Structure and evolution of the largest chloroplast gene (ORF2280): internal plasticity and multiple gene loss during angiosperm evolution

Stephen R. Downie*, Deborah S. Katz-Downie*, Kenneth H. Wolfe**, Patrick J. Calie***, Jeffrey D. Palmer

Department of Biology, Indiana University, Bloomington, IN 47405, USA

Abstract. We have determined the nucleotide sequence of the *Pelargonium* × *hortorum* ORF2280 homolog, the largest gene in the plastid genome of most land plants, and compared it to published homologs from *Nicotiana tabacum*, *Epifagus virginiana*, *Spinacia oleracea*, and *Marchantia polymorpha*. Multiple alignment of protein sequences requires an extraordinary number of gaps, indicating a very high frequency of insertion/deletion events during the evolution of the protein; however, the overall predicted size of the protein varies relatively little among the five species. At 2 109 codons, the *Pelargonium* gene is smaller than other land plant ORF2280 homologs and exhibits a rate of nucleotide substitution several times higher relative to *Nicotiana*, *Epifagus*, and *Spinacia*. Southern-blot and restriction-mapping studies were carried out to uncover length variation in ORF2280 homologs from 279 species (representing 111 families) of angiosperms. In many independent angiosperm lineages, this gene has sustained deletions ranging in size from 200 bp to almost 6 kb. Based on the severity of deletions, we postulate that the chloroplast homolog of ORF2280 has become nonfunctional in at least four independent lineages of angiosperms.

Key words: Chloroplast DNA – ORF2280 – Multiple gene loss – Deletion

Introduction

Chloroplast gene content is highly conserved in land plants (Palmer 1991). All but one of the 113 genes found in to-bacco (*Nicotiana tabacum*) chloroplast DNA (cpDNA; Wolfe et al. 1991) are also present in the genome of the bryophyte *Marchantia polymorpha*, while the latter genome contains six genes not found in tobacco cpDNA. Tobacco and rice (*Oryza sativa*) cpDNAs, which diverged roughly 150–200 million years ago (Wolfe et al. 1989), differ by only three genes, all of which represent losses from the rice genome (Hiratsuka et al. 1989).

One of the missing rice genes is ORF2280[1], which is also the largest gene in the chloroplast genomes of tobacco, *Epifagus* (ORF2216; Wolfe et al. 1992 a), spinach (*Spinacia oleracea*, ORF2131; Zhou et al. 1988), and *Marchantia* (ORF2136; Ohyama et al. 1986). The gene is well conserved in sequence between the dicots tobacco and spinach (Zhou et al. 1988), where it is buffered from change by virtue of its location within the mutationally-retarded inverted repeat (IR). In contrast, it exhibits only a low level of conservation (primarily at its C-terminus) between dicots and *Marchantia*, where it is a single copy gene. ORF2280 is also conserved in immediate location in all examined cpDNAs, being flanked by *trnI*-CAU and *trnL*-CAA located on the opposite strand.

Partial sequences for ORF2280 are known from broad bean, tomato, and *Oenothera* (Blasko et al. 1988; Herdenberger et al. 1988; Richards et al. 1991; Nimzyk et al. 1993). The absence of this gene from rice cpDNA (Hiratsuka et al. 1989) results from several major deletions within the region occupied by ORF2280 in tobacco. Two short regions of residual ORF2280 sequence, recognized as ORF28 and ORF64, have been reported in rice (Hiratsuka et al. 1989; Shimada and Sugiura 1991).

