a ‘Rosales sister to Fabales’ topology (Ravi et al. 2007).

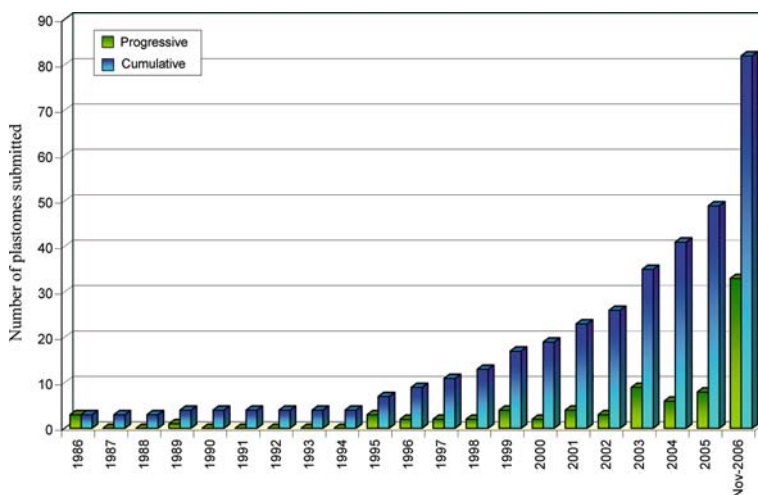


Fig. 6. Rate of submission of complete chloroplast genomes to the GenBank (1986-2006) as per the submission dates provided in the NCBI Plastid Genomes website (<http://www.ncbi.nlm.nih.gov/genomes/ORGANELLES/plastids.html>)

DNA Barcoding

Recently, genes and spacers typically used for resolving phylogenetic issues have gained recognition as “internal species tags” (Hebert and Gregory 2005) which would prove to be the building blocks of a new field termed as DNA Barcoding. The long-term goal of DNA Barcoding is to provide a rapid, accurate and automatable method for species identification using a standardized DNA region acting as a unique species-specific ‘tag’ (Hebert et al. 2003, Chase et al. 2005, Hebert and Gregory 2005). The mitochondrial *cox1* gene encoding subunit 1 of cytochrome oxidase has been used for barcoding of diverse groups of animals and has also proved useful in some groups of algae (Saunders 2005). This gene, however, is highly invariant and therefore unsuitable for barcoding in land plants. In an effort to identify regions that could serve as barcoding candidates in plants, Kress et al. (2005) compared the plastid genomes of *Nicotiana* and *Atropa* (members of Solanaceae). Regions with raw sequence differences $\geq 2\%$ were considered the most promising regions for barcoding. The nuclear internal transcribed spacer (ITS) and plastid *rbcL* gene were taken as baseline controls for these test regions. It was found that intergenic spacer regions were the best barcoding candidates. The intergenic region between genes *trnH* and *psbA* (*trnH-psbA*) was the most preferred region in terms of per cent sequence divergence and amplification success. In a similar study, Taberlet et al. (2007) evaluated the utility of *trnL* (UAA) intron for barcoding purposes. They checked both the full intron and a portion of the intron (P6 loop). Although the resolution of both these regions is relatively low, they had several advantages which include the ease of amplification owing to the high degree of conservation of the primers and the robustness of the amplification system which can aid in amplification from highly degraded samples. Newmaster et al. (2006) proposed a tiered approach in which a first tier region would provide resolution at a higher level (e.g. family or genus) and another region (preferably more variable) would be used for lower levels (e.g. species). Alignment

of difficult regions due to highly divergent taxa would be overcome by this approach. The use of first tier regions would create subgroups within which alignments, even with non-coding regions, would not be a problem. They analysed more than 10,000 *rbcL* gene sequences to demonstrate that this region could serve as the first tier or “core region”. Besides the identification of new regions which would serve as barcoding candidates, there is also a need for tools which would help in faster sequence data analyses and thus help in identification of useful regions. Towards this end, Steinke et al. (2005) developed a flexible tool known as TaxI. This program calculates sequence divergences between a query sequence (taxon to be barcoded) and each sequence of a dataset of reference sequences defined by the user.

Gene transfer to the nucleus

It is now well established that chloroplasts and mitochondria were once free-living cyanobacteria and proteobacteria, respectively. However, when the coding capacity and size of the genomes of these organelles are compared to their predecessors, there is a vast difference. Organelles have a much-reduced genome as compared to their predecessors, but still retain many proteins not coded by them. Chloroplast genomes encode only about 5–10% as many proteins as free-living cyanobacteria while mitochondrial genomes encode about 1–3% as many proteins as free-living alpha-proteobacteria. This discrepancy is due to the fact that many of the genes have been transferred to the host nucleus and a complex transport system has evolved where proteins required by the organelles are encoded by the host cell nucleus and then transported into the organelles—a phenomenon known as the endosymbiotic gene transfer (Timmis et al. 2004). This loss of autonomy from a free-living organism to an endosymbiont was a gradual process where duplicate copies of genes were generated and one of the copies was transferred to the nucleus. Thus, there was an intermediary stage where the gene was present in

