J Mol Evol (2008) 66:555–564
DOI 10.1007/s00239-008-9091-7

Complete Sequence of the Duckweed (*Lemna minor*) Chloroplast Genome: Structural Organization and Phylogenetic Relationships to Other Angiosperms

Andrey V. Mardanov · Nikolai V. Ravin · Boris B. Kuznetsov · Tahir H. Samigullin · Andrey S. Antonov · Tatiana V. Kolganova · Konstantin G. Skyabin

Received: 10 May 2007 / Accepted: 21 February 2008 / Published online: 8 May 2008

Abstract The complete nucleotide sequence of the duckweed (*Lemna minor*) chloroplast genome (cpDNA) was determined. The cpDNA is a circular molecule of 165,955 bp containing a pair of 31,223-bp inverted repeat regions (IRs), which are separated by small and large single-copy regions of 89,906 and 13,603 bp, respectively. The entire gene pool and relative positions of 112 genes (78 protein-encoding genes, 30 tRNA genes, and 4 rRNA genes) are almost identical to those of *Amborella trichopoda* cpDNA; the minor difference is the absence of *infA* and *ycf15* genes in the duckweed cpDNA. The inverted repeat is expanded to include *ycf1* and *rps15* genes; this pattern is unique and does not occur in any other sequenced cpDNA of land plants. As in basal angiosperms and eudicots, but not in other monocots, the borders between IRs and a large single-copy region are located upstream of *rps19* and downstream of *trnH*, so that *trnH* is not included in IRs. The model of rearrangements of the chloroplast genome during the evolution of monocots is proposed as the result of the comparison of cpDNA structures in duckweed and other monocots. The phylogenetic analyses of 61 protein-coding genes from 38 plastid genome sequences provided strong support for the monophyly of monocots and position of *Lemna* as the next diverging lineage of monocots after Acorales. Our analyses also provided support for *Amborella* as a sister to all other angiosperms, but in the bayesian phylogeny inference based on the first two codon positions *Amborella* united with Nymphaeales.

Keywords Chloroplast genome · *Lemna minor* · Monocots · Phylogeny · Angiosperms

Electronic supplementary material The online version of this article (doi:10.1007/s00239-008-9091-7) contains supplementary material, which is available to authorized users.

A. V. Mardanov · N. V. Ravin · B. B. Kuznetsov · T. V. Kolganova · K. G. Skyabin (✉)
Centre "Bioengineering", Russian Academy of Sciences, Prosp. 60-let Oktyabrya, bld.7-1, Moscow 117312, Russia
e-mail: nravin@biengi.ac.ru

T. H. Samigullin · A. S. Antonov
A. N. Belozersky Institute of Physico-Chemical Biology, Moscow State University, Moscow, Russia

Introduction

Chloroplast is a plant organelle containing the entire enzymatic machinery for photosynthesis; it is supposed to have evolved from ancient endosymbiotic cyanobacteria. Chloroplast genes are responsible for >50% of the total leaf soluble protein encoded by both nuclear and chloroplast genomes. Over the past two decades, the plastid genome and its structure, expression, and evolution have been extensively studied using molecular methods (Sugiura 1992; Wakasugi et al. 2001). The rapid progress in this area has been determined by a permanent increase in the number of newly sequenced chloroplast genomes. The complete nucleotide sequences of the chloroplast genomes are available for more than 40 plants, including 2 ferns, *Psilotum nudum* and *Adiantum capillus-veneris*; 2 gymnosperms, *Pinus thunbergii* and *Pinus koraiensis;* and more than 20 dicots, including the putative basal angiosperms (Soltis et al. 1999) *Amborella trichopoda* (Goremykin et al. 2003a), *Calycanthus floridus* (Goremykin et al. 2003b), and *Nymphaea alba* (Goremykin et al. 2004). Among monocots, which have been suggested to be the most ancient branch of angiosperms (Goremykin et al. 2003a),

genome sequences are available for six species: *Acorus calamus* (Goremykin et al. 2005) (*Acoraceae*), *Phalaenopsis aphrodite* (Chang et al. 2006) (*Orchidaceae*), and four grasses (*Poaceae*)—*Oryza sativa* (Hiratsuka et al. 1989), *Triticum aestivum* (Ogihara et al. 2002), *Zea mays* (Maier et al. 1995), and *Saccharum officinarum* (Asano et al. 2004).

Physical mapping studies and available sequence data revealed that chloroplast genomes of most land plants are highly conserved with respect to their size, ranging from 120 to 217 kb, and structure (Palmer et al. 1987). The presence of a large inverted repeat (IR), which ranges from 5 to 76 kb in length (Palmer 1991), is one of the conserved structural features of chloroplast genomes. The majority of size variations between the genomes can be accounted for by variations in the size of IR and intergenic spacers (Wakasugi et al. 2001; Raubeson and Jansen 2005). IR consists of two completely identical segments, IRA and IRB, which are typically 10–25 kb long but may range from 6 to 76 kb (Palmer 1985; Shinozaki et al. 1986). The repeated segments are separated by long single-copy (LSC) and small single-copy (SSC) regions. The completely sequenced algal and plant chloroplast genomes contain from 63 to 209 genes, with an average number of 110–130 (Jansen and Palmer 1987). The gene content and the polycistronic transcription units of the chloroplast genome are also conserved among the majority of vascular plant species (Kim and Lee 2004). The gene order in chloroplast genome is relatively conserved but sometimes disturbed by invertive mutations that can be mediated by intramolecular recombination or by multiple extensions or reductions of the IR sequences (Perry et al. 2002).

In this work, we present the complete sequence of the cpDNA (165955 bp; minimally 3 and up to 15 reads for each base pair) of duckweed *Lemna minor*—a floating aquatic plant belonging to the monocot family *Lemnaceae*. *L. minor* is one of the most primitively organized flowering plants. Although the *Lemnaceae* have long been associated with the *Araceae* (Les et al. 2002), relationships between the *Lemnaceae* and other monocots remain uncertain (Mayo et al. 1997). We compared *Lemna minor* cpDNA with available cpDNA sequences of vascular plants to review the evolutionary modes of chloroplast genomes, with particular emphasis on the evolution of chloroplast genomes of monocots.