Our interest in ORF2280 evolution started several years ago when we noticed that certain ORF2280-derived probes

---

* *Present address:* Department of Plant Biology, University of Illinois, Urbana, IL 61801, USA
** Department of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland
*** Department of Biological Sciences, Eastern Kentucky University, Richmond, KY 40475, USA

Correspondence to: S. R. Downie

---

[1] ORF2280 was originally reported as two ORFs of 581 and 1708 codons in tobacco (Shinozaki et al. 1986) due to a frameshift sequencing error (Shimada and Sugiura 1991). We have followed Zhou et al. (1988) in assuming that the error is due to omission of a G (guanine) nucleotide after position 90 586 in the tobacco cpDNA sequence

from tobacco hybridized surprisingly weakly to cpDNA from *Pelargonium × hortorum* (the common cultivated geranium), a species whose cpDNA has been found by physical and gene mapping studies to be extensively rearranged relative to most other land plants (Palmer et al. 1987). Subsequently, and following the report of ORF2280 absence in rice cpDNA (Hiratsuka et al. 1989), we observed that the cpDNAs of other angiosperms failed to hybridize to at least one of several probes internal to tobacco ORF2280 (Downie and Palmer 1992 a). These results suggested that the gene had changed to an unusual extent (in length and/or sequence) in a number of lineages and, as a result, may no longer be functional in some of them. In this study, we (1) report the nucleotide sequence of the ORF2280 homolog from *Pelargonium*, (2) describe the tempo and pattern of ORF2280 evolution by comparing the five sequenced forms of the gene, and (3) examine the extent of length mutation and loss of ORF2280 homologs from 279 species of angiosperms.

Materials and methods

Restriction fragments containing the ORF2280 homolog from *Pelargonium* cpDNA were cloned into BlueScript vectors (Stratagene, Inc.) following standard methods and sequenced by the dideoxy method using ^{35}S and Sequenase (USB Corporation). Some 21% of the *Pelargonium* ORF2109 sequence was determined on both strands, 51% was sequenced from two or more independent and overlapping sequencing runs on the same strand, and 28% was sequenced only once. Ambiguities were resolved using deaza-GTP, dITP, and second-strand sequencing. Furthermore, the sequence was rigorously checked for possible frame-shifting errors by dot-matrix comparison, in all three reading frames, to the four published homologous ORF sequences. The GenBank accession number for the *Pelargonium* ORF2109 sequence is M83200. ORF2280 amino-acid sequence homologs from the five species described in Fig. 1 were aligned using CLUSTAL (Higgins and Sharp 1988) on a microVAX computer. The alignment was adjusted by eye after comparing the results obtained by altering the order of sequence addition as well as the results of pairwise dot-matrix plots. The structure of ORF2280 homolog sequences was investigated in 279 species of angiosperms (Table 1). These species represent 111 families and comprise members of all six subclasses of dicots and four of the five subclasses of monocots (*sensu* Cronquist 1981). The isolation of cpDNA or total cellular DNA from leaf material was accomplished using the sucrose-gradient technique of Palmer (1986) or the modified CTAB procedure of Doyle and Doyle (1987), respectively. All DNAs were further purified by centrifugation in cesium chloride-ethidium bromide gradients. Restriction endonuclease digestions, agarose-gel electrophoresis, bidirectional transfer of DNA fragments from agarose gels to nylon filters (Zetabind, AMF Cuno), labeling of recombinant plasmids with ^{32}P by nick-translation or random priming, filter hybridizations, and autoradiography were performed according to Downie and Palmer (1992 a), being modified from the approach outlined in Palmer et al. (1988). All DNAs were digested singly with two or four restriction enzymes (*Bam*HI and *Hind*III or *Bam*HI, *Hind*III, *Bgl*II, and *Eco*RV), with the exception of seven species of Geraniaceae, which were digested singly with *Bam*HI, *Hind*III, *Bst*XI, and *Sac*I. All filters were washed in 2 × SSC, 0.5% SDS, twice for 5–15 min at room temperature, and then two to three times for 30–60 min at 65 °C. To assess gene structure, five cloned restriction fragments (described in Fig. 5 and illustrated in Figs. 1, 2 and 5) from tobacco cpDNA were used as hybridization probes against all blots containing DNAs from all 279 taxa. Larger, parental clones from which these clones were generated were graciously provided by M. Sugiura (Sugiura et al. 1986). Because this study is part of a larger investigation to study chloroplast genome organization in an-

giosperms (S. Downie and J. Palmer 1992 a, and unpublished data), our original numbering system of the probes is maintained. These probes are numbered from 79 to 83, range in size between 320 bp and 2.949 kb, and together comprise virtually the entire ORF2280 region from tobacco. Probe 78 (Figs. 2 and 5), at 213 bp in size, contains the 5′ 108 bp of ORF2280. Because this small probe hybridized weakly to many angiosperm cpDNAs, it was not used to investigate gene structure. Seven probes adjacent to ORF2280 (numbered 74–77 and 84–86 in Fig. 5) were used to test for linkage between ORF2280 and the genes that normally flank it.

