Complete chloroplast DNA sequence of the moss *Physcomitrella patens*: evidence for the loss and relocation of *rpoA* from the chloroplast to the nucleus

Chika Sugiura, Yuki Kobayashi, Setsuyuki Aoki¹, Chieko Sugita and Mamoru Sugita*

Received June 12, 2003; Revised and Accepted July 21, 2003 DDBJ/EMBL/GenBank accession nos*

**ABSTRACT**

The complete chloroplast DNA sequence (122 890 bp) of the moss *Physcomitrella patens* has been determined. The genome contains 83 protein, 31 tRNA and four rRNA genes, and a pseudogene. Four protein genes (*rpoA*, *cysA*, *cysT* and *ccsA*) found in the liverwort *Marchantia polymorpha* and the hornwort *Anthoceros formosae* are absent from *P. patens*. The overall structure of *P. patens* chloroplast DNA (cpDNA) differs substantially from that of liverwort and hornwort. Compared with its close relatives, a 71 kb region from *petD* to *rpoB* of *P. patens* is inverted. To investigate whether this large inversion and the loss of *rpoA* usually occur in moss plants, we analyzed amplified cpDNA fragments from four moss species. Our data indicate that the large inversion occurs only in *P. patens*, whereas the loss of the *rpoA* gene occurs in all mosses. Moreover, we have isolated and characterized the nuclear *rpoA* gene encoding the α subunit of RNA polymerase (RNAP) from *P. patens* and examined its subcellular localization. When fused to green fluorescent protein, RpoA was observed in the chloroplasts of live moss protonemata cells. This indicates that chloroplast RNAP is encoded separately by chloroplast and nuclear genomes in the moss. These data provide new insights into the regulation and evolution of chloroplast transcription.

**INTRODUCTION**

Chloroplasts evolved from cyanobacteria through endosymbiosis, and possess independent genomes (1). To date, complete chloroplast DNA (cpDNA) sequences from two dozen land plants and algae have been determined (2,3). Algal cpDNAs range from 89 (*Codium fragile*) to over 1500 kb (*Acetabularia*), and their gene content and organization vary from species to species (2). In contrast, land plant cpDNAs are relatively uniform in size, from 120 to 160 kb, and their gene content and organization are well conserved (3). Bryophytes consist of three classes (liverworts, hornworts and mosses) and have been identified as the earliest land plants (4–6). Among the bryophytes, the cpDNA (121 024 bp) of the liverwort *Marchantia polymorpha* was the first to be sequenced (7), and the hornwort *Anthoceros formosae* cpDNA sequence (161 162 bp) has recently been published (8). The larger size of the cpDNA genome of *A. formosae* results from a larger inverted repeat (IR) sequence compared with the liverwort *M. polymorpha*. The overall gene organization is highly conserved between the two bryophytes. Single introns, however, are inserted into the 23S rRNA genes (*rrn23*) of *A. formosae* (8) but not into those of *M. polymorpha* (7), the charophyte *Chaetosphaeridium globosum* (9) or vascular plants (3). Over 900 RNA editing sites have been identified in the hornwort *A. formosae* chloroplast (10). In contrast, no RNA editing site has been found in the cpDNA of the liverwort *M. polymorpha* (11). Therefore, RNA editing events differ substantially between bryophyte species compared with the high degree of conservation of editing sites in the chloroplasts of vascular plants (12). Significant divergence in RNA editing events among the bryophytes supports bryophyte paraphyly, as suggested by molecular phylogenetic studies (13–15).

The remaining bryophyte group, the mosses, has been reported to be most closely related to the early lycophod land plants, the ferns, and the vascular plant lineage (4–6,15). The exact phylogenetic status of this group, however, remains unresolved. To answer this question, molecular data on moss chloroplast genes are required. There is little information on the gene structure (16–18) or RNA editing (19) of the moss *P. patens*. Therefore, we have determined the complete nucleotide sequence of the *P. patens* cpDNA. Moreover, we
have isolated a nuclear rpoA gene encoding an α subunit of RNA polymerase (RNAP) and discuss the relocation of rpoA from the chloroplast to the nucleus.

