DNA Research **12**, 215–220 (2005)

Short Communication
doi:10.1093/dnares/dsi006

Complete Nucleotide Sequence of the Chloroplast Genome from the Tasmanian Blue Gum, *Eucalyptus globulus* (Myrtaceae)

Dorothy A. Steane*

Cooperative Research Centre for Sustainable Production Forestry, School of Plant Science, University of Tasmania, Private Bag 55, Hobart, Tasmania 7001, Australia

(Received 22 November 2004; revised 4 April 2005)

Abstract

The complete nucleotide sequence of the chloroplast genome of the hardwood species *Eucalyptus globulus* is presented and compared with chloroplast genomes of tree and non-tree angiosperms and two softwood tree species. The 160 286 bp genome is similar in gene order to that of *Nicotiana*, with an inverted repeat (IR) (26 393 bp) separated by a large single copy (LSC) region of 89 012 bp and a small single copy region of 18 488 bp. There are 128 genes (112 individual gene species and 16 genes duplicated in the inverted repeat) coding for 30 transfer RNAs, 4 ribosomal RNAs and 78 proteins. One pseudogene (*ψ-inf*A) and one pseudo-ycf (*ψ-ycf*15) were identified. The chloroplast genome of *E. globulus* is essentially co-linear with that of another hardwood tree species, *Populus trichocarpa*, except that the latter lacks *rps*16 and *rpl*32, and the IR has expanded in *Populus* to include *rps*19 (part of the LSC in *E. globulus*). Since the chloroplast genome of *E. globulus* is not significantly different from other tree and non-tree angiosperm taxa, a comparison of hardwood and softwood chloroplasts becomes, in essence, a comparison of angiosperm and gymnosperm chloroplasts. When compared with *E. globulus*, *Pinus* chloroplasts have a very small IR, two extra tRNAs and four additional photosynthetic genes, lack any functional *ndh* genes and have a significantly different genome arrangement. There does not appear to be any correlation between plant habit and chloroplast genome composition and arrangement.

Key words: eucalypt; Myrtaceae; chloroplast DNA; pseudogene; gymnosperm

Eucalyptus globulus is one of the most economically important species for hardwood forestry plantations in temperate regions of the world.[1] It has been studied intensively by quantitative, population and evolutionary geneticists and is becoming a model species for genetic research in *Eucalyptus*. Chloroplast DNA has been essential to many studies of population genetics and phylogeography in *Eucalyptus*. This paper presents the complete chloroplast genome from *E. globulus* and compares it with chloroplast genomes from other angiosperm taxa [including the hardwood tree species, *Populus trichocarpa* (B. Heinz, S. DiFazio, K. Ritland et al., manuscript in preparation)] and softwood tree species (*Pinus thunbergii*[2] and *Pinus koraiensis*).

The complete chloroplast genome of *E. globulus* (GenBank accession no. AY780259) may be represented as a circular chromosome (Fig. 1), although this is likely to be a rare form of the molecule, as most chloroplast DNA is, in fact, linear.[3,4] Comprising 160 286 bp, it ranks among the larger land plant chloroplast genomes. Most land plant plastids sequenced to date have genomes of 116–163 kb, and the longest belongs to *Oenothera elata* (163 935 bp[5]). The structure of the *E. globulus* chloroplast genome is typical of most plastids: a large single copy (LSC) region (89 012 bp) and a small single copy (SSC) region (18 488 bp) are separated by an inverted repeat (IR) (26 393 bp). The relative sizes of the LSC, SSC and IR regions remain reasonably constant across genomes of angiosperms (approximately 55, 12 and 16.5% of the total genome size, respectively), regardless of the overall size of the genome. The relative size of the IR in gymnosperms varies much more. For example, in *Ginkgo biloba* the IR is 17 kb, but in *P. thunbergii* it is just 495 bp[2] containing *trn*I-CAU and 83 bp from the 3′ end of *psb*A, but lacking the ribosomal RNA genes that characterize other land plant IRs.