Results and discussion

ORF2109 in Pelargonium

Multiple alignment of the *Pelargonium* chloroplast ORF and four other land plant ORF2280 homolog sequences (Fig. 1) requires an extraordinary number of gaps, indicating a very high frequency of insertions/deletions during the evolution of the protein. There is a gap in at least one sequence at 39% of the positions in the alignment (1049 out of 2637 sites; Fig. 2). The only other chloroplast gene with a comparable level of internal length variation is ORF1901 (Wolfe et al. 1992 b). The two largest gaps are deletions of 206 and 108 residues near the N-terminus of the protein in *Pelargonium* and spinach, respectively (Fig. 2). *Pelargonium* ORF2109 is smaller than any of the other four sequenced genes. Given the abundance of length mutations, the overall predicted size of the protein varies remarkably little among the examined species (2109–2280 residues).

Repetitive elements

Seven regions of the protein contain directly-repeated motifs of 4–8 amino acids, and all but one of these repeats are species-specific (Fig. 1). Five of the repeats are present in the *Pelargonium* sequence, and all but one of these correspond to insertions unique to that species [e.g. the sequence (E/Q)VSKILIP at position 780–795, and the more complicated repeat from 1851–1921 based on the sequence EEAELQD]. The fifth *Pelargonium* repeat – the duplication of RTLLSK at positions 913–926 – has apparently arisen without a change in length, i.e., by obliteration of a flanking sequence, possibly by gene conversion. These repeats are reminiscent of the sequence flux observed at two distinct sites in the ORF2280 homolog of *Oenothera*, detected either as a length polymorphism among isolates of a single species (*O. hookeri*; Blasko et al. 1988) or as a length difference between ORF2280 sequences from two species (*O. odorata* and *O. berteriana*; Nimzyk et al. 1993). It seems likely that these length mutations are effectively neutral in terms of the function of the protein. Similar repetitive insertions have been identified in other chloroplast genes: *rpoC2* and *rps18* sequences from corn and rice (Igloi et al. 1990; Shimada et al. 1990; Weglöhner and Subramanian 1991), pea *accD* (formerly called ORF587 and *zfpA*; Smith et al. 1991), and the *Oenothera* homolog of tobacco ORF1901 (formerly called ORF1244; Nimzyk et al. 1993). In many of these cases the repetitive unit is seven or eight amino-acid residues.

Table 1. Angiosperms surveyed for length variation in chloroplast ORF2280 by blot hybridization. Members of those families failing to hybridize to at least one of the five tobacco ORF2280 probes (i.e., probes 79–83) are indicated by boldface. For these and related taxa, the hybridization results are summarized in Table 2. Those families for which at least one member exhibits length variation within the gene, as ascertained by blot hybridization, are underlined. System of classification follows that of Cronquist (1981). Unless indicated, only one species per family was surveyed; otherwise, the numbers in parentheses indicate number of genera/number of species examined. Voucher information is available from S.R.D. upon request