MATERIALS AND METHODS

Accession numbers
The sequences reported in this paper have been deposited in the DDBJ/GenBank/EMBL databases with accession numbers: P. patens cpDNA, AP005672; cpDNA fragments amplified from four mosses, AB098724–AB098727; PpRpoA gene, AB110071; PpRpoA cDNA, AB110072.

Plant materials
The moss P. patens subspecies patens was grown at 25°C under continuous illumination. Total cellular DNA and RNA were isolated from P. patens protistoma, as described previously (20).

Preparation of cpDNA fragments
A P. patens genomic library was screened using the tobacco ndhD sequence (21) as a probe, as described previously (22), and the genomic clone A-D5 was isolated. The λ-D5 insert DNA was amplified by polymerase chain reaction (PCR) using primers designed from AfiI-II vector sequences, AGCTCTAACTACACTATAGGGCGTCGA and GAGCTCAATTAACCTCCTAAAGGGAGTCGA. The reaction of long and accurate (LA) PCR was performed with 30 cycles of 10 s denaturation at 98°C and 15 min annealing and extension at 68°C using LA Taq DNA polymerase (Takara Shuzo). The amplified DNA contained a 12 640 bp region from rbcL to psbD, 13 774 bp DNA region from TACTAACTTATTTACTTCTATTG, AB066297); for a region from CAACAAACCCTA, AB066297); for a 14 473 bp DNA region from trnH to psbB, 19 156 bp DNA region from rpoC1 to rps7 to chlL, LA15 (GTTCCTCTTTTTTCTTATTGCTCTCCAC, this study) and LA16 (GTAAATACACTTCTCAGGCACAAACACTTCCAT, this study). The eight amplified cpDNA fragments and the λ-D5 insert DNA encompassed the entire P. patens chloroplast genome.

Sequence analysis
The amplified cpDNA fragments were sheared, cloned into pUC18, and shotgun-sequenced using Shimadzu RISA384 sequencers by the Department of Genomic Research, Shimadzu Co., Japan. Nucleotide sequence files were assembled using the Phrap program (Phil Green, University of Washington, Seattle, WA, USA) and the resulting sequences were analyzed with Genetyx-Mac v.9.0 (Software Development) and Sequencer software (Hitachi Software Engineering). When the nucleotide sequences of either protein- or RNA-coding regions contained ambiguous bases or apparent small gaps, they were corrected by sequencing of re-amplified P. patens DNA fragments (19,23). Fifteen cpDNA regions with small gaps and 33 protein- or RNA-coding regions were amplified from total cellular DNA, and subcloned into pGEM-T Easy (Promega). At least three independent clones were picked up and sequenced using the DYEnamic ET Terminator Cycle Sequencing Kit (Amersham Bioscience) and an ABI PRISM 3100 DNA sequencer (Perkin-Elmer, Applied Biosystems). Database searches and sequence analysis data were performed as described previously (19).

DNA analysis of four mosses
Total cellular DNA from the mosses Hylcomium splendens (Hedw.) Bruch and Schimp., Plagiothecium euryphyllum (Card. and Ther.) Iwats., Bartramia pomiformis Hedw. and Ceratodon purpureus (Hedw.) Bruch and Schimp. were used as templates for PCR. Primers P1–P4 were designed from the sequences that have been determined for the P. patens genes rps11, rpoB, petD and petN, respectively: P1, 5′-TTTTTGTTCGTGATATTACCTCTCTATAGTGTA-3′; P2, 5′-CTACCATAGCTCCTCAGCGAGTAGGT-3′; P3, 5′-CTAAATGGTAAAGGTGATGGGTGC-3′; and P4, 5′-TAATCTAACCTTTTCTATTCCGCTTCC-3′. The amplified DNA fragments were subcloned into pGEM-T Easy (Promega) and sequenced as described previously (19, 24).

Isolation and sequence analysis of cDNA and genomic DNA clones
Full-length cDNA encoding PpRpoA was prepared by 5′-rapid amplification of cDNA ends using primers designed from an expressed sequence tag (EST) clone as described (24). Genomic clones were isolated using the cDNA probe from a P. patens genomic library (19). Sequencing was performed with an ABI PRISM 3100 sequencer and the DYEnamic ET Terminator Cycle Sequencing Kit (Amersham Bioscience) using appropriate sequencing primers.