Communicated by Katsumi Isono

* Tel. +61-3-62261828, Fax. +61-3-62262698, E-mail: dorothy.steane@utas.edu.au

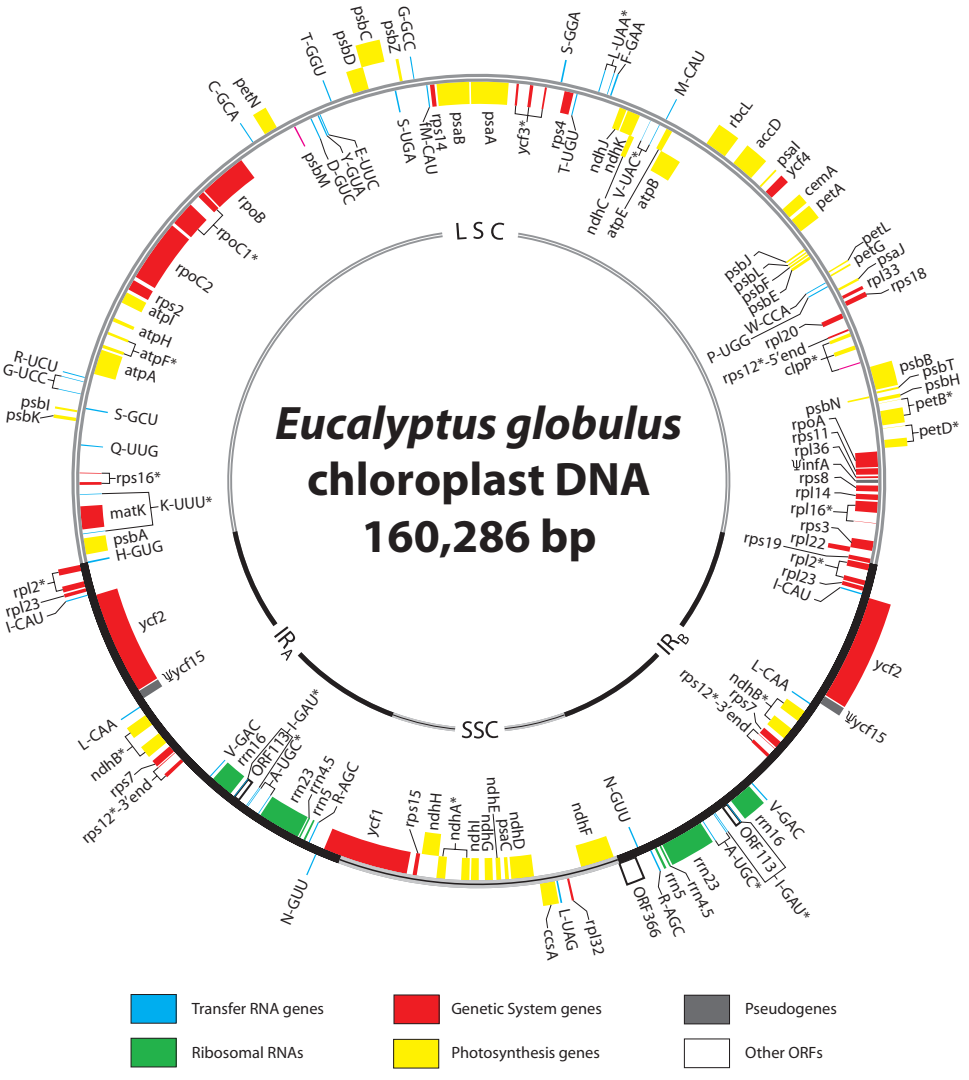

Figure 1. Gene map of the plastid chromosome of *Eucalyptus globulus*. Genes belonging to different functional groups are color coded (see key). Genes drawn inside the circle are transcribed clockwise; those outside the circle are transcribed anti-clockwise. In cases where two genes overlap, one of them is shifted off the map to show its position. Asterisks indicate genes that contain introns. Pseudogenes are marked by ψ. ORF366 in IR_B is a truncated form of *ycf*1.