Duckweed is a very promising object for biotech applications: it can be used as an efficient gene expression system. *Lemna* is one of the fastest-growing higher plants, doubling its biomass every 1.5 days; it achieves this high growth rate through clonal proliferation. The *Lemna* biomass protein averages 35% dry weight of the plant. It is possible to obtain transgenic duckweed using an *Agrobacterium*-mediated method (Yamamoto et al. 2001). Taking into account the recent improvements in transplastomic techniques, duckweed chloroplast transformation will be possible in the very near future. Therefore, knowledge of the nucleotide sequence of *Lemna minor* chloroplast genome would be helpful in the construction of the expression cassettes for the stable expression of heterologous proteins upon chloroplast transformation (Maliga 2002).

Materials and Methods

The *Lemna minor* specimen was obtained from its natural habitat. Its identity to *Lemna minor* species was confirmed by morphological analysis with subsequent sequencing of the *trnL-trnF* chloroplast intergenic spacer, which can be used for discrimination between different *Lemna* species (Rothwell et al. 2004). About 40 g of the green tissue was harvested after vegetative propagation of a single duckweed plant over 3 months. Total DNA was extracted using the CTAB-based method (Murray and Thompson 1980) and purified by electrophoresis in low-melting-point agarose.

The fragments of chloroplast genomic DNA were amplified by PCR. In brief, PCR primers were designed using the alignment of known chloroplast genomic sequences of angiosperms (sequences of primers are available online as supplementary material). Using these primers, we covered the entire chloroplast genome of *Lemna minor* with overlapping PCR fragments ranging in size from 1 to 8 kb. Each fragment was sequenced independently. Some fragments, nonamplifiable with the initial "consensus" primers, were amplified with primers designed based on the newly determined sequences of adjacent regions. Automated sequencing was performed on ABI 3100 and ABI 3730 sequencers using the Big Dye Terminator v.3.1 sequencing kit (ABI, USA). The sequence fragments were assembled using the Gene Studio program (http://www.genestudio.com). All fragments were sequenced 3–15 times (6 times on average). The GenBank accession number for the nucleotide sequence determined in this study is DQ400350.

Structural RNA genes were identified by BLAST search (Altschul et al. 1990) against the GenBank database. The tRNAscan-SE program (Lowe and Eddy 1997) was applied for the assignment of tRNA genes. Gene annotations were performed using the chloroplast annotation package DOGMA (Wyman et al. 2004) (http://phylocluster.biosci.utexas.edu/dogma/). The correctness of the annotation for all genes was additionally verified by similarity search against the available plant chloroplast genome sequences. Sequence alignments were performed using the ClustalX software package (Thompson et al. 1997).

Our phylogenetic analysis did not include all available plastid genomes, representatives of eudicots were restricted to 18, because the position of *Lemna* among monocots was of main interest in this study. A set of 61 protein-coding genes from 38 plastid genome sequences (Supplementary Table S2) was collected. Nucleotide sequences of the genes were checked for frameshift mutations and corrected when necessary, then translated into amino acid sequences, which were aligned using MUSCLE ver. 3.6 (Edgar 2004) with manual correction. Nucleotide sequences were aligned according to the aligned amino acid sequences.

Gap regions were excluded from the analysis when gapped positions were more than one-third of a column. The 5% chi-square test of nucleotide or amino acid compositional homogeneity and alternative topologies test were performed with the Tree-Puzzle program (Schmidt et al. 2002). Phylogenetic analyses using maximum parsimony (MP) method were performed using PAUP* ver. 4.0b10 (Swofford 2003). Bayesian inference of phylogeny was explored using the MrBayes program ver. 3.1.2 (Ronquist and Huelsenbeck 2003).

MP analysis involved a heuristic search using TBR branch swapping and 100 random addition replicates. Nonparametric bootstrap analysis (Felsenstein 1985) was performed with 100 replicates with TBR branch swapping.

Bayesian approach was applied for both the amino acid and the nucleotide data set. The amino acid data set was divided into 61 partitions, and for each partition the most appropriate model of subsititutions was determined by the BIC in Modelgenerator ver. 0.43 (Keane et al. 2006). The models CPREV and JTT with the presence of rate variation among sites (+Γ) and/or invariable sites (+I) in some partitions were chosen. For one partition the MTREV24 model was specified (see details in Supplementary Table S3). The Bayesian inference was performed with two runs with three chains in each; 2,500,000 replicates were generated, and trees were sampled every 100 generations. The proportion of the invariable sites and the shape of gamma-distribution of the rates were unlinked across the partitions. The number of discarded trees was determined using the convergence diagnostic.

For the nucleotide data set partitioned and unpartitioned approaches were applied. The data set was divided into 61 partitions. Nucleotide frequencies and parameters of the substitution matrix were unlinked across the partitions. For each partition the most appropriate model of the nucleotide substitution was determined by the AIC in Modeltest ver. 3.7 (Posada and Crandall 1998). The models GTR+Γ, HKY+Γ, K2P+Γ, and SYM+Γ, with the presence of invariable sites (+I) for some partitions, were chosen. In unpartitioned analysis the GTR+I+Γ model was used. Bayesian analysis of nucleotide sequences was performed with two runs with three chains in each. Four million replicates were generated, and trees were sampled every 100 generations.

To achieve the nucleotide composition homogenity, the third codon positions were excluded and the representatives of Fabales (*Medicago, Lotus, Glycine*) were deleted from the data matrix, then the most appropriate model of the nucleotide substitution for the whole data set was determined and the Bayesian analysis was repeated.