Construction of fusion genes and microscopic observation
A DNA fragment encoding the N-terminal 94 amino acid residues of PpRpoA was amplified from the cDNA as above with oligonucleotides containing restriction sites (underlined), 5′-ATAATCCATTTGCAGCGCTCAGCAGCAGCATC-3′ and
5'-AGCTTCTAGAATGGCAGTCATGAGCGC-3'. SalI-NcoI-digested amplified DNA fragments were introduced into SalI-NcoI-cleaved CaMV35S-sGFP(S65T)-nos3' (23) to create the PpRpoA-gfp construct. The reporter construct was introduced into the protoplasts prepared from P. patens protonemata by polyethylene glycol-mediated transformation (20). One day after transformation, the fluorescence of green fluorescent protein (GFP) was monitored using a BX-50 fluorescence microscope (Olympus).

RESULTS

Overall structure of Physcomitrella cpDNA

The circular cpDNA of P. patens is 122,890 bp long, and its gene map is shown in Figure 1. Overall GC content of the cpDNA is 28.5%, which is closer to those of the liverwort M. polymorpha (28.8%) (7) and the hornwort A. formosae (32.9%) (8), the charophyte Chaetosphaeridium (29.6%) (9) and the algae (30.4–33%) (2), than to the GC content of the vascular plants (38–39%) (3). Two identical copies of a large sequence (9589 bp) containing the rDNA operon (rrn16–rrn23–rrn4.5–rrn5) are present in an inverted orientation (IR_A and IR_B). A large single-copy (LSC) region of 85,212 bp and a small single-copy (SSC) region of 18,501 bp separate these sequences from one another. This quadripartite structure of cpDNA is well conserved in the bryophytes (7,8), a charophyte (9) and most vascular plants (3). BglII, ClaI and SacI restriction maps predicted from the complete nucleotide sequence coincide with the physical maps constructed previously (25).

Relative to hepatic bryophytes (M. polymorpha and A. formosae), an extensive rearrangement of the P. patens cpDNA is evident within the LSC region. A large inversion is present that encompasses a 71 kb region from petD to rpoB, its apparent endpoints lying between rps11 and rpoB at one end and petD and rnc-GCA at the other (Fig. 1, arrows). To investigate whether this large inversion occurs in other moss species, we used PCR with different combinations of primers on the DNAs of C. purpurea, P. eurycyphllum, H. splendens and

Figure 1. Gene map of the moss P. patens chloroplast genome. IR sequences (IR_A and IR_B) are separated by the SSC and LSC regions. Genes on the inside of the map are transcribed clockwise and genes on the outside are transcribed counter-clockwise. Genes with related functions are shown in the same color. Asterisks denote intron-containing genes or split genes. Arrows indicate positions of the endpoint of large inversion relative to liverwort M. polymorpha and hornwort A. formosae cpDNA.
were used. This indicates that the large inversion occurs only in *P. patens*.

**Gene content**

One hundred and eighteen identified genes are listed in Table 1. Eighty-three protein-coding genes, including the hypothetical chloroplast reading frame (*ycfS*), were identified. The differences in genes in the cpDNAs of the three classes of bryophyte are summarized in Table 2. Four protein genes, rpoA, cysA, cysT, and *ccsA* (*ycf5*) present in the liverwort *M. polymorpha* (7) and the hornwort *A. formosae* cpDNA (8), are absent from *P. patens* cpDNA. Two genes, cysA and cysT, that encode sulfate transport proteins, are present in the green algae *Chlorella* (26) and *Nephroselmis* (27), but are absent in charophytes (9) and vascular plants (4). *ccsA* is usually found in land plants (28±30) and algae (2,26), except *Euglena gracilis* (31). Furthermore, two open reading frames (ORFs) encoding 197 and 40 amino acid residues are located downstream of petA. The 70 N-terminal amino acid residues predicted from ORF 197 show ~30% identity with those of *Chaetosphaeridium* ORF 231 (9) and *Mesostigma* ORF 167 (32), which also occur downstream of petA. However, no sequences similar to ORF 197 are found in the cpDNAs of other plants or algae. A sequence homologous to ORF 40 (~40% amino acid identity) has been found in the hornwort *A. formosae* (8).