The *Eucalyptus* chloroplast genome has a GC-content of 36.9%, which is comparable with that of other vascular plant plastids (e.g. 36.7% in *Populus*, 37.8% in *Nicotiana*, 38.4% in *Zea*, 39.2% in *Oenothera* and 38.5% in *P. thunbergii*). The genome is AT-rich in both the non-coding intergenic regions (67% AT) and the coding regions (62% AT), where there is an AT bias (73% ± 4.5%) in the third base positions of all amino acid codons. This phenomenon is also observed in other plastid genomes. In contrast, the tRNA genes show less of an AT bias (58%), and the rRNA genes have a slight GC bias (55%). The latter is characteristic of rRNA genes in other plants.[6]

Table 1 lists all genes detected in the chloroplast genome of *E. globulus*. The start codons for the protein-coding genes were based on alignments with plastomes of other taxa. The start positions are, therefore, hypothetical and can be confirmed only through analysis of gene transcripts. The genome is essentially co-linear with that of the annual angiosperm *Nicotiana tabacum* (Fig. 2a), with all the same genes except *spr*A, which is absent from *E. globulus*. The chloroplast genome of *E. globulus* is also virtually co-linear with that from another hardwood tree species, *P. trichocarpa* (Fig. 2b), except for three notable differences: (i) *rps*16 and flanking intergenic sequences (~1800 bp of LSC) are missing from *Populus*; (ii) the gene *rpl*32 and flanking sequences (~1100 bp of SSC) are absent from *Populus*; and (iii) the IR in *Populus* has expanded to include *rps*19, with the 'extra' copy of this gene located close to J_{LA}. As in other angiosperms, the *E. globulus* plastome has four ribosomal RNA (rRNA) genes and 30 transfer RNA (tRNA) genes (of which seven are located in the IRs) that provide tRNAs for all 20 amino acids (Table 1). There are 78 protein-coding

Table 1. List of genes found in *Eucalyptus globulus* chloroplast genome (GenBank accession no. AY780259; herbarium accession no. HO528199)[a].

RNA genes						
Transfer RNAs	*trn*A-UGC[b,c]	*trn*C-GCA	*trn*D-GUC	*trn*E-UUC	*trn*F-GAA	
	*trn*fM-CAU	*trn*G-GCC	*trn*G-UCC[b]	*trn*H-GUG	*trn*I-CAU	
	*trn*I-GAU[b,c]	*trn*K-UUU[b]	*trn*L-CAA[c]	*trn*L-UAA[b]	*trn*L-UAG	
	*trn*M-CAU	*trn*N-GUU[c]	*trn*P-UGG	*trn*Q-UUG	*trn*R-ACG[c]	
	*trn*R-UCU	*trn*S-GCU	*trn*S-GGA	*trn*S-UGA	*trn*T-GGU	
	*trn*T-UGU	*trn*V-GAC[c]	*trn*V-UAC[b]	*trn*W-CCA	*trn*Y-GUA	
Ribosomal RNAs	*rrn*16S	*rrn*23S[c]	*rrn*4.5S[c]	*rrn*5S[c]		
Genetic system genes						
Conserved ORFs[d]	*ycf*1	*ycf*2[c]	*ycf*3[e]	*ycf*4		
Intron maturase	*mat*K					
RNA polymerase	*rpo*A	*rpo*B	*rpo*C1[b]	*rpo*C2		
Ribosomal proteins						
Large subunit	*rpl*14	*rpl*16[b]	*rpl*2[b,c]	*rpl*20	*rpl*22	
	*rpl*23[c]	*rpl*32	*rpl*33	*rpl*36		
Small subunit	*rps*11	*rps*12[e,f]	*rps*14	*rps*15	*rps*16[b]	*rps*18
	*rps*19	*rps*2	*rps*3	*rps*4	*rps*7[c]	*rps*8
Photosynthesis genes						
Acetyl-CoA carboxylase	*acc*D					
ATP-dependent protease	*clp*P[e]					
ATP synthase	*atp*A	*atp*B	*atp*E	*atp*F[b]	*atp*H	*atp*I
Cytochrome *b/f*	*pet*A	*pet*B[b]	*pet*D[b]	*pet*G	*pet*L	*pet*N
Cytochrome *c* biogenesis	*ccs*A					
Membrane protein	*cem*A					
NADH dehydrogenase	*ndh*A[b]	*ndh*B[b,c]	*ndh*C	*ndh*D	*ndh*E	*ndh*F
	*ndh*G	*ndh*H	*ndh*I	*ndh*J	*ndh*K	
Photosystem I	*psa*A	*psa*B	*psa*C	*psa*I	*psa*J	
Photosystem II	*psb*A	*psb*B	*psb*C	*psb*D	*psb*E	*psb*F
	*psb*H	*psb*I	*psb*J	*psb*K	*psb*L	*psb*M
	*psb*N	*psb*T	*psb*Z			
Rubisco	*rbc*L					
Open reading frames	ORF113[c]	ORF366[g]				
Pseudogenes	Pseudo-*inf*A	Pseudo-*ycf*15[c]				