Dicotyledons	**Passifloraceae**	Caprifoliaceae (6/7)
Magnoliidae	Primulaceae	Convolvulaceae (3/3)
Aristolochiaceae	Salicaceae	Dipsacaceae (3/4)
Berberidaceae	Sarraceniaceae	Gentianaceae (4/4)
Calycanthaceae	Violaceae	Gesneriaceae (3/3)
Lauraceae		Globulariaceae
Magnoliaceae	Rosidae	Goodeniaceae (2/2)
Papaveraceae	Aceraceae	Hydrophyllaceae (3/3)
Piperaceae	Anacardiaceae	Lamiaceae (9/9)
Ranunuclaceae	Apiaceae	Lentibulariaceae
Saururaceae	Araliaceae (2/2)	Lobeliaceae (4/9)
Winteraceae	Balsaminaceae	Loganiaceae (4/4)
	Cornaceae (2/3)	Menyanthaceae (4/4)
Hamamelidae	Crassulaceae	Myoporaceae (3/4)
Ceridiphyllaceae	Euphorbiaceae	Nolanaceae
Juglandaceae	Fabaceae (12/17)	Oleaceae (4/6)
Ulmaceae (4/5)	Francoaceae	Orobanchaceae (2/2)
Urticaceae	**Geraniaceae (4/8)**	Pedaliaceae (2/2)
	Grossulariaceae	Plantaginaceae (1/2)
Caryophyllidae	Hippocastanaceae	Polemoniaceae (3/4)
Aizoaceae	Hydrangeaceae	Rubiaceae (4/4)
Amaranthaceae (3/4)	Iteaceae	Scrophulariaceae (5/5)
Basellaceae	Krameriaceae	Solanaceae (4/4)
Cactaceae	Linaceae	Valerianaceae
Caryophyllaceae (4/4)	Onagraceae (9/9)	Verbenaceae (7/8)
Chenopodiaceae (5/5)	Oxalidaceae	
Didiereaceae (2/2)	Penthoraceae	Monocotyledons
Molluginaceae	Polygalaceae	Arecidae
Nyctaginaceae (2/2)	Rosaceae (3/3)	Araceae
Phytolaccaceae (3/3)	Sapindaceae	Arecaceae
Plumbaginaceae	Saxifragaceae (4/4)	
Polygonaceae (2/2)	Tropaeolaceae	Commelinidae
Portulacaceae (2/3)		Commelinaceae
	Asteridae	**Poaceae (8/8)**
Dilleniidae	Acanthaceae (3/3)	
Begoniaceae	Apocynaceae (5/5)	Zingiberidae
Bombacaceae	Asclepiadaceae (2/3)	Bromeliaceae (2/2)
Brassicaceae (3/3)	Asteraceae (2/2)	Strelitziaceae
Droseraceae	Bignoniaceae (3/3)	
Fouquieriaceae	Boraginaceae (4/4)	Liliidae
Loasaceae	Buddlejaceae (1/2)	Iridaceae
Malvaceae	Callitrichiaceae	Liliaceae (4/5)
Nepenthaceae	Calyceraceae (2/2)	Orchidaceae (2/2)
Paeoniaceae	**Campanulaceae (3/6)**	Pontederiaceae

Functional domains

The function of ORF2280 is unknown; however, it shares a few short amino-acid motifs, including parts of a nucleotide-binding site, with members of the CDC48 family of proteins and may possibly be a proteolytic ATPase (Wolfe 1994). Its presence in the plastid genome of the nongreen parasitic plant *Epifagus* implies that it is not involved in photosynthetic metabolism (Wolfe et al. 1992 a). Richards et al. (1991) have reported a relatively high level of transcription of ORF2280 in fruits of tomato, suggesting a chromoplast-specific function. Glick and Sears (1993) have shown that a large protein product derived from

ORF2280 is present in the soluble fraction of chloroplast from tobacco, spinach and *Oenothera*.