Results and Discussion

The Overall Structure and Gene Pool of the *L. minor* Chloroplast Genome

The *L. minor* chloroplast genome includes a pair of inverted repeats of 31,223 bp (IRA and IRB) separated by 13,603-bp-long SSC and 89,906-bp-long LSC. The total genome size is 165,955 bp; thus, it is one of the largest chloroplast genomes among recently sequenced ones.

The overall A+T content of *L. minor* cpDNA is 64.3%, a value similar to those of tobacco (62.2%), *A. thaliana* (63.7%), rice (61.1%), and maize (61.5%). The A+T contents of the LSC and SSC regions were 66.5% and 69.9%, respectively, whereas that of the IR regions was 59.9%. The lower A+T contents of the IR regions reflect the low A+T content of rRNA genes in this region.

The assignment of the potential genes was performed by similarity search, and the positions of 112 genes, including 95 unique and 17 duplicated ones in the inverted repeat regions, were localized on the map (Table 1 and Fig. 1). *L. minor* chloroplast genome is colinear to those of tobacco (Shinozuki et al. 1986) and the basal angiosperm *A. trichopoda* (Goremykin et al. 2003a) with respect to the gene order and overall homology. A number of rearrangements and deletions specific to the completely sequenced plastomes of monocots of the Poaceae family — *Oryza* (Hiratsuka et al. 1989), *Zea* (Maier et al. 1995), *Triticum* (Ogihara et al. 2002), and *Saccharum* (Asano et al. 2004)—were not found in *L. minor* cpDNA: namely, three inversions in the LSC, the translocation of *rpl*23 gene from the inverted repeat to the LSC, loss of an intron in the *rpo*C1 gene, and a large insertion in the *rpo*C2 gene.

The gene pool of the *L. minor* chloroplast genome is almost identical to that of *Amborella* with two exceptions: absence of the *inf*A gene, encoding translation initiation factor, and absence of the conserved open reading frame (ORF) *ycf*15. The *inf*A gene is absent in the *Arabidopsis*, *Lotus*, and *Medicago* chloroplast genomes but is present as a truncated pseudogene in several other genomes (Millen et al. 2001). The putative *ycf*15 gene has also been lost several times during the evolution of vascular plants (it is absent in *Psilotum, Adiantum, Pinus, Oryza, Triticum, Zea,*

Table 1 Genes contained in the *Lemna minor* chloroplast genome (a total of 112 genes)

Category	Group of genes	Genes
Transcription and translation (59)	rRNA genes (4)	rrn16(x2), rrn23(x2)
		rrn4.5(x2), rrn5(x2)
	tRNA genes (30)	trnF-GAA, trnL-UAA*, trnL-CAA(x2), trnL-UAG, trnI-GAU*(x2), trnI-CAU(x2), trnM-CAU, trnfM-CAU, trnV-GAC(x2), trnV-UAC*, trnS-GGA, trnS-UGA, trnP-UGG, trnT-GGU, trnT-UGU, trnA-UGC*(x2), trnY-GUA, trnH-GUG, trnQ-UUG, trnN-GUU(x2), trnK-UUU*, trnD-GUC, trnE-UUC, trnC-GCA, trnW-CCA, trnR-ACG(x2), trnS-GCU, trnR-UCU, trnG-GCC, trnG-UCC*
	Small subunit of ribosome (12)	rps2, rps3, rps4, rps7(x2), rps8, rps11, rps12_5′end, rps12_3′end*(x2), rps14, rps15(x2), rps16*, rps18, rps19
	Large subunit of ribosome(9)	rpl2*(x2), rpl14, rpl16*, rpl20, rpl22, rpl23(x2), rpl32, rpl33, rpl36
	RNA polymerase (4)	rpoA, rpoB, rpoC1*, rpoC2
Photosynthesis (45)	Large subunit of RuBisCo (1)	rbcL
	Photosystem I (6)	psaA, psaB, psaC, psaI, psaJ, ycf3**
	Photosystem II (15)	psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ
	NADH dehydro-genase (11)	ndhA*, ndhB*(x2), ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK
	Cytochrome b/f complex (6)	petA, petB*, petD*, petG, petL, petN
	ATP synthase (6)	atpA, atpB, atpE, atpF*, atpH, atpI
Other genes (8)	Maturase (1)	matK
	Protease (1)	clpP**
	Envelope membrane protein (1)	cemA
	Subunit of acetyl-CoA-carboxylase (1)	accD
	c-type cytochrome synthesis gene (1)	ccsA
	Conserved genes with unknown functions (3)	ycf1(x2), ycf2(x2), ycf4
Putative pseudogenes		ycf15(x2), ycf68 (x2)

Note. One and two superscript asterisks indicate one- and two-intron-containing genes, respectively. Genes located in the IR region are indicated by (x2) after the gene name

Lotus, Medicago, and *Arabidopsis*, [Kim and Lee 2004]). Recent evidences (Cai et al. 2006) suggest that *ycf*15 is not a functional protein-coding gene; in *L. minor* the *ycf*15 is apparently a pseudogene, since its sequence is interrupted by a stop codon located only 108 bp downstream from the start. Another conserved ORF in the duckweed chloroplast genome, *ycf*68 (Stoebe et al. 1998), is also a pseudogene, since this ORF contains a frameshift mutation 42 bp downstream from the start codon. Unlike another monocot, *P. aphrodite*, the duckweed chloroplast genome contains a full set of *ndh* genes.

RNA-coding genes were identified by similarity search. Two identical copies of rRNA gene clusters (16S–23S–4.5S–5S) were found in inverted repeat regions. Each cluster was intervened by two tRNA genes, *trn*I and *trn*A, in the 16S–23S spacer region. A total of 30 tRNA genes, six of them having additional copies in the inverted repeats, were identified (Table 2). These 30 tRNA types can recognize all the codons present in the chloroplast genes, and therefore, no import of nuclear-encoded tRNAs is necessary to complement the chloroplast tRNA set. Six tRNA genes, *trn*K-UUU, *trn*V-UAC, *trn*L-UAA, *trn*G-UCC, *trn*I-GAU, and *trn*A-UGC, contained introns.