[a] Chloroplasts were isolated using sucrose gradients[20,21] and cpDNA was extracted using a modified CTAB method.[22] The DNA was randomly sheared, producing fragments of 2–4 kb. Fragments were ligated into pSMART-LC vector (Lucigen) and were inserted into One Shot GeneHogs Electrocompetent *Escherichia coli* (Invitrogen). Recombinant clones were sequenced using Big Dye Terminator v.3.1 chemistry and an ABI 3730 xl capillary sequencer. Sequences were quality scored using Phred software[23] and contigs were assembled using Phrap software (see http://www.phrap.org/). The 22 resulting contigs were aligned with conserved regions of the tobacco chloroplast genome[24] to gain an estimate of gene order. PCR primers were designed to fill the gaps, and the PCR amplified fragments were sequenced in both directions on a CEQ 8000 Genetic Analysis System (Beckman Coulter). Sequence assembly was carried out using Sequencher 3.1 (Gene Codes Corporation, USA) and Sequence Navigator 1.0.1 (Applied Biosystems, Inc., USA). From two to six times coverage was obtained for all regions of the chloroplast genome. The software package DOGMA[15] was used to locate putative genes. The coordinates and composition of genes were checked against GenBank by using BLASTX and BLASTN. Genomic analyses were conducted using the Biomanager suite of programs available from ANGIS (Australian National Genome Information Service).

[b] Gene containing one intron.

[c] Two gene copies due to IR.

[d] *ycf*1 and *ycf*2 are known to be essential chloroplast genes, although their exact functions remain unclear; *ycf* 3 and *ycf*4 are hypothesized to be involved in biogenesis of photosystem I (see text).

[e] Gene containing two introns.

[f] Divided gene.

[g] ORF366 occurs in IR_B and is a truncated form of *ycf*1 that spans IR_A and SSC.