both the organelle and nuclear genomes till the nuclear copy acquired suitable regulatory elements and became functional. It was then that the plastid copy became redundant and might have accumulated deleterious mutations and finally got deleted. Transient states where both the copies are functional are still not known. However, cases where the functional transferred gene exists in the nucleus while a degenerate copy persists in the organelle (Brennicke et al. 1993) are known. Sanchez et al. (1996) showed that in *Arabidopsis*, the mitochondrial copy of *rps19* is defective while the recently transferred nuclear copy has acquired domains to make it functional. The transferred organelle gene has two main obstacles to cross before the nuclear copy becomes functional permitting the loss of the organelle copy. The first one is expression and the second is targeting (Martin and Herrmann 1998). If the product of the organelle gene has to return back to the organelle, then the nuclear copy of the gene has to acquire suitable regulatory elements, organelle targeting signals and transit peptides. If this happens, then the nuclear copy can become functional and supply the organelle with the required product, thereby increasing the chances that the organelle gene would get eliminated. There are several reports supporting this phenomenon. A genome-wide phylogenetic survey by Martin et al. (2002) revealed that ~18% of the protein-coding genes in the *Arabidopsis* genome were acquired from the cyanobacterial ancestor of plastids. In a similar study, two completely sequenced plant genomes – rice and *Arabidopsis*, were found to have an abundance of such insertions (Shahmuradov et al. 2003). Matsuo et al. (2005) analyzed nuclear-localized plastid DNA fragments – now known as “nupDNA” – in the *japonica* rice nuclear genome. They studied the fragments with respect to their age, size, structure and integration sites. Their work revealed the fact that plastid DNA has been transferred continuously to the nucleus. They also observed that after integration into the nuclear genome, there is fragmentation, shuffling and most of these (80%) are actually eliminated within a span of million years. Most of the insertions were near the pericentromeric regions.

Chromosome 1 had the maximum number of such insertions and chromosomes 9, 10 and 11 had the least. When the amount of insertions (kilobase pairs) was compared, chromosome 10 had the maximum amount while 11 had the least. nupDNA represents 0.2% of the total rice nuclear genome. In another study, Huang et al. (2005) analyzed two largest organelle genome copies in the nuclear genome. These two, namely, the 131 kb nupDNA (or nuptDNA) in rice and the 262 kb numtDNA (nuclear-localized mitochondrial DNA) in *Arabidopsis* were compared with their organelle counterparts. They found out that these transferred fragments are subject to mutational decay (predominantly by 5-methylcytosine) and are non-functional as shown by the absence of purifying selection. The rice 131 kb nuptDNA is absent from the *indica* subspecies and *O. rufipogon* suggesting that the transfer was after the divergence from the *indica* lineage. The time of transfer for these fragments was estimated to be around 148,000 years ago for the rice nuptDNA and 88,000 years ago for the *Arabidopsis* numtDNA (Huang et al. 2005). Thus, organellar gene transfer to the nucleus is a continuous process, which has resulted in bulk of the coding capacity of plastomes to be transferred to the host nucleus (Timmis et al. 2004). There are some forces, however, which seem to be holding back the remaining genes within the chloroplast genomes (Allen 2003). Figure 5 shows the different possible mechanisms of gene transfer to the nucleus and targeting of the protein back to the chloroplast.

Future

Chloroplast genomics is in its exponential phase, with the number of plastomes submitted to GenBank (Fig. 6) crossing the “one plastome a month” rate in the year 2005. It is estimated that nearly 200 plastomes would be completed in a span of 5 years (Jansen et al. 2005). Emergence of new techniques like Rolling Circle Amplification (RCA), tools like DOGMA (Wyman et al. 2004) for annotation, MEGA (Kumar et al. 2004), TREECON (Van de Peer 1997) and BIOEDIT (Hall 1999) have eased the pressure

on the scientific community and will aid in speeding up ongoing projects. Databases such as ChloroplastDB (Cui et al. 2006) and GOBASE (O'Brien et al. 2006) have brought together sequence data in an organized way to help in systematic data extraction.

To summarize, chloroplast genomes are excellent phylogenetic tools. Their conservation, both with respect to gene content and order, has made them a favourite among plant evolutionary biologists. Rapid progress in sequencing technology has seen a tremendous increase in the number of sequenced plastomes. Plastomes from various groups are now sequenced and the plant 'tree of life' seems an achievable goal in the near future. Problematic regions within the plant tree of life and phylogenetic hurdles like Long Branch Artifacts are bound to be solved with the introduction of improved phylogenetic methods and models of substitutions and it is only a matter of time when we would have representative(s) from each and every order and family of the plant kingdom with a completely sequenced plastome. Besides their use in phylogeny, these plastid genomes would help in rapid, accurate and automated identification of species providing plant systematists/taxonomists with an easy to use species-specific DNA barcode catalogue.

We thank an anonymous reviewer for critical reading and valuable comments and suggestions on an earlier version of this manuscript. The research work of our group is funded by the Department of Biotechnology, Government of India, New Delhi.