Data summarized in Fig. 2 indicate that the sequence of the N-terminal 1000–1500 residues of the protein cannot be critical; the level of identity between the dicots and *Marchantia* in this region is little greater that would be expected between random protein sequences. Given this lack of conservation, the maintenance of an ORF through this region (approximately 4 kb of A+T-rich DNA) in five species is surprising. The most plausible explanation for the lack of stop codons or frameshifts is that the overall size of the protein is subject to some form of constraint. Alternatively, there may be many very small essential domains
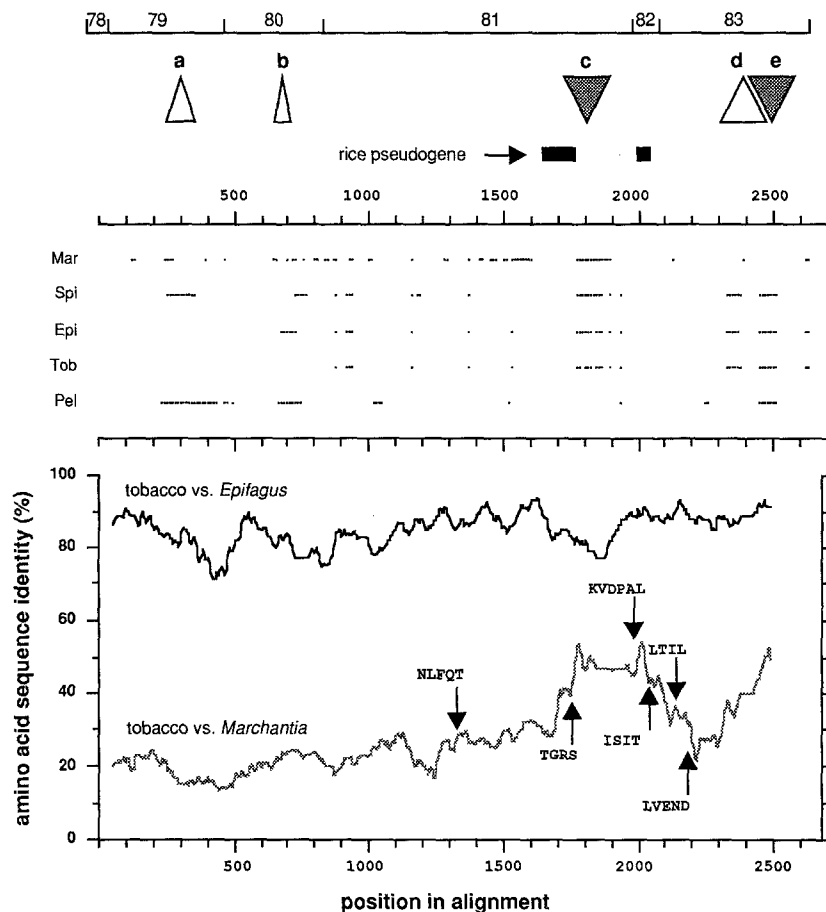
```
                                                           78 ◆ 79
Pel     1 MDEQEFKLFLKFFILELKGILREIKNSHQLLDSWTKLNSLGTLIQIFFNPERAFQLLDPRVWKILLSRNSRGWRKTRHFTIGRVLLFVVVVLMYRMSRRNMVENKKSYLTGVLPIPLNPI
Tob     1 MRGHQF----KSWIFELREILREILREIKNSHHFLDSWTQFNSAGSFIHIFFHQERFLKLFDPRIWSILLSRNSQGSTSNRYFTIKGVILFVVAVLIYRINNRNMVERKNLYLIGLLPIPMNSI
Epi     1 MKEHPFPY--KSWILE----LREIKNSHYFLDSWTKFNSVGSYINIFSHQERFIKLFDPRILSILLSRNSQGSTSNRYFTIKGVILFVVAVLIYRINNRNMVEIKNIYWRGLLPIPMNSI
Spi     1 MKGHQF----KSWIFELREILREIKNSHYFLDSWTQFNSVGSFIHIFFYQERFLKLFDPRIWSILLSPNSQGSTSNRYFTIKGVULFVVVVLIYRITNRNMVERKNLYLIGLFPIPMNSI
Mar     1 MKQKLP----KKKSLYKNLDLDEIQKIQNLGNPYTKWSLIRLLIAIFSNKRNFSTLLDFQILTSLFFRDLYNSKKKKKFLLNILVFLTLPFFVYILDKSIVEQQNF--DFLKIQKQNF
          +                 . +           ++           + +        + + +++         +        +  +      .+....+     +  .+.

Pel   121 GHRNDTLEESIGFSNINRLILPLLYLPKGKKIPESSFLDPNESTW----ALP-ITKKSIMPESR-WGSRWWCSWVGKRRDSSCKRAHKTVAGIEISFKEKKSKYLEFL-----------
Tob   117 GPRNDTLEESVGSSNINRLIVSLLYLPKGKKISESCFLNPKESTW----VLP-ITKKCSMPESN-WGSRWWRNWIGKKRRDSSCKISNETVAGIEILFKEKDLKYLEFLFVYYMDDPIRKD
Epi   115 GPRNDTLEELVGSYNINRFIVSLLYLTKGKKISESFFLNLKESTL----VLP-ITKKCSMPESN-WGSRWWRNWTGKNRDYSCKISNETVAGIEILFKEKDKKYLEFIFFYYMDDPIRKD