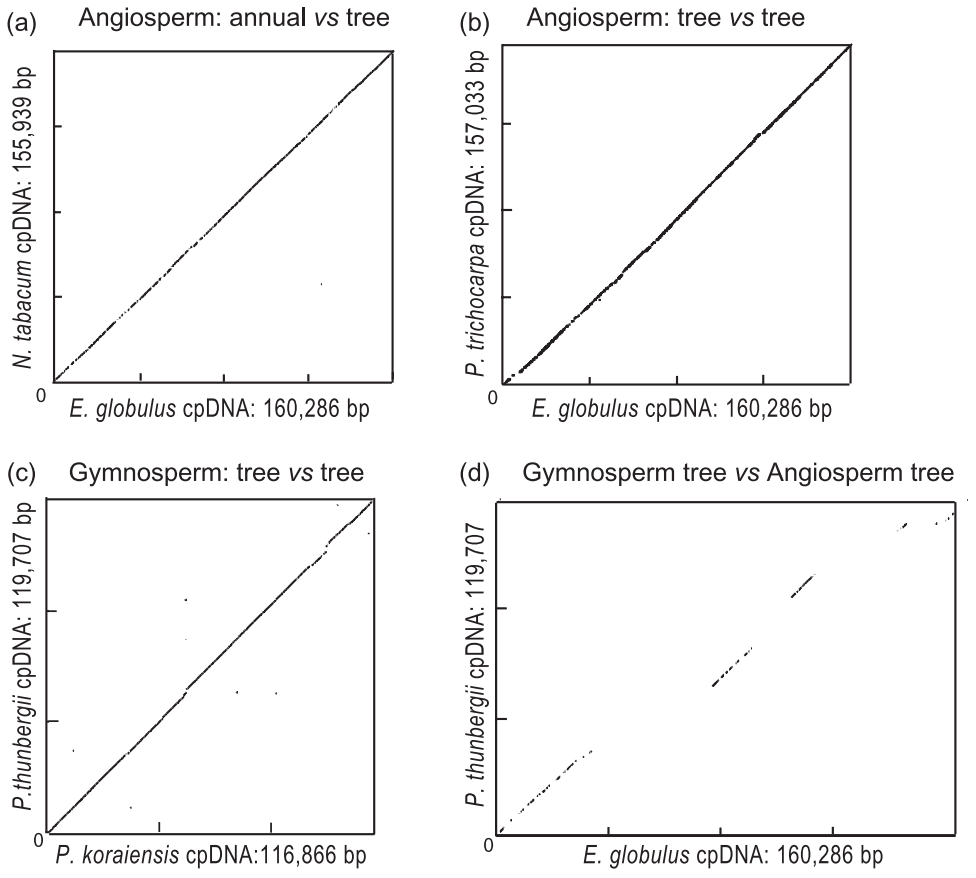

Figure 2. Harr plot analysis comparing chloroplast genomes from an annual angiosperm, hardwood (angiosperm) trees and softwood (gymnosperm) trees: a) *Nicotiana tabacum* and *Eucalyptus globulus*; b) *E. globulus* and *Populus trichocarpa*; c) *Pinus koraiensis* and *Pinus thunbergii*; and d) *E. globulus* and *P. thunbergii*. Plots were constructed using COMPARE (GCG) and DOTPLOT (GCG). Each dot represents a position where 45 out of 50 nucleotides match in both sequences. All genomes are available from GenBank, except for that of *Populus*, which can be viewed on-line (http://genome.ornl.gov/poplar_chloroplast/).

genes, including four conserved open reading frames (ORFs) ('ycfs'). Approximately 74 protein-coding genes are common to most angiosperm chloroplast genomes, and an additional 5 are present in only some species.[7] Of these five, four (*acc*D, *ycf*1, *ycf*2 and *rpl*23) appear to be functional in the plastome of *E. globulus*, but the fifth, *inf*A, is a pseudogene (ψ), as in *Populus*, *Nicotiana*, *Arabidopsis* and *Oenothera*.[7] One other pseudogene was detected, that of a hypothetical chloroplast protein, ψ*ycf*15. One open reading frame, ORF113, has high homology to regions of *ycf*68 in rice, maize and *Pinus*, as well as to hypothetical proteins ORF119 and ORF58 in the *trn*I intron of *Oenothera*. A second open reading frame, ORF366, is found in IR_B at the junction with the SSC. It is a truncated inverted repeat of *ycf*1 and is probably non-functional.