References

- Allen JF (2003) The function of genomes in bioenergetic organelles. *Philos Trans Roy Soc Lond B Biol Sci* 358: 19–37
- Archibald JM (2005) Jumping genes and shrinking genomes – probing the evolution of eukaryotic photosynthesis with genomics. *IUBMB Life* 57: 539–547
- Asano T, Tsudzuki T, Takahashi S, Shimada H, Kadowaki K (2004) Complete nucleotide sequence of the sugarcane (*Saccharum officinarum*) chloroplast genome: a comparative analysis of four monocot chloroplast genomes. *DNA Res* 11: 93–99
- Bausher MG, Singh ND, Lee SB, Jansen RK, Daniell H (2006) The complete chloroplast genome sequence of *Citrus sinensis* (L.) Osbeck var 'Ridge Pineapple': organization and phylogenetic relationships to other angiosperms. *BMC Pl Biol* 6: 21
- Bedbrook JR, Bogorad L (1976) Endonuclease recognition sites mapped on *Zea mays* chloroplast DNA. *Proc Natl Acad Sci USA* 73: 4309–4313
- Belanger AS, Brouard JS, Charlebois P, Otis C, Lemieux C, Turmel M (2006) Distinctive architecture of the chloroplast genome in the chlorophycean green alga *Stigeoclonium helveticum*. *Molec Genet Genomics* 276: 464–477
- Bendich AJ (2004) Circular chloroplast chromosomes: the grand illusion. *Pl Cell* 16: 1661–1666
- Boudreau E, Takahashi Y, Lemieux C, Turmel M, Rochaix JD (1997) The chloroplast *ycf3* and *ycf4* open reading frames of *Chlamydomonas reinhardtii* are required for the accumulation of the photosystem I complex. *EMBO J* 16: 6095–6104
- Brennicke A, Grohmann L, Hiesel R, Knoop V, Schuster W (1993) The mitochondrial genome on its way to the nucleus: different stages of gene transfer in higher plants. *FEBS Lett* 325: 140–145
- Brocks JJ, Logan GA, Buick R, Summons RE (1999) Archean molecular fossils and the early rise of eukaryotes. *Science* 285: 1033–1036
- Butterfield NJ (2000) *Bangiomorpha pubescens* n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. *Paleobiology* 26: 386–404
- Cai X, Fuller AL, McDougald LR, Zhu G (2003) Apicoplast genome of the coccidian *Eimeria tenella*. *Gene* 321: 39–46
- Cai Z, Penafior C, Kuehl JV, Leebens-Mack J, Carlson JE, dePamphilis CW, Boore JL, Jansen R K (2006) Complete plastid genome sequences of *Drimys*, *Liriodendron*, and *Piper*: implications for the phylogenetic relationships of magnoliids. *BMC Evol Biol* 6: 77
- Chang CC, Lin HC, Lin IP, Chow TY, Chen HH, Chen WH, Cheng CH, Lin CY, Liu SM, Chang CC, Chaw SM (2006) The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. *Molec Biol Evol* 23: 279–291
- Chase MW, Salamin N, Wilkinson M, Dunwell JM, Kesanakurthi RP, Haidar N, Savolainen V (2005) Land plants and DNA barcodes: short-term and