Thirty-five RNA genes were identified, of which 31 are tRNA genes and four are rRNA genes. A tRNA-like sequence is present between *rpl32* and *trnL*-UAG in the SSC region (Fig. 1). The nucleotide sequence shows 71.6% identity with that of *A. formosae* trnP-GGG (8) and 62.1% with that of *M. polymorpha* pseudo trnP-GGG (7), which are located in similar positions to the *P. patens* trnA-like sequence. However, the *P. patens* sequence encodes a tRNA with an AAC anticodon, which is complementary to the GTT codon (encodes valine). No tRNAs or tRNA genes with AAC

---

**Table 1.** Genes encoded by *P. patens* cpDNA

<table>
<thead>
<tr>
<th>Gene products</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photosystem I</td>
<td>psaA, B, C, I, J, M</td>
</tr>
<tr>
<td>Cytochrome b6/f</td>
<td>petA, B\textsuperscript{a}, D\textsuperscript{a}, G, L, N</td>
</tr>
<tr>
<td>ATP synthase</td>
<td>atpA, B, E, F, H, I</td>
</tr>
<tr>
<td>Chlorophyll biosynthesis</td>
<td>chlB, L, N</td>
</tr>
<tr>
<td>Rubisco</td>
<td>rbcL</td>
</tr>
<tr>
<td>NADH oxidoreductase</td>
<td>ndhA\textsuperscript{a}, B\textsuperscript{a}, C, D, E, F, G, H, I, J, K</td>
</tr>
<tr>
<td>Large subunit ribosomal proteins</td>
<td>rps2\textsuperscript{a}, 3, 4, 7, 8, 11, 12\textsuperscript{b}, 14, 15, 18, 19</td>
</tr>
<tr>
<td>Small subunit ribosomal proteins</td>
<td>rps11, rpoB, petD, trnC</td>
</tr>
<tr>
<td>RNAP</td>
<td>clpP\textsuperscript{a}, C2</td>
</tr>
<tr>
<td>Translation factor</td>
<td>infA</td>
</tr>
<tr>
<td>Other proteins</td>
<td>accD, clpP\textsuperscript{a}, matK</td>
</tr>
<tr>
<td>Proteins of unknown function</td>
<td>ycf1, 2, 3\textsuperscript{a}, 4, 10, 12, 66\textsuperscript{b}</td>
</tr>
<tr>
<td>Transfer RNAs</td>
<td>rrl\textsuperscript{a}, 28\textsuperscript{a}, 55\textsuperscript{a}</td>
</tr>
<tr>
<td>Ribosomal RNAs</td>
<td>tRNA(UCC)\textsuperscript{a}, D(GCA), D(GUC), E(UUC), F(GAA), G(GCC), G(UCC)\textsuperscript{a}, H(GUG), H(CAU), H(GAU)\textsuperscript{a}, K(UUU)\textsuperscript{a}, L(CAA), L(UAA)\textsuperscript{a}, L(UAG), f(MCAU), M(CAU), N(GUU)\textsuperscript{a}, P(UGG), Q(UUG), R(ACG)\textsuperscript{a}, R(CCG), R(UCU), S(GCU), S(UGA), T(GGU), T(UGU), V(GAC)\textsuperscript{a}, V(UAC)\textsuperscript{a}, W(CCA), Y(GUA)</td>
</tr>
</tbody>
</table>

\*Gene containing a single intron.
\textsuperscript{a}Gene divided into two independent transcription units.
\textsuperscript{b}Gene containing two introns.
\textsuperscript{c}Two gene copies due to the IR.
anticodon sequences have been found in other organelles or cyanobacteria. Therefore, we tentatively classified this tRNA-like sequence as a pseudogene. Genes encoding stable RNAs other than rRNAs and tRNAs have been identified for tscA in Chlamydomonas reinhardtii (33), for rnpB and tmRNA genes in the non-green algae (34), and for sprA in tobacco (35), but sequences homologous to these genes are not present in this moss.