There are three classes of ORFs in plastid DNA: (i) genes of known function; (ii) hypothetical chloroplast reading frames (ycfs) that are highly conserved between species; and (iii) species-specific or rapidly diverging ORFs. Four major ycfs have been partially characterized, but their precise functions are not yet understood. Two highly conserved ycfs, *ycf*1 and *ycf*2, have been demonstrated to be essential to cellular function in dicots;[8] they are not involved in photosynthesis, but are speculated to be involved in cellular metabolism or to have a structural role in the plastid.[8] Two more ycfs, *ycf* 3 and *ycf* 4, are believed to be involved in the formation of photosystem I.[9,10] The functionality of some other ycfs, however, has been brought into question by the relatively frequent occurrence of pseudo-ycf loci. For example, although *ycf*15 in tobacco appears to be a potentially functional protein-coding gene, in many other species—including *E. globulus*—a variable insertion of ~250 bp (295 bp in *E. globulus*) introduces premature stop codons. Schmitz-Linneweber *et al.*[11] showed that although the *ycf*15 cistron may be transcribed, splicing of the two conserved ends does not occur; hence, *ycf*15 is probably not a protein-coding gene. The *ycf*15 sequences of *E. globulus* and *Oenothera* are very similar after the removal of their insertions. However, both with and without the intervening sequence, *ycf*15 of both taxa have premature stop codons, providing further evidence that *ycf*15 is probably not a functional protein-coding gene. Another example of a ycf that has highly conserved domains, but often is not completely conserved, is

*ycf*68. In *E. globulus*, ORF113 is highly homologous to a small region of *ycf*68 in rice and maize, ORF75 in *P. koraiensis*, ORF75a in *P. thunbergii* and a hypothetical protein in *O. elata* (ORF58). All these ORFs have some homology to *ycf*68. Such ORFs and ycfs that have some highly conserved regions may have roles in gene regulation (e.g. as promotor or terminator sequences) or may be genes specifying a structural RNA[11] (as was at first proposed for *spr*A in tobacco chloroplasts,[12] but was later discounted[13]).

The *psb*L gene that codes for a 38 amino acid peptide of photosystem II is highly conserved among many higher plants. This gene is unusual because in *Eucalyptus*, as well as in some other taxa (e.g. *Nicotiana* and *Spinacia*, but not *Populus*), transcription of the gene does not require any of the standard chloroplast initiation codons [i.e. leucine (TTG, CTG), isoleucine (ATT, ATC, ATA), valine (GTG) or, the most common, methionine (ATG)]. Instead, ACG appears at the beginning of the gene. It has been shown in *Nicotiana* that a translatable *psb*L mRNA containing an AUG initiator codon is formed by C to U editing of the ACG codon,[14] and it is possible that a similar mechanism exists in *Eucalyptus*.