- long-term goals. *Philos Trans Roy Soc Lond, B, Biol Sci* 360: 1889–1895
- Chiba Y (1951) Cytochemical studies on chloroplasts I. Cytologic demonstration of nucleic acids in chloroplasts. *Cytologia (Tokyo)* 16: 259–264
- Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, Jansen RK (2006) The complete chloroplast genome sequence of *Pelargonium × hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. *Molec Biol Evol* 23: 2175–2190
- Cui L, Veeraraghavan N, Richter A, Wall K, Jansen R K, Leebens-Mack J, Makalowska I, dePamphilis C W (2006) ChloroplastDB: the chloroplast genome database. *Nucl Acids Res* 34: D692–D696
- Daniell H, Lee SB, Grevich J, Sasaki C, Quesada-Vargas T, Guda C, Tomkins J, Jansen RK (2006) Complete chloroplast genome sequences of *Solanum bulbocastanum*, *Solanum lycopersicum* and comparative analyses with other Solanaceae genomes. *Theor Appl Genet* 112: 1503–1518
- de Cambiaire JC, Otis C, Lemieux C, Turmel M (2006) The complete chloroplast genome sequence of the chlorophycean green alga *Scenedesmus obliquus* reveals a compact gene organization and a biased distribution of genes on the two DNA strands. *BMC Evol Biol* 6: 37
- de Koning AP, Keeling PJ (2006) The complete plastid genome sequence of the parasitic green alga *Helicosporidium* sp. is highly reduced and structured. *BMC Biol* 4: 12
- Douglas SE, Penny SL (1999) The plastid genome of the cryptophyte alga, *Guillardia theta*: complete sequence and conserved synteny groups confirm its common ancestry with red algae. *J Molec Evol* 48: 236–244
- Doyle JJ, Davis JJ, Soreng RJ, Garvin D, Anderson M J (1992) Chloroplast DNA inversions and the origin of the grass family (Poaceae). *Proc Natl Acad Sci USA* 89: 7722–7726
- Drescher A, Ruf S, Calsa T Jr, Carrer H, Bock R (2000) The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. *Pl J* 22: 97–104
- Embley TM, Martin W (2006) Eukaryotic evolution, changes and challenges. *Nature* 440: 623–630
- Gardner MJ, Bishop R, Shah T, de Villiers EP, Carlton JM, Hall N, Ren Q, Paulsen IT, Pain A, Berriman M, Wilson R J, Sato S, Ralph SA, Mann DJ, Xiong Z, Shallom SJ, Weidman J, Jiang L, Lynn J, Weaver B, Shoaibi A, Domingo AR, Wasawo D, Crabtree J, Wortman JR, Haas B, Angiuoli SV, Creasy TH, Lu C, Suh B, Silva JC, Utterback TR, Feldblyum TV, Pertea M, Allen J, Nierman WC, Taracha EL, Salzberg SL, White OR, Fitzhugh HA, Morzaria S, Venter JC, Fraser CM, Nene V (2005) Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. *Science* 309: 134–137
- Gargano D, Vezzi A, Scotti N, Gray JC, Valle G, Grillo S, Cardi T (2005) The complete nucleotide sequence of potato (*Solanum tuberosum* cv. Désirée) chloroplast DNA In: Abstracts of the 2nd Solanaceae Genome Workshop 2005: 107
- Glöckner G, Rosenthal A, Valentin K (2000) The structure and gene repertoire of an ancient red algal plastid genome. *J Molec Evol* 51: 382–390
- Gockel G, Hachtel W (2000) Complete gene map of the plastid genome of the nonphotosynthetic euglenoid flagellate *Astasia longa*. *Protist* 151: 347–351
- Goremykin VV, Hansmann S, Martin WF (1997) Evolutionary analysis of 58 proteins encoded in six completely sequenced chloroplast genomes: Revised molecular estimates of two seed plant divergence times. *Pl Syst Evol* 206: 337–351
- Goremykin VV, Hirsch-Ernst KI, Wolf S, Hellwig FH (2003a) The chloroplast genome of the “basal” angiosperm *Calycanthus fertilis* – structural and phylogenetic analysis. *Pl Syst Evol* 242: 119–135
- Goremykin VV, Hirsch-Ernst KI, Wolf S, Hellwig FH (2003b) Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. *Molec Biol Evol* 20: 1499–1505
- Goremykin VV, Hirsch-Ernst KI, Wolf S, Hellwig FH (2004) The chloroplast genome of *Nymphaea alba*: whole-genome analyses and the problem of identifying the most basal angiosperm. *Molec Biol Evol* 21: 1445–1454
- Goremykin VV, Holland B, Hirsch-Ernst KI, Hellwig FH (2005) Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. *Molec Biol Evol* 22: 1813–1822
- Goulding SE, Olmstead RG, Morden CW, Wolfe KH (1996) Ebb and flow of the chloroplast inverted repeat. *Molec Gen Genet* 252: 195–206
- Hager M, Biehler K, Illerhaus J, Ruf S, Bock R (1999) Targeted inactivation of the smallest plastid genome-encoded open reading frame reveals a novel and essential subunit of the cytochrome b(6) complex. *EMBO J* 18: 5834–5842
- Hagopian JC, Reis M, Kitajima JP, Bhattacharya D, de Oliveira MC (2004) Comparative analysis of