**Intron content**

Eighteen genes for six tRNAs and 12 proteins contain introns, as shown in Figure 1 and Table 1. The intron of the trnL gene encoding tRNA\(^{1\text{-leu(UAA)}}\) is classified into the group I-type intron and the remaining introns of the 17 genes are categorized into the group II intron (8, 21). A maturase-like polypeptide is encoded within the longest intron (2168 bp) of trnK coding for tRNA Lys-UUU. The gene for ribosomal protein S12 is divided into 5\(^{¢}\) and 3\(^{¢}\)-rps12, and each gene segment can be transcribed independently and the transcripts trans-spliced. clpP and ycf3 contain two introns, as in the corresponding tobacco and rice genes, and the other intron-containing genes have a single intron. Anthoceros formosae ycf3 also has two introns, whereas M. polymorpha ycf3 has only the first of these introns. ycf66 is absent in A. formosae, but present in P. patens and M. polymorpha. The intron of rrn23 is present in A. formosae as well as in the green algae Chlorella (26) and Chlamydomonas (36), but not in P. patens, M. polymorpha (8) or the prasinophytes Mesostigma (32) and Nephroselmis (27), which constitute the most ancient green-plant lineage (5).

**Loss of rpoA from moss cpDNA**

The rpoA gene is normally present immediately downstream from a ribosomal-protein gene cluster (rpl23 to rps11) and is located between rps11 and petD in most land plants. In contrast, P. patens rpoB is located downstream from rps11, and rpoA is completely absent from the cpDNA (Fig. 1). Interestingly, in the other moss species, H. splendens, P. eurhythrum, B. pomiformis and C. purpureus, petD is located downstream from rps11, and there is no rpoA gene located between them (Fig. 2B). This observation strongly suggests that the loss of the rpoA gene is a general occurrence in mosses. However, the possibility that the rpoA gene is present at some other locus on the cpDNA cannot be ruled out for other mosses.

**Identification of the nuclear rpoA gene**

The absence of rpoA in the moss cpDNA strongly suggests the transfer of the rpoA gene from the chloroplast to the nuclear genome. To isolate the nuclear rpoA counterpart, we searched the DNA databases and identified an EST (GenBank accession no. BI740521) encoding the C-terminal 57 amino acid residues of the α subunit. We next obtained and sequenced the full-length cDNA. The predicted protein contains 450 amino acid residues with the characteristic arrangement of protein domains identified in the Escherichia coli α subunit of RNAP (37) and shows 50% amino acid identity with chloroplast-encoded liverwort M. polymorpha RpoA and 29% identity with the E. coli α subunit (Fig. 3). Therefore, we designated this protein and gene PpRpoA and M. polymorpha chloroplast RpoA, αNTD and αCTD denote α subunit N-terminal and C-terminal domains, respectively. The gray boxes denote regions conserved in sequence between α homologs of prokaryotic, archaeabacterial, chloroplast and eukaryotic RNAPs.

**Chloroplast localization of the moss PpRpoA**

The N-terminal sequence of PpRpoA was predicted to specify chloroplast targeting (a score of 0.911) by the TargetP
program for protein sorting (38). To localize PpRpoA in cells we further investigated the cellular localization in protoplasts of P. patens protonemata through the use of a chimeric protein comprised of the PpRpoA N-terminal 94 amino acid residues fused to GFP. The PpRpoA-gfp construct produced GFP fluorescence associated with chloroplasts (Fig. 4). This clearly demonstrated that the N-terminal extension of the PpRpoA functions as a chloroplast-targeting signal. This strongly suggests that the nuclear-encoded PpRpoA is a component of chloroplast RNAP.

DISCUSSION

The size and quadripartite structure of the chloroplast genome of P. patens is very similar to that of M. polymorpha. The cpDNA of P. patens, like the vascular plants, however, lacks cysA and cysT and contains a ycf3 gene with two introns, whereas that of the liverwort does not. This implies that the moss P. patens diverged from the hepatic bryophytes and is more closely related to the vascular plants.