In general, the chloroplast genome of *E. globulus* is not significantly different from most other angiosperms, so a comparison of hardwood and softwood chloroplasts becomes, in essence, a comparison of angiosperm and gymnosperm chloroplasts. Chloroplast DNA sequences are available for two gymnosperms, *P. thunbergii* (119 707 bp) and *P. koraiensis* (116 866 bp). Both genomes are significantly smaller than those of most angiosperms sequenced so far. Pairwise comparisons using Harr plots (Fig. 2c) and DOGMA software[15] (data not shown) show that the chloroplast DNA sequences of the two pine species are very similar. In contrast, those same analytical techniques indicate that the chloroplast genomes of *P. thunbergii* and *E. globulus* are arranged very differently (Fig. 2d). Relative to *Eucalyptus*, *rbc*L and its neighboring regions in the LSC region are inverted in the pines, and a large region from the LSC, including *psa*A and *psa*B, occurs in the SSC.[2] The rRNA genes from *rrn*16 to *trn*R-AGC that are in the inverted repeat in angiosperms form a cluster in the middle of the SSC in *P. thunbergii*.[2] In addition to the 30 tRNA genes found in angiosperms, the two pine species have two unusual tRNAs, *trn*P-GGG and *trn*R-CCG. The first of these is also found in hornworts[16] and ferns[17], and *trn*R-CCG has been found in moss, although it is not essential for plastid function in moss and may not be a functional gene.[18] Angiosperms and pines have the same suite of ribosomal protein genes, except that the pines lack *rps*16. Pines have an intact *inf*A gene, in contrast to the pseudogene found in *Eucalyptus* and many other angiosperms (see above). In addition to the 29 genes encoding components of the photosynthetic apparatus in angiosperms, pines have 4 more genes that exist in some lower plants: *psa*M, *chl*B, *chl*L and *chl*N. The *psa*M gene (which is duplicated in the LSC of *P. thunbergii*,[2] but not in *P. koraiensis*) has been found in non-vascular plants, but is absent from ferns and angiosperms, suggesting parallel losses in the latter two groups during their evolution.[17] The genes *chl*B, *chl*L and *chl*N may be associated with the ability of pines to synthesize chlorophyll in the dark (as in *Chlamydomonas*[19]). A major difference in the gene content between pines and angiosperms is the complete absence of functional *ndh* genes from pine chloroplasts.[2] It is unclear whether chloroplast *ndh* genes have been transferred to the nuclear genome of pines, or whether pine chloroplasts lack an NADH dehydrogenase altogether. *Eucalyptus* and *Nicotiana* have 21 introns, 5 more than *P. thunbergii* and *P. koraiensis*. Of these five, three occur in genes that are absent from pines (*rps*16, *ndh*A and *ndh*B), and two occur in *clp*P that, in pines, has no introns. The 16 remaining split genes are conserved between pines and angiosperms.[2]

In conclusion, there does not appear to be any correlation between plant habit and plastome composition and arrangement. Differences between chloroplast genomes of tree and non-tree angiosperm species are slight. In contrast, although angiosperm and gymnosperm chloroplasts share many genes, there are significant differences in genome size, arrangement and gene content.

Acknowledgements: The author thanks Peter Wilson and other staff at the Australian Genome Research Facility (AGRF); Natalie Papworth and Alan McFadden (Royal Tasmanian Botanical Garden); Peter Boyer (SouthWind Writing and Publishing Services, Tasmania); Bob Elliott, Adam Smolenski, Natalie Conod, Rebecca Jones, Catherine Phillips, Briony Patterson, Gay McKinnon, Brad Potts and René Vaillancourt (University of Tasmania). This research was funded by the Cooperative Research Centre for Sustainable Production Forestry (CRC-SPF).

References

1. Eldridge, K. G., Davidson, J., Harwood, C., and van Wyk, G. 1993, Eucalypt Domestication and Breeding, Clarendon Press, Oxford.
2. Wakasugi, T., Tsudzuki, J., Ito, S., Nakashima, K., Tsudzuki, T., and 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.
3. Oldenburg, D. J. and Bendich, A. J. 2004, Most chloroplast DNA of maize seedlings in linear molecules with defined ends and branched forms, *J. Mol. Biol.*, **335**, 953–970.
4. Bendich, A. J. 2004, Circular chloroplast chromosomes: the grand illusion, *Plant Cell*, **16**, 1661–1666.
5. Hupfer, H., Swiatek, M., Hornung, S., et al. 2000, Complete nucleotide sequence of the *Oenothera elata* plastid chromosome, representing plastome I of the five