Eighteen *L. minor* chloroplast genes contained one or two introns. Only one of them, the *trn*L-UAA gene intron, belongs to the self-splicing group I, while all the others belong to group II. Two genes, *clp*P and *rps*12, possess two introns. The *rps*12 gene, as in the tobacco chloroplast genome, is a uniquely divided gene in which the 5′ exon is located in the LSC far away from its second and third exons, which are located as duplicates in the IR regions, thus requiring a *trans*-splicing mechanism between exon I and exon II to produce mature *rps*12 mRNA (Sugiura et al. 1987).

Earlier it was shown (Hoch et al. 1991) that RNA editing plays an important role in the translation process in chloroplasts. In at least two cases, RNA editing is expected to be performed in duckweed. Two genes, *rpl*2 and *ndh*D, possess ACG instead of ATG at the translation initiation site, as observed in maize (for *rpl2*), rice (for *rpl2*), *Acorus calamus (*for both *rpl2* and *ndhD)*, and several other plants. Creation of the AUG initiator codon

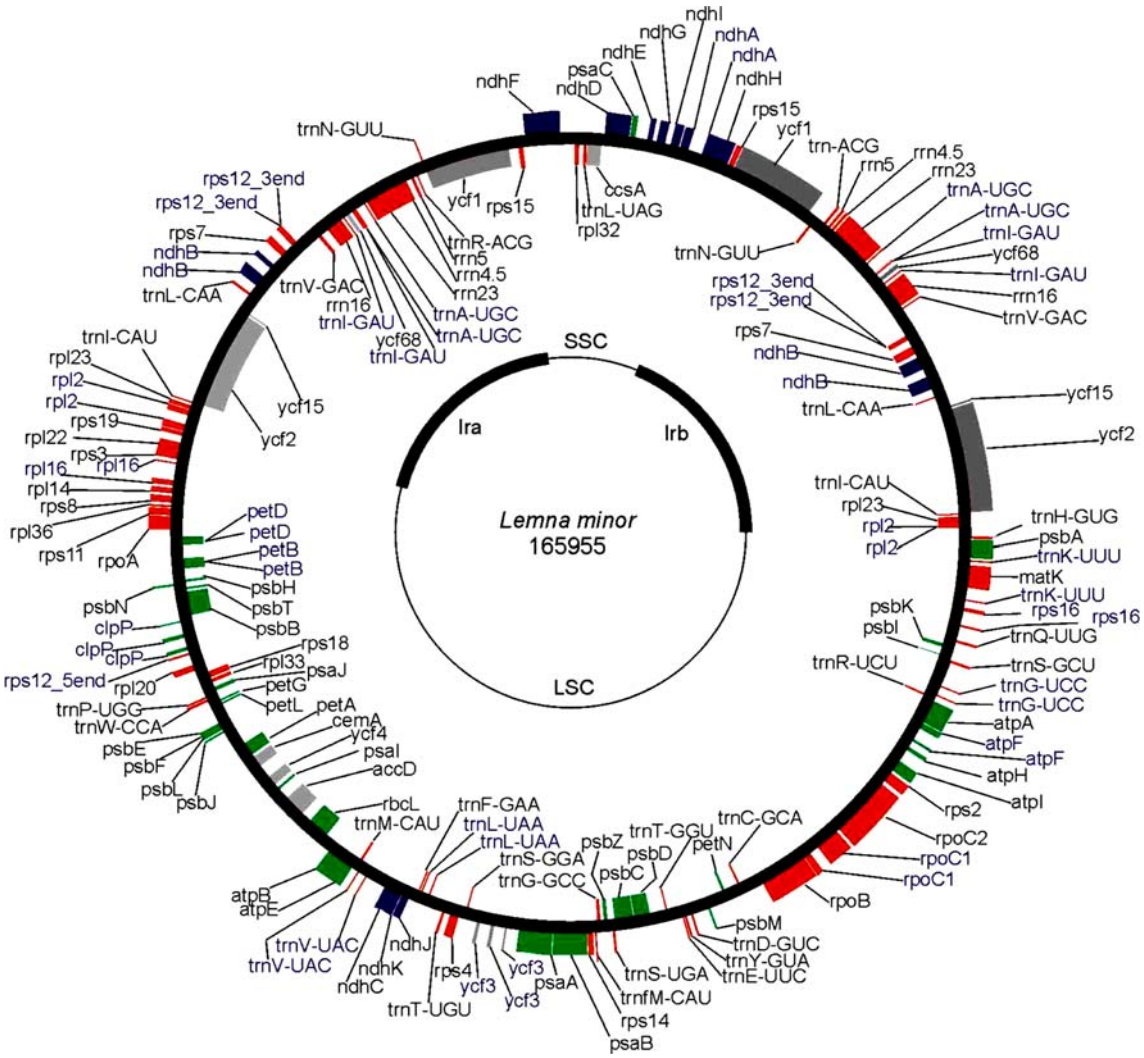

Fig. 1 *Lemna minor* cp DNA. The genes shown inside the circle are transcribed clockwise; those outside the circle are transcribed counterclockwise. The genes of the genetic apparatus are shown in red, photosynthesis genes are shown in green, and genes of NADH dehydrogenase are shown in violet. Gray color marks ORFs, ycfs, and genes of unknown function. Intron-containing genes are represented by their exons. Their names are given in blue

from the ACG codon by C-to-U editing has been shown to occur in tobacco, spinach, and snapdragon (Neckermann et al. 1994).

Phylogenetic Implications

The data set for phylogenetic analyses consisted of 61 concatenated protein-coding gene sequences for 38 taxa, including 36 angiosperms and 2 gymnosperms as outgroups. The data matrix after exclusion of ambiguously aligned positions contained 42522 nucleotide positions.