- the complete plastid genome sequence of the red alga *Gracilaria tenuistipitata* var. *liui* provides insights into the evolution of rhodoplasts and their relationship to other plastids. *J Molec Evol* 59: 464–477
- Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acids Symp Ser* 41: 95–98
- Hallick RB, Bairoch A (1994) Proposal for the naming of chloroplast genes. III. Nomenclature for open reading frames encoded in chloroplast genomes. *Pl Molec Biol Rep* 12: S29–S30
- Hallick RB, Hong L, Drager RG, Favreau MR, Monfort A, Orsat B, Spielmann A, Stutz E (1993) Complete sequence of *Euglena gracilis* chloroplast DNA. *Nucl Acids Res* 21: 3537–3544
- Hebert PD, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Proc Roy Soc Lond B Biol Sci* 270: 313–321
- Hebert PD, Gregory TR (2005) The promise of DNA barcoding for taxonomy. *Syst Biol* 54: 852–859
- Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, Kondo C, Honji Y, Sun C-R, Meng B-Y, Li Y-Q, Kanno A, Nishizawa Y, Hirai A, Shinozaki K, Sugiura M (1989) The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. *Molec Gen Genet* 217: 185–194
- Hirose T, Kusumegi T, Tsudzuki T, Sugiura M (1999) RNA editing sites in tobacco chloroplast transcripts: editing as a possible regulator of chloroplast RNA polymerase activity. *Molec Gen Genet* 262: 462–467
- Huang CY, Grünheit N, Ahmadinejad N, Timmis JN, Martin W (2005) Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. *Pl Physiol* 138: 1723–1733
- Hupfer H, Swiatek M, Hornung S, Herrmann RG, Maier RM, Chiu WL, Sears B (2000) Complete nucleotide sequence of the *Oenothera elata* plastid chromosome, representing plastome I of the five distinguishable euoenothera plastomes. *Molec Gen Genet* 263: 581–585
- Jansen RK, Kaitanis C, Lee S B, Saski C, Tomkins J, Alverson AJ, Daniell H (2006) Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. *BMC Evol Biol* 6: 32
- Jansen RK, Raubeson LA, Boore JL, dePamphilis CW, Chumley TW, Haberle RC, Wyman SK, Alverson AJ, Peery R, Herman SJ, Fourcade HM, Kuehl JV, McNeal JR, Leebens-Mack J, Cui L (2005) Methods for obtaining and analyzing whole chloroplast genome sequences. *Meth Enzymol* 395: 348–384
- Jarvis P, Soll J (2001) Toc, Tic, and chloroplast protein import. *Biochim Biophys Acta* 1541: 64–79
- Kahlau S, Aspinall S, Gray JC, Bock R (2006) Sequence of the tomato chloroplast DNA and evolutionary comparison of solanaceous plastid genomes. *J Molec Evol* 63: 194–207
- Källersjö M, Farris JS, Chase MW, Bremer B, Fay MF, Humphries CJ, Petersen G, Seberg O, Bremer K (1998) Simultaneous parsimony jackknife analysis of 2538 *rbcL* DNA sequences reveals support for major clades of green plants, land plants, seed plants and flowering plants. *Pl Syst Evol* 213: 259–287
- Kato T, Kaneko T, Sato S, Nakamura Y, Tabata S (2000) Complete structure of the chloroplast genome of a legume, *Lotus japonicus*. *DNA Res* 7: 323–330
- Kim J-S, Jung JD, Lee J-A, Park H-W, Oh K-H, Jeong WJ, Choi DW, Liu JR, Cho KY (2006) Complete sequence and organization of the cucumber (*Cucumis sativus* L. cv. Baekmibaekdadagi) chloroplast genome. *Pl Cell Rep* 25: 334–340
- Kim KJ, Lee HL (2004) Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. *DNA Res* 11: 247–261
- Kostianovsky M (2000) Evolutionary origin of eukaryotic cells. *Ultrastruct Pathol* 24: 59–66
- Kowallik KV, Stoebe B, Schaffran I, Kroth-Pancic P, Freier U (1995) The chloroplast genome of a chlorophyll a+c-containing alga, *Odontella sinensis*. *Pl Molec Biol Rep* 13: 336–342
- Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. *Proc Natl Acad Sci USA* 102: 8369–8374
- Kugita M, Kaneko A, Yamamoto Y, Takeya Y, Matsumoto T, Yoshinaga K (2003a) The complete nucleotide sequence of the hornwort (*Anthoceros formosae*) chloroplast genome: insight into the earliest land plants. *Nucl Acids Res* 31: 716–721