distinguishable Euoenothera plastomes, *Mol. Gen. Genet.*, **263**, 581–585.
6. Goremykin, V. V., Hirsch-Ernst, K. I., Wolfl, S., and Hellwig, F. H. 2003, Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm, *Mol. Biol. Evol.*, **20**, 1499–1505.
7. Millen, R. S., Olmstead, R. G., Adams, K. L., et al. 2001, Many parallel losses of *inf*A from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus, *Plant Cell*, **13**, 645–658.
8. Drescher, A., Ruf, S., Calsa, T., Carrer, H., and Bock, R. 2000, The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes, *Plant J.*, **22**, 97–104.
9. Boudreau, E., Takahashi, Y., Lemieux, C., Turmel, M., and Rochaix, J. D. 1997, The chloroplast *ycf*3 and *ycf*4 open reading frames of *Chlamydomonas reinhardtii* are required for the accumulation of the photosystem I complex, *EMBO J.*, **16**, 6095–6104.
10. Ruf, S., Kossel, H., and Bock, R. 1997, Targeted inactivation of a tobacco intron-containing open reading frame reveals a novel chloroplast-encoded photosystem I-related gene, *J. Cell Biol.*, **139**, 95–102.
11. Schmitz-Linneweber, C., Maier, R. M., Alcaraz, J. P., Cottet, A., Herrmann, R. G., and Mache, R. 2001, The plastid chromosome of spinach (*Spinacia oleracea*): complete nucleotide sequence and gene organization, *Plant Mol. Biol.*, **45**, 307–315.
12. Vera, A. and Sugiura, M. 1994, A novel RNA gene in the tobacco plastid genome: its possible role in the maturation of 16S ribosomal RNA, *EMBO J*, **13**, 2211–2217.
13. Sugita, M., Svab, Z., Maliga, P., and Sugiura, M. 1997, Targeted deletion of *spr*A from the tobacco plastid genome indicates that the encoded small RNA is not essential for pre-16S rRNA maturation in plastids, *Mol. Gen. Genet.*, **257**, 23–27.
14. Kudla, J., Igloi, G., Metzlaff, M., Hagemann, R., and Kossel, H. 1992, RNA editing in tobacco chloroplasts leads to the formation of a translatable *psb*L mRNA by a C to U substitution within the initiation codon, *EMBO J.*, **11**, 1099–1103.
15. Wyman, S. K., Jansen, R. K., and Boore, J. L. 2004, Automatic annotation of organellar genomes with DOGMA., *Bioinformatics*, **20**, 3252–3255.
16. Kugita, M., Kaneko, A., Yamamoto, Y., Takeya, Y., Matsumoto, T., and Yoshinaga, K. 2003, The complete nucleotide sequence of the hornwort (*Anthoceros formosae*) chloroplast genome: insight into the earliest land plants, *Nucleic Acids Res.*, **31**, 716–721.
17. Wolf, P. G., Rowe, C. A., Sinclair, R. B., and Hasebe, M. 2003, Complete nucleotide sequence of the chloroplast genome from a leptosporangiate fern, *Adiantum capillus-veneris* L., *DNA Res.*, **10**, 59–65.
18. Sugiura, C. and Sugita, M. 2004, Plastid transformation reveals that moss tRNA(Arg)-CCG is not essential for plastid function, *Plant J.*, **40**, 314–321.
19. Liu, X. Q., Xu, H., and Huang, C. Z. 1993, Chloroplast *chl*B gene is required for light-independent chlorophyll accumulation in *Chlamydomonas reinhardtii*, *Plant Mol. Biol.*, **23**, 297–308.
20. Palmer, J. D. 1986, In Weissbach, A. and Weissbach, H. (eds) Methods in Enzymology. Academic Press, New York, pp. 167–186.
21. Steane, D. A., West, A. K., Potts, B. M., Ovenden, J. R., and Reid, J. B. 1991, Restriction fragment length polymorphisms in chloroplast DNA from six species of *Eucalyptus*, *Aust. J. Bot.*, **39**, 399–414.
22. Doyle, J. J. and Doyle, J. L. 1990, Isolation of plant DNA from fresh tissue, *Focus*, **12**, 13–15.
23. Ewing, B. and Green, P. 1998, Base-calling of automated sequencer traces using Phred. II. Error probabilities, *Genome Res.*, **8**, 186–194.
24. Wakasugi, T., Sugita, M., Tsudzuki, T., and Sugiura, M. 1998, Updated gene map of tobacco chloroplast DNA, *Plant Mol. Biol. Rep.*, **16**, 231–241.