Analysis of nucleotide composition shows that within each of 59 genes, compositional homogeneity is present (*rpoC1* and *ccsA* are heterogeneous), but 15 of 38 concatenated sequences do not pass the 5% chi-square test of compositional homogeneity: these are *Ginkgo*, basal angiosperms, three of four magnoliids, *Lemna*, and seven eudicots. Amino acid composition is stationary and all amino acid sequences pass the 5% chi-square test.

MP analyses of all aligned nucleotide positions resulted in a single fully resolved tree, most nodes of which gained high support in bootstrap analyses (Fig. 2A). The topologies of the Bayesian trees derived from the partitioned and unpartitioned full-length nucleotide matrix and amino acid sequences analyses are identical (Fig. 2B). Although analysis of the unpartitioned and partitioned full-length nucleotide matrix gave the same tree topology, partitioned Bayesian analysis looks more preferable if we compare the harmonic means of the marginal likelihoods (−387,220.98 and −388,889.80 for partitioned and unpartitioned analysis, respectively). The resulting tree of the Bayesian analysis of the unpartitioned nucleotide matrix after exclusion of the

Table 2 The codon-anticodon recognition pattern and tRNA genes identified in the *L. minor* chloroplast genome

UUU	Phe	trnF-GAA	UCU	Ser	trnS-GGA	UAU	Tyr	trnY-GUA	UGU	Cys	trnC-GCA
UUC	Phe		UCC	Ser		UAC	Tyr		UGC	Cys	
UUA	Leu	trnLUAA*	UCA	Ser	trnS-UGA	UAA	—		UGA	—	
UUG	Leu	trnL-CAA(x2)	UCG	Ser		UAG	—		UGG	Trp	trnW-CCA
CUU	Leu	trnL-UAG	CCU	Pro	trnP-UGG	CAU	His	trnH-GUG	CGU	Arg	trnR-ACG(x2)
CUC	Leu		CCC	Pro		CAC	His		CGC	Arg	
CUA	Leu		CCA	Pro		CAA	Gln	trnQ-UUG	CGA	Arg	
CUG	Leu		CCG	Pro		CAG	Gln		CGG	Arg	
AUU	Ile	trnI-GAU*(x2)	ACU	Thr	trnT-GGU	AAU	Asn	trnN-GUU(x2)	AGU	Ser	trnS-GCU
AUC	Ile		ACC	Thr		AAC	Asn		AGC	Ser	
AUA	Ile	trnI-CAU(2)	ACA	Thr	trnT-UGU	AAA	Lys	trnK-UUU*	AGA	Arg	trnR-UCU
AUG	Met	trnM-CA trnfM-CAU	ACG	Thr		AAG	Lys		AGG	Arg	
GUU	Val	trnV-GAC(x2)	GCU	Ala	TrnA-UGC*(x2)	GAU	Asp	trnD-GUC	GGU	Gly	trnG-GCC
GUC	Val		GCC	Ala		GAC	Asp		GGC	Gly	
GUA	Val	trnV-UAC*	GCA	Ala		GAA	Glu	trnE-UUC	GGA	Gly	trnG-UCC*
GUG	Val		GCG	Ala		GAG	Glu		GGG	Gly	

Note. One and two superscript asterisks indicate one- and two-intron-containing genes, respectively. Genes located in the IR region are indicated by (x2) after the gene name

third codon positions and representatives of Fabaceae is shown in Fig. 2C.

In all obtained phylogenetic trees, monocots are monophyletic and relationships among them are insensitive to the method of phylogeny inference and amount of data under analysis, *Lemna minor* represents the next branch after separation of *Acorus*. Among rosids, placement of *Cucumis* is unstable: *Cucumis* is either united with