- Kugita M, Yamamoto Y, Fujikawa T, Matsumoto T, Yoshinaga K (2003b) RNA editing in hornwort chloroplasts makes more than half the genes functional. *Nucl Acids Res* 31: 2417–2423
- Kumar S, Tamura K, Nei M (2004) MEGA3: integrated software for molecular evolutionary analysis and sequence alignment. *Brief Bioinf* 5: 150–163
- Lee SB, Kaittani C, Jansen RK, Hostetler JB, Tallon LJ, Town CD, Daniell H (2006) The complete chloroplast genome sequence of *Gossypium hirsutum*: organization and phylogenetic relationships to other angiosperms. *BMC Genomics* 7: 61
- Leister D (2003) Chloroplast research in the genomic age. *Trends Genet* 19: 47–56
- Lemieux C, Otis C, Turmel M (2000) Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. *Nature* 403: 649–652
- Lockhart PJ, Penny D (2005) The place of *Amborella* within the radiation of angiosperms. *Trends Pl Sci* 10: 201–202
- Mäenpää P, Gonzalez EB, Chen L, Khan MS, Gray JC, Aro EM (2000) The *ycf9* (*orf 62*) gene in the plant chloroplast genome encodes a hydrophobic protein of stromal thylakoid membranes. *J Exp Bot* 51: 375–382
- Maier RM, Neckermann K, Igloi GL, Kossel H (1995) Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. *J Molec Biol* 251: 614–628
- Manning JE, Wolstenholme DR, Ryan RS, Hunter JA, Richards OC (1971) Circular chloroplast DNA from *Euglena gracilis*. *Proc Natl Acad Sci USA* 68: 1169–1173
- Margulis L (1970) *Origin of Eukaryotic Cells*, Yale University Press, New Haven
- Martin W, Deusch O, Stawski N, Grunheit N, Goremykin V (2005) Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. *Trends Pl Sci* 10: 203–209
- Martin W, Herrmann RG (1998) Gene transfer from organelles to the nucleus: how much, what happens, and why? *Pl Physiol* 118: 9–17
- Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D (2002) Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci USA* 99: 12246–12251
- Martin W, Stoebe B, Goremykin V, Hapsmann S, Hasegawa M, Kowallik K V (1998) Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* 393: 162–165
- Matsuo M, Ito Y, Yamauchi R, Obokata J (2005) The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear DNA flux. *Pl Cell* 17: 665–675
- Maul JE, Lilly JW, Cui L, dePamphilis CW, Miller W, Harris EH, Stern DB (2002) The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats. *Pl Cell* 14: 2659–2679
- Mereschkowsky C (1905) Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. *Biol. Centralbl.* 25:593–604. English translation. In: Martin W, Kowallik KV (1999) Annotated English translation of Mereschkowsky's 1905 paper 'Über Natur und Ursprung der Chromatophoren im Pflanzenreiche'. *Eur J Phycol* 34: 287–295
- Monod C, Takahashi Y, Goldschmidt-Clermont M, Rochaix JD (1994) The chloroplast *ycf8* open reading frame encodes a photosystem II polypeptide which maintains photosynthetic activity under adverse growth conditions. *EMBO J* 13: 2747–2754
- Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, Foltá KM, Soltis DE (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Pl Biol* 6: 17
- Newmaster SG, Fazekas AJ, Ragupathy S (2006) DNA barcoding in land plants: evaluation of *rbcL* in a multigene tiered approach. *Canad J Bot* 84: 335–341
- O'Brien EA, Zhang Y, Yang LS, Wang E, Marie V, Lang BF, Burger G (2006) GOBASE - a database of organelle and bacterial genome information. *Nucl Acids Res* 34: D697–D699
- Ogihara Y, Isono K, Kojima T, Endo A, Hanaoka M, Shiina T, Terachi T, Utsugi S, Murata M, Mori N, Takumi S, Ieko K, Gojobori T, Murai R, Murai K, Matsuoka Y, Ohnishi Y, Tajiri H, Tsunewaki K (2002) Structural features of a wheat plastome as revealed by complete sequencing of chloroplast DNA. *Molec Genet Genomics* 266: 740–746
- Ohta N, Matsuzaki M, Misumi O, Miyagishima S Y, Nozaki H, Tanaka K, Shin-I T, Kohara Y, Kuroiwa T (2003) Complete sequence and analysis of the plastid genome of the unicellular red alga *Cyanidioschyzon merolae*. *DNA Res* 10: 67–77
- Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, Sano S, Umeson K, Shiki Y, Takeuchi M, Chang Z, Aota S-I, Inokuchi H, Ozeki H (1986) Chloroplast gene organization deduced from complete

- sequence of liverwort *Marchantia polymorpha* chloroplast DNA. *Nature* 322: 572–574
- Palmer JD, Nugent JM, Herbon LA (1987) Unusual structure of geranium chloroplast DNA: A triple-sized inverted repeat, extensive gene duplications, multiple inversions, and two repeat families. *Proc Natl Acad Sci USA* 84: 769–773
- Perry AS, Wolfe KH (2002) Nucleotide substitution rates in legume chloroplast DNA depend on the presence of the inverted repeat. *J Molec Evol* 55: 501–508
- Pombert JF, Otis C, Lemieux C, Turmel M (2005) The chloroplast genome sequence of the green alga *Pseudoclonium akinetum* (Ulvophyceae) reveals unusual structural features and new insights into the branching order of Chlorophyte lineages. *Molec Biol Evol* 22: 1903–1918
- Race HL, Herrmann RG, Martin W (1999) Why have organelles retained genomes?. *Trends Genet* 15: 364–370
- Ravi V, Khurana JP, Tyagi AK, Khurana P (2006) The chloroplast genome of mulberry: complete nucleotide sequence, gene organization and comparative analysis. *Tree Genet Genomes* 3: 49–59
- Ravi V, Khurana JP, Tyagi AK, Khurana P (2007) Rosales sister to Fabales: towards resolving the rosid puzzle. *Molec Phylogenet Evol* 44: 488–493
- Reith ME, Munholland J (1995) Complete nucleotide sequence of *Porphyra purpurea* chloroplast genome. *Pl Molec Biol Rep* 13: 333–335
- Robbens S, Derelle E, Ferraz C, Wuyts J, Moreau H, Van de Peer Y (2007) The chloroplast and mitochondrial DNA sequence of *Ostreococcus tauri*: organelle genomes of the smallest eukaryote are examples of compaction. *Molec Biol Evol* 24: 956–968
- Rogers MB, Gilson PR, Su V, McFadden GI, Keeling PJ (2007) The complete chloroplast genome of the chlorarachniophyte *Bigeloviella natans*: evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. *Molec Biol Evol* 24: 54–62
- Rolland N, Dorne AJ, Amoroso G, Sültmeyer D, Joyard J, Rochaix JD (1997) Disruption of the plastid *ycf10* open reading frame affects uptake of inorganic carbon in the chloroplasts of *Chlamydomonas*. *EMBO J* 16: 6713–6726
- Ruhlman T, Lee SB, Jansen RK, Hostetler JB, Tallon LJ, Town CD, Daniell H (2006) Complete plastid genome sequence of *Daucus carota*: implications for biotechnology and phylogeny of angiosperms. *BMC Genomics* 7: 222
- Sager R, Ishida MR (1963) Chloroplast DNA in *Chlamydomonas*. *Proc Natl Acad Sci USA* 50: 725–730
- Samson N, Bausher MG, Lee SB, Jansen RK, Daniell H (2007) The complete nucleotide sequence of the coffee (*Coffea arabica* L.) chloroplast genome: organization and implications for biotechnology and phylogenetic relationships amongst angiosperms. *Pl Biotech J* 5: 339–353
- Sanchez H, Fester T, Kloska S, Schroder W, Schuster W (1996) Transfer of *rps19* to the nucleus involves the gain of an RNP-binding motif which may functionally replace RPS13 in *Arabidopsis* mitochondria. *EMBO J* 15: 2138–2149
- Sánchez Puerta MV, Bachvaroff TR, Delwiche CF (2005) The complete plastid genome sequence of the haptophyte *Emiliania huxleyi*: a comparison to other plastid genomes. *DNA Res* 12: 151–156
- Sanderson MJ, Driskell AC (2003) The challenge of constructing large phylogenetic trees. *Trends Pl Sci* 8: 374–379
- Saski C, Lee S-B, Daniell H, Wood TC, Tomkins J, Kim HG, Jansen RK (2005) Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes. *Pl Molec Biol* 59: 309–322
- Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S (1999) Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Res* 6: 283–290
- Saunders GW (2005) Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future applications. *Philos Trans Roy Soc Lond B Biol Sci* 360: 1879–1888
- Schimper AFW (1883) Über die Entwicklung der Chlorophyllkörner und Farbkörper. *Bot. Zeitung* 41: 105–114, 121–131, 137–146, 153–162
- Schmitz-Linneweber C, Maier RM, Alcaraz JP, Cottet A, Herrmann RG, Mache R (2001) The plastid chromosome of spinach (*Spinacia oleracea*): complete nucleotide sequence and gene organization. *Pl Molec Biol* 45: 307–315
- Schmitz-Linneweber C, Regel R, Du TG, Hupfer H, Herrmann RG, Maier RM (2002) The plastid chromosome of *Atropa belladonna* and its comparison with that of *Nicotiana tabacum*: The role of RNA editing in generating divergence in the process of speciation. *Molec Biol Evol* 19: 1602–1612
- Shahid Masood M, Nishikawa T, Fukuoka S, Njenga PK, Tsudzuki T, Kadowaki K-I (2004) The complete nucleotide sequence of wild rice (*Oryza*