Spi   117 GPRNDTLEKSFGSSNINRLIVSLLYLPKGKKISESYFLDPKESTW----FLP-ITKKCIMPESN-RGSRWWRNWIGKRRDSSCKISNETVAGIEISFKEKDIQYLEFPFVYYMDDPIRKD
Mar   114 IEKN---------NKSILKNNFYFLNTKFDIFLHNFFSLKKKWYKNSLLNLIDFRSILKKKEILNLHWWKFLVLEQIQSNWKISEESLSELKIVLEQKNIDELKHFFEFYINQKIYPN
          . .+....    . +.........+                      .      +  +............           ...++  ..... ......+...+  .+.. +.+...+....

Pel   223 -------------------------------------------------------------------------------------------------------------------------
Tob   231 HDWELFDRLSLRKSRNRINLNSGPLFEILVKHWISYLMSAFREKIPIEV-EGFFKQQGAGSTIQ---SNDIEHVSHLFSR---NKWAISLQNCAQFHMWQFRQDLF--VSWGKNPPESDF
Epi   229 HDWELFDRLSPSKRLNKINFYSGPLFEILVKRRIYYLMSAFREKIPIEVVKGFFKQQKVGSTIQ---SNDIEHVSHFFSR---NKRAISLKNSAQFNMWQFRQDLL--VSWGENPHESDF
Spi   231 HDWELFDCLSL------------------------------------------------------------------------------------------------------------F
Mar   224 NNWE--------------------------------YYFYSIFINQLKIDIKNSKYNKNSIGFEVFLAFCEKLLFEVEFLSKPNNNNLQMKLNCLENFSFLDIFCILNKKLPWVNKKIFKN-
          .++...              ...+         ..+.+......+..   ......   +....   .........+.   ...+   ......+  ..+  . ..+..+...

                                                                                        79 ◆ 80
Pel   223 ------------------------------------------------------------------------EDSEYPTLINQREIQQIKEESILWHPS----------------
Tob   342 LRNVSRENWIWLDNVWLVN-KDRFFSKVQNVSSNIQYDS----TRSSFVQVTDSSQLKGSSDQSRDHL-DSISNEDSEYHTLINQREIQQRKERSILWDPSFLQTERKEIESGRFPKCLS
Epi   341 LRNVSRANWIWLNNVWLVN-KYRFCRKVRNVSSNIKYKYDSTRSRSSFVQVTDSSQLKGSYYKSSGHFYSVISNEDSEYHTLINQREIKPLK--SIFFDPSFLQTEATEIESDQLQKRPS
Spi   243 LRNVSRENWIWLDNVWLVLN-KDRFFSKVRNVSSNIQYDF----TRSSFVQVTDSSQLKGSYYKSSDRS-NSISNADSEYHTLINQREIILRDPSFLQTEGTEIESDRFPKCLS
Mar   313 LQNFNESDKKLIESFFLLKIKGNLYFK------NYIEFV----TWQSYKK--DCLDFNKFNELNNSEIYIKIEELFSDYIYKFSKYILYEGKKSKTIIKQSF--------NNNIYYKKLN
          +.+...  ......+..  +.. +..  + . ....