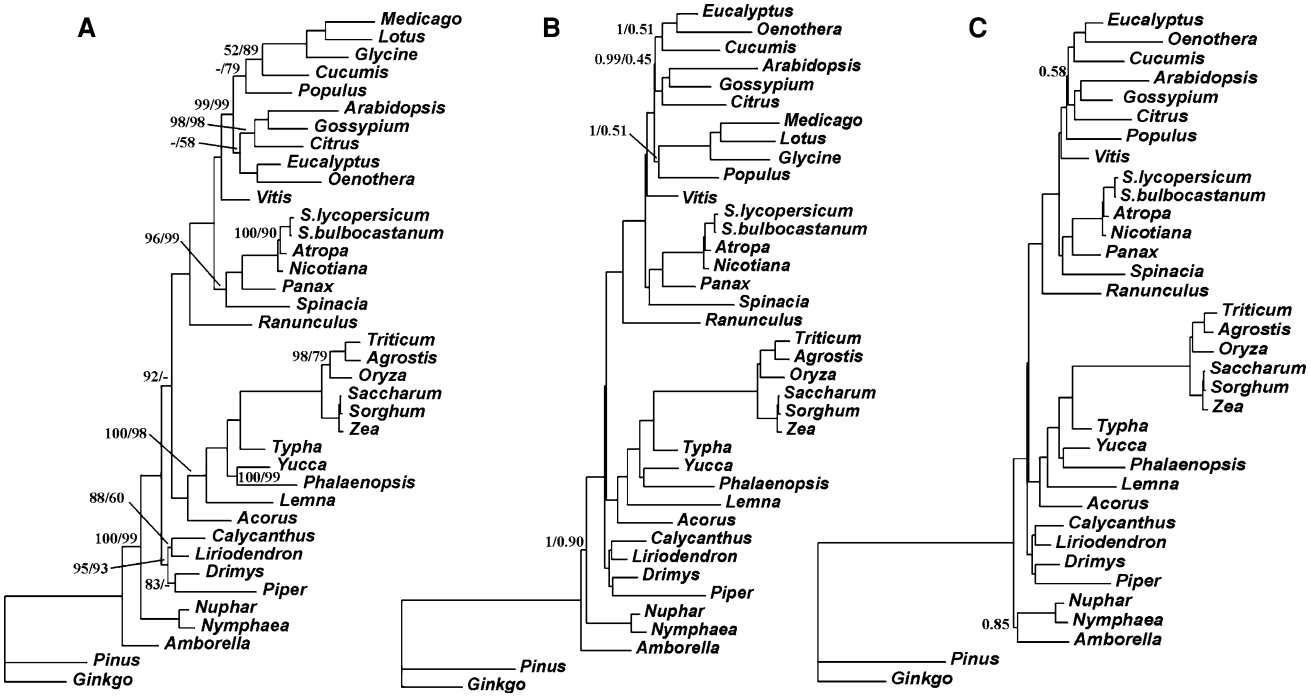

Fig. 2 (**A**) Phylogenetic tree of 61 concatenated plastid protein-coding sequences from 38 taxa, derived from equally weighted maximum parsimony analysis of the full-length data. Numbers at nodes indicate maximum parsimony bootstrap support estimates for full-length matrix analysis and after exclusion of third codon positions, respectively; other branches have 100/100 support. The length of the tree is 71,119 steps, CI = 0.457, and RI = 0.602. (**B**) Bayesian tree derived from the partitioned full-length nucleotide sequences analyses. Numbers at nodes represent posterior probabilities for nucleotide and amino acid data, respectively. Other branches have 1/1 posterior probabilities. (**C**) Bayesian tree derived from the unpartitioned nucleotide matrix analysis after exclusion of third codon positions and representatives of Fabaceae. Numbers at nodes represent posterior probabilities; only values <1 are shown

Myrtaceae (*Eucalyptus, Oenothera*) in the Bayesian trees or nested within eurosids I in the MP tree. Another difference between the trees obtained is the position of *Amborella:* in the Bayesian and MP trees derived from analyses of all codon positions, *Amborella* occupies basal position among angiosperms, whereas the Bayesian phylogeny inference is based on the first two codon positions united *Amborella* with Nymphaeales.

Considering the fact, that alternative placement of *Amborella* differs solely in root placement and two currently available gymnosperm sequences may be too divergent for appropriate rooting of the angiosperm tree, we estimated the angiosperm root using nonstationary substitution model, which does not imply the stationarity condition and does not require an outgroup (Yap and Speed 2005). Site variation in the substitution rates was handled by assigning the sites into three classes. We used a program designed for analysis of nine sequences, therefore the initial data set was reduced to nine plant representatives, which included three basal angiosperms (*Amborella, Nuphar, Nymphaea*), two magnoliids (*Calycanthus, Drimys*), two monocots (*Acorus, Lemna*), and two eudicots (*Vitis, Ranunculus*) and the likelihoods of the competing rooted topologies were compared. In this analysis rooting at a branch leading to *Amborella* had higher likelihood.

To determine if these topologies with alternative placement of *Amborella* can be distinguished using the full-length data, a Shimodaira-Hasegawa (1999) test was conducted and the expected-likelihood weights (Strimmer and Rambaut 2002) were calculated using RELL optimization (Kishino and Hasegawa 1989) as implemented in the Tree-Puzzle program. In none of our analyses a basal position of monocots was recovered, but we tested it too, because in several earlier phylogenomic analyses, under certain conditions, monocots were shown as a sister to all other angiosperms (Goremykin et al. 2003a, 2004, 2005; Chang et al. 2006). According to the results of the tests, the basalmost position of monocots is significantly worse than the optimal topology with *Amborella* as a sister to other angiosperms but a close relationship of *Amborella* and Nymphaeales cannot be rejected ($\Delta \log L = 9.1$, $p = 0.56$, $c = 0.21$).

In general, our inferred chloroplast phylogenies are congruent with recently published molecular trees, in which monophyly of magnoliids, as well as monocots and eudicots, were strongly supported (Cai et al. 2006; Saarela et al. 2007; Jansen et al. 2007).

Structural Rearrangements of the Chloroplast Genome in the Course of Evolution of Monocots

With few exceptions, the gene sets, orders, and nucleotide sequence of chloroplast genomes are highly conserved among land plants (Palmer 1991). The location of borders between the two IR and LSC and the two IR and SSC are known to vary among various cpDNAs (Maier et al. 1995; Goulding et al. 1996; Kim and Lee 2004), even among closely related species (e.g., *Nicotiana tabacum* and *Atropa belladonna* [Kim and Lee 2004]). Generally, these variations are restricted to "movement" of borders within 1 kb, but more considerable expansions and contractions of IR regions are also known. This phenomenon is responsible for wide size variations in chloroplast genomes and apparent inversions (Perry et al. 2002) in different groups of plants. However, such rearrangements are rather rare events and, therefore, are usually viewed as highly reliable markers of common ancestry for the taxa in which they are found (Jansen and Palmer 1987; Raubeson and Jansen 1992). We compared the positions of IR borders in duckweed, tobacco, *Arabidopsis*, basal angiosperm *A. trichopoda*, and several monocots (*P. aphrodite*, *A. calamus*, and grasses represented by rice) aiming to recognize general evolutionary implications from the available chloroplast genomes (Fig. 3).

In both the basal angiosperm *Amborella* and tobacco, the border between LSC and IRb is located between *rps*19 and *rpl*2, and the LSC/IRa border occurs between *rpl*2 and *trn*H. The IRa/SSC border is located in the 3′ region of the *ycf*1 gene and produces the *ycf*1 pseudogene at the IRb/SSC border. In all analyzed monocots other than *Lemna minor*, the IR sequences are expanded within LSC so that IRs encompass *trn*H (*Acorus*), *trn*H and *rps*19 (grasses), or even *trn*H, *rps*19, and the 5′ region of *rpl*22 (*Phalaenopsis)*. This configuration may result from a two-step expansion of IR. At the first step, the IRa/LSC border is moved into LSC, resulting in inclusion of *trn*H into IR. The second step is expansion of IRb into LSC, resulting in incorporation of *rps*19 (*Acorus* and grasses) and a part of *rpl*22 (*Phalaenopsis*) into IR. The result is the appearance of the *trn*H gene between *rps*19 and *rpl*2; this is a structural rearrangement specific for monocot genomes, which therefore must have occurred in the early evolution of monocots.