- nivara*) chloroplast genome: first genome wide comparative sequence analysis of wild and cultivated rice. *Gene* 340: 133–139
- Shahmuradov IA, Akbarova YY, Solovyev VV, Aliyev JA (2003) Abundance of plastid DNA insertions in nuclear genomes of rice and *Arabidopsis*. *Pl Molec Biol* 52: 923–934
- Shimada H, Sugiura M (1991) Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes. *Nucl Acids Res* 19: 983–995
- Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, Ohto C, Torazawa K, Meng BY, Sugita M, Deno H, Kamogashira T, Yamada K, Kusuda J, Takaiwa F, Kato A, Tohdoh N, Shimada H, Sugiura M (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *EMBO J* 5: 2043–2049
- Soltis DE, Albert VA, Savolainen V, Hilu K, Qiu YL, Chase MW, Farris JS, Stefanovic S, Rice DW, Palmer JD, Soltis PS (2004) Genome-scale data, angiosperm relationships, and ‘ending incongruence’: a cautionary tale in phylogenetics. *Trends Pl Sci* 9: 477–483
- Steane DA (2005) Complete nucleotide sequence of the chloroplast genome from the Tasmanian blue gum, *Eucalyptus globulus* (Myrtaceae). *DNA Res* 12: 215–220
- Stefanovic S, Rice DW, Palmer JD (2004) Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? *BMC Evol Biol* 4: 35
- Steinke D, Vences M, Salzburger W, Meyer A (2005) TaxI: a software tool for DNA barcoding using distance methods. *Philos Trans Roy Soc Lond B Biol Sci* 360: 1975–1980
- Stocking C, Gifford E (1959) Incorporation of thymidine into chloroplasts of *Spirogyra*. *Biochem Biophys Res Commun* 1: 159–164
- Stoebe B, Martin W, Kowallik K V (1998) Distribution and nomenclature of protein-coding genes in 12 sequenced chloroplast genomes. *Pl Molec Biol Rep* 16: 243–255
- Sugiura C, Kobayashi Y, Aoki S, Sugita C, Sugita M (2003) Complete chloroplast DNA sequence of the moss *Physcomitrella patens*: evidence for the loss and relocation of *rpoA* from the chloroplast to the nucleus. *Nucl Acids Res* 31: 5324–5331
- Sugiura M (1992) The chloroplast genome. *Plant Mol Biol* 19: 149–168
- Swiatek M, Kuras R, Sokolenko A, Higgs D, Olive J, Cinque G, Muller B, Eichacker LA, Stern DB, Bassi R, Herrmann RG, Wollman FA (2001) The chloroplast gene *ycf9* encodes a photosystem II (PSII) core subunit, *psbZ*, that participates in PSII supramolecular architecture. *Pl Cell* 13: 1347–1368
- Taberlet P, Coissac E, Pompanon F, Gielly L, Miquel C, Valentini A, Vermet T, Corthier G, Brochmann C, Willerslev E (2007) Power and limitations of the chloroplast *trnL* (UAA) intron for plant DNA barcoding. *Nucl Acids Res* 35: e14
- Takahashi Y, Rahire M, Breyton C, Popot JL, Joliot P, Rochaix JD (1996) The chloroplast *ycf7* (*petL*) open reading frame of *Chlamydomonas reinhardtii* encodes a small functionally important subunit of the cytochrome b6f complex. *EMBO J* 15: 3498–3506
- Tang J, Xia H, Cao M, Zhang X, Zeng W, Hu S, Tong W, Wang J, Wang J, Yu J, Yang H, Zhu L (2004) A comparison of rice chloroplast genomes. *Pl Physiol* 135: 412–420
- Taylor F (1987) An overview of the status of evolutionary cell symbiosis theories. *Ann N Y Acad Sci* 503: 1–16
- Tillich M, Lehwark P, Morton BR, Maier U.G (2006) The evolution of chloroplast RNA editing. *Molec Biol Evol* 23: 1912–1921
- Timme RE, Kuehl JV, Boore JL, Jansen RK (2007) A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: identification of divergent regions and categorization of shared repeats. *Amer J Bot* 94: 302–312
- Timmis JN, Ayliffe MA, Huang CY, Martin W (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet* 5: 123–135
- Turmel M, Otis C, Lemieux C (1999) The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: insights into the architecture of ancestral chloroplast genomes. *Proc Natl Acad Sci USA* 96: 10248–10253
- Turmel M, Otis C, Lemieux C (2002) The chloroplast and mitochondrial genome sequences of the charophyte *Chaetosphaeridium globosum*: insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. *Proc Natl Acad Sci USA* 99: 11275–11280
- Turmel M, Otis C, Lemieux C (2005) The complete chloroplast DNA sequences of the charophycean green algae *Staurastrum* and *Zygnema* reveal that the chloroplast genome underwent extensive changes during the evolution of the Zygnematales. *BMC Biol* 3: 22