The most striking feature of the cpDNA of the moss P. patens is the absence of rpoA. After the divergence of the mosses from the hepatic bryophytes, rpoA was lost from the cpDNA, together with ycf5, cysA and cysT. The loss of rpoA gene appears to be typical in mosses, and poses the question as to whether moss chloroplasts possess a transcription system unique among the green land plants. rpoA is already known to be absent from cpDNAs of the photosynthetic plant geranium (Pelargonium hortorum) (39), the parasitic vascular plant Epifagus virginiana (40), and from the plastid-like genome of the malaria parasite Plasmodium falciparum (41). The retained 60% of rpoA in the geranium cpDNA is extensively fragmented but an intact rpoA gene was not detected in the nuclear DNA by Southern hybridization analysis (39). In contrast, rpoB, C1 and C2 are most likely intact genes in P. hortorum (39) and P. falciparum (41). This suggests that rpoA gene was preferentially lost from the cpDNA, rather than rpoB, C1 and C2. The loss of single genes or groups of functionally related genes occurred similarly in specific clades, e.g. the loss of ycf1, ycf2 and accD in monocotyledonous plants (42,43), the loss of all ndh genes in the gymnosperm Pinus (29) and the loss of infA during angiosperm evolution (44).

In this study we have identified a nuclear rpoA gene encoding a protein predicted to contain 450 amino acids with high sequence identity to the α subunit of E. coli RNAP and the cpDNA-encoded RpoA (Fig. 3). Our findings clearly indicate that a functional rpoA has been transferred from the chloroplast to the nucleus. This is the first evidence for the presence of a nuclear rpoA. Nuclear rpoA genes with a chloroplast transit sequence are not found in Arabidopsis thaliana (45) and rice (46), whose cpDNAs retain an intact rpoA.

There are 171 identical amino acid residues between PpRpoA and liverwort M. polymorpha RpoA, of which 98 are encoded by codons that differ at synonymous positions. For example, PpRpoA preferentially uses GC residues at the third position in Leu, and generally uses the codons TTC for Phe, AAG for Lys and GAG for Glu. These codons are also preferentially used in P. patens ESTs (18). In contrast, A or T residues are preferentially used for the M. polymorpha rpoA as well as P. patens chloroplast protein genes. This indicates that the transfer of the moss rpoA to the nucleus was not a recent event and may have occurred immediately after the divergence of mosses from the hepatic bryophytes. There is also an EST (AW098196) in the databases from the moss C. purpureus which is predicted to encode a polypeptide similar to the P. patens PpRpoA. This indicates that mosses probably possess in general a transferred nuclear rpoA gene.

The α subunit is well characterized to be essential for RNAP assembly and basal transcription in E. coli (47,48). Therefore, in the moss chloroplast, the α subunit is indispensable for the function of plastid-encoded plastid RNAP (PEP), and the PpRpoA probably functions as the α subunit of PEP. Many P. patens chloroplast genes have canonical ±10 and ±35 sequences similar to those of E. coli promoter sequences (49). Moreover, cDNAs encoding σ-like transcription initiation factor for PEP have been isolated from P. patens, and the encoded transit peptide of σ-like factor has been shown to function as a chloroplast-targeting signal (20,50). Therefore, there is no doubt that PEP also transcribes photosynthetic genes in the P. patens chloroplasts. Our data indicate that the biosynthesis and assembly of the core subunits of PEP are controlled cooperatively by the chloroplast and nuclear genomes, and a relatively more complicated transcription system operates in mosses than in higher plants.

SUPPLEMENTARY MATERIAL

The amino acid sequence alignment of PpRpoA with the α subunit from E. coli, Synechococcus PCC6301, tobacco and liverwort M. polymorpha chloroplasts is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Drs M. Higuchi and T. Omata for providing us with DNAs of mosses, Dr T. Tsuzuki for drawing the gene map of cpDNA, and Drs K. Yoshinaga and M. Hasebe for valuable discussions. We also thank Drs A. C. Cuming and...
S. Bashiardes for the Physcomitrella genomic library as part of the Physcomitrella EST Programme at the University of Leeds and Washington University in St Louis. This work was supported by Grants-in-Aid from the Ministry of Education, Science, Sports, Culture and Technology (12874107 and 13206027 to M.S.) and of the Ministry of Agriculture, Forestry and Fisheries (Bio-Design Project).

REFERENCES