The proposed rearrangement in the chloroplast genomes of monocots fits the scheme presented in Fig. 3. At first, the above mentioned two-step expansion of IRb results in the inclusion of *trn*H into IR between *rps*19 and *rpl*2 and the movement of the IRb/LSC border inside *rps*19. The resulting structure corresponds to the *A. calamus* chloroplast genome. The further evolution toward modern chloroplast genomes of grasses and *Phalaenopsis* involves, according to this model, the expansion of IRa into SSC in *Acorus*-like genome so that *ycf1* and *rps15* appear within IRs ("intermediate"). Subsequent deletion of *ycf1* in IRs, expansion of IRb into LSC with inclusion of *rps19* into IRs, and a set of specific rearrangements and deletions within

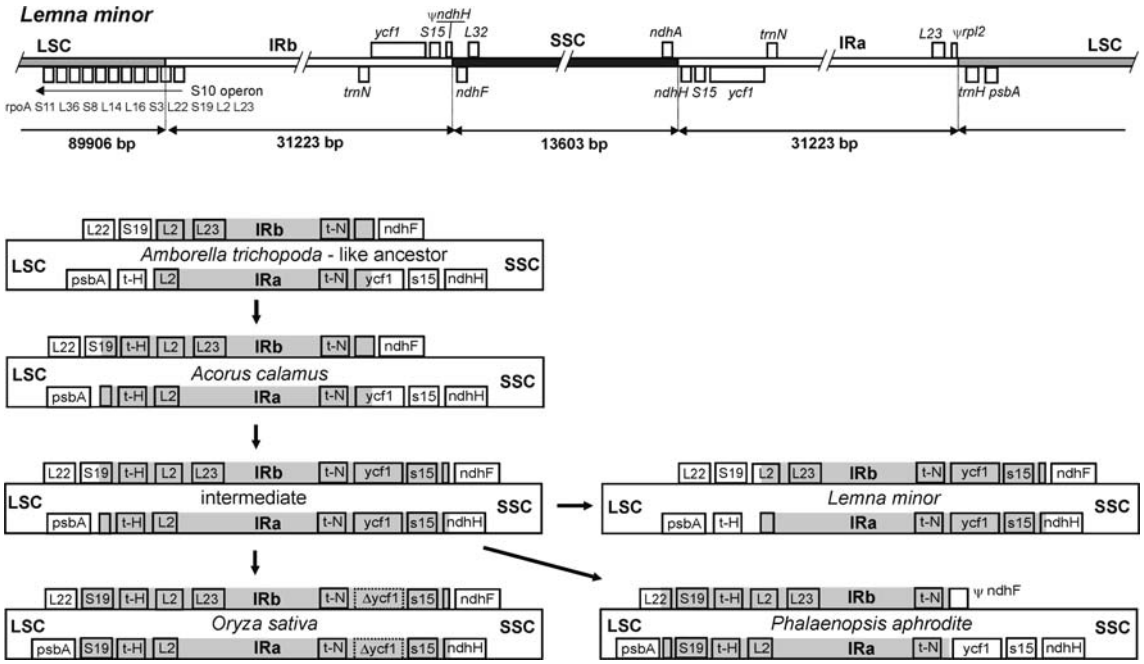

Fig. 3 Comparison of the border positions of LSC, SSC, and IR regions and the proposed model of the evolution of the modern arrangement of cpDNAs in monocots from an *Amborella*-like ancestor. The orientation and order of genes near the borders in *Lemna minor* are shown at the top. Open boxes immediately above and below the main line represent predicted genes that are transcribed rightward and leftward, respectively; pseudogenes at the borders are shown by the ψ(letter. The bottom shows the proposed model of rearrangements. IR regions are shaded. Abbreviations of gene names are as follows: S, *rps*; L, *rpl*; t-N, *trnN*; t-H, *trnH*

LSC produce the chloroplast genome of grasses. The *Phalaenopsis*-like genome could result from two processes: (i) contraction of IRb that would transfer *ycf1* and *rps15* to SSC and (ii) expansion of IRb into the LSC up to *rpl22*. The final step toward *P. aphrodite* genome is shortening of SSC due to the deletion of *ycf2* and several *ndh* genes.

The structure of duckweed chloroplast genome is unique since it combines two features: (i) contrary to other monocots, the LSC/IRb border is located within *rpl2* so that the normal sequence of the S10 operon remains intact, while *trnH* is present only in LSC region; (ii) the IRs are extended into a SSC region to include *ycf1*, *rps15*, and the 5′ part of the *ndhH* gene—a pattern not observed in other completely sequenced cpDNAs of land plants. However, the structure of the *Lemna minor* chloroplast genome may be explained within the above model. One explanation of the formation mechanism of the duckweed chloroplast genome is an extension of IRa into SSC in an *Amborella*-like genome, occurring independently of a similar expansion, which has produced an *Acorus*-like genome. This mechanism seems to be hardly probable since the SSC/IRa borders in *L. minor* and grasses are located almost in the same sites within the *ndh*H gene, which is unlikely to result from evolutionarily independent events. The second, more plausible, explanation is that *L. minor* genome results from contraction of IRb in the "intermediate" genome, resulting in deletion of *trnH* from IRb so that the original *Amborella*-like gene order at the IRa/LSC junction is restored. This model of rearrangements of the chloroplast genomes of monocots fits the phylogenetic relations of monocot plants obtained in this work (Fig. 2).

Acknowledgments The expert technical assistance of Taisia Strakhova is greatly appreciated. This work was supported by the program "Dynamics of Genomes of Plants, Animals and Humans" of the Russian Academy of Sciences.