- Van de Peer Y, De Wachter R (1997) Construction of evolutionary distance trees with TREECON for Windows: accounting for variation in nucleotide substitution rate among sites. *Comput Appl Biosci* 13: 227–230
- Wakasugi T, Nagai T, Kapoor M, Sugita M, Ito M, Ito S, Tsudzuki J, Nakashima K, Tsudzuki T, Suzuki Y, Hamada A, Ohta T, Inamura A, Yoshinaga K, Sugiura M (1997) Complete nucleotide sequence of the chloroplast genome from the green alga *Chlorella vulgaris*: the existence of genes possibly involved in chloroplast division. *Proc Natl Acad Sci USA* 94: 5967–5972
- Wakasugi T, Tsudzuki J, Ito S, Nakashima K, Tsudzuki T, Sugiura M (1994) Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. *Proc Natl Acad Sci USA* 91: 9794–9798
- Wolf PG, Karol KG, Mandoli DF, Kuehl J, Arumuganathan K, Ellis MW, Mishler BD, Kelch DG, Olmstead RG, Boore JL (2005) The first complete chloroplast genome sequence of a lycophyte, *Huperzia lucidula* (Lycopodiaceae). *Gene* 350: 117–128
- Wolf PG, Rowe CA, Hasebe M (2004) High levels of RNA editing in a vascular plant chloroplast genome: analysis of transcripts from the fern *Adiantum capillus-veneris*. *Gene* 339: 89–97
- Wolf PG, Rowe CA, Sinclair RB, Hasebe M (2003) Complete nucleotide sequence of the chloroplast genome from a leptosporangiate fern, *Adiantum capillus-veneris* L. *DNA Res* 10: 59–65
- Wolfe KH, Morden CW, Palmer JD (1992) Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proc Natl Acad Sci USA* 89: 10648–10652
- Wyman S, Jansen R, Boore J (2004) Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20: 3252–3255
- Xie Z, Merchant S (1996) The plastid-encoded *ccsA* gene is required for heme attachment to chloroplast c-type cytochromes. *J Biol Chem* 271: 4632–4639
- Yoshinaga K, Iinuma H, Masuzawa T, Uedal K (1996) Extensive RNA editing of U to C in addition to C to U substitution in the *rbcL* transcripts of hornwort chloroplasts and the origin of RNA editing in green plants. *Nucl Acids Res* 24: 1008–1014
- Yoshinaga K, Kakehi T, Shima Y, Iinuma H, Masuzawa T, Ueno M (1997) Extensive RNA editing and possible double-stranded structures determining editing sites in the *atpB* transcripts of hornwort chloroplasts. *Nucl Acids Res* 25: 4830–4834
- Yukawa M, Tsudzuki T, Sugiura M (2006) The chloroplast genome of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*: complete sequencing confirms that the *Nicotiana sylvestris* progenitor is the maternal genome donor of *Nicotiana tabacum*. *Molec Genet Genomics* 275: 367–373