References

Altschul SF, Gish W, Mille W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

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

Cai Z, Penaflor C, Kuehl JV, Leebens-Mack J, Carlson JE, dePamphilis CW, Boore JL, Jansen RK (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. Mol Biol Evol 23:279–291

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH (2003a) Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. Mol Biol Evol 20:1499–1505

Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH (2003b) The chloroplast genome of the basal angiosperm *Calycanthus floridus*—structural and phylogenetic analyses. Plant Syst Evol 242:119–135

Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH (2004) The chloroplast genome of Nymphaea alba: whole-genome analyses and the problem of identifying the most basal angiosperm. Mol 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. Mol Biol Evol 22:1813–1822

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

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. Mol Gen Genet 217:185–194

Hoch B, Maier RM, Appel K, Igloi GL, Kössel H (1991) Editing of a chloroplast mRNA by creation of an initiation codon. Nature 353:178–180

Jansen RK, Palmer JD (1987) A chloroplast DNA inversion mark an ancient evolutionary split in the sunflower family (Asteraceae). Proc Natl Acad Sci USA 84:5818–5822

Jansen RK, Cai Z, Raubeson LA, Daniell H, dePamphilis CW, Leebens-Mack J, Müller KF, Guisinger-Bellian M, Haberle RC, Hansen AK, Chumley TW, Lee S-B, Peery R, McNeal JR, Kuehl JV, Boore JL (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA 104:19369–19374

Keane TM, Creevey CJ, Pentony MM, Naughton TJ, McInerney JO (2006) Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evol Biol 6:29

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

Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J Mo. Evol 29:170–179

Les HL, Crawford DJ, Landolt E, Gabel JD, Timball RT (2002) Phylogeny and systematics of Lemnaceae, the duckweed family. Syst Bot 27:221–240

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964

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 Mol Biol 251:614–628

Maliga P (2002) Engineering the plastid genome of higher plants. Curr Opin Plant Biol 5:164–172

Mayo SJ, Bogner J, Boyce PC (1997) Genera of Araceae. Royal Botanic Gardens, Kew, UK

Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, Heggie L, Kavanagh TA, Hibberd JM, Gray JC, Morden CW, Calie PJ, Jermiin LS, Wolfe KH. (2001) Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. Plant Cell 13:645–658

Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight DNA. Nucleic Acids Res 8:4321–4325

Neckermann K, Zeltz P, Igloi GL, Kossel H, Maier RM (1994) The role of RNA editing in conservation of start codons in chloroplast genomes. Gene 146:177–182

Ogihara Y, Isono K, Kojima T, Endo A, Hanaoka M, Shiina T, Terachi T, Utsugi S, Murata M, Mori N, Takumi S, Ikeo 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. Mol Genet Genomics 266:740–746

Palmer JD (1985) Comparative organization of chloroplast genomes. Annu Rev Genet 19:325–354

Palmer JD (1991) Plastid chromosome: structure and evolution. In: Hermann RG (ed) The molecular biology of plastids. Cell culture and somatic cell genetics of plants. Springer-Verlag, Vienna, pp 5–53

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, Brennan S, Murphy DJ, Kavanagh TA, Wolfe KH (2002) Evolutionary re-organisation of a large operon in adzuki bean chloroplast DNA caused by inverted repeat movement. DNA Res 9:157–162

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817–818

Raubeson LA, Jansen RK (1992) Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. Science 255:1697–1699

Raubeson LA, Jansen RK (2005) Chloroplast genomes of plants. In: Henry RJ (ed) Plant diversity and evolution: genotypic and phenotypic variation in higher plants. CAB International, Wallingford, UK, pp 45–68

Ronquist F, Huelsenbeck JP (2003) MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574

Rothwell GW, Van Atta MR, Ballard HE Jr, Stockey RA (2004) Molecular phylogenetic relationships among Lemnaceae and Araceae using the chloroplast *trnL-trnF* intergenic spacer. Mol Phylogenet Evol 30:378–385

Saarela FM, Rai HS, Doyle JA, Endress PK, Mathews S, Marchant AD, Briggs BG, Graham SW (2007) Hydatellaceae identified as a new branch near the base of the angiosperm phylogenetic tree. Nature 446:312–315

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504

Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 16:1114–1116

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 tobacco chloroplast genome: its gene organisation and expression, EMBO J 5:2043–2049

Soltis PS, Soltis DE, Chase MW (1999) Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature 402:402–404

Stoebe B, Martin W, Kowallik KV (1998) Distribution and nomenclature of protein-coding genes in 12 sequenced chloroplast genomes. Plant Mol Biol Rep 16:243–255

Strimmer K, Rambaut A (2002) Inferring confidence sets of possibly misspecified gene trees. Proc Biol Sci 269:137–142

Sugiura M (1992) The chloroplast genome. Plant Mol Biol 19:149–168

Sugiura M, Shinozaki K, Tanaka M, Hayashida N, Wakasugi T, Matsubayashi T, Ohto C, Torazawa K, Meng BY, Hidaka T, Zaita N (1987) Split genes and cis/trans splicing in tobacco chloroplasts. In: von Wettstein D, Chua N-H (eds) Plant molecular biology. Plenum Press, New York, pp 65–76

Swofford DL (2003) PAUP*: Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, MA

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 24:4876–4882

Wakasugi T, Sugita M, Tsudzuki T, Sugiura M (2001) The genomics of land plant chloroplasts: gene content and alteration of genomic information by RNA editing. Photosyn Res 70:107–118

Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252–3255

Yamamoto YT, Rajbhandari N, Lin X, Bergmann BA, Nishimura Y, Stomp AM (2001) Genetic transformation of duckweed *Lemna gibba* and *Lemna minor*. In Vitro Cell Dev Biol Plant 37:349–353

Yap VB, Speed T (2005) Rooting a phylogenetic tree with nonreversible substitution models. BMC Evol Biol 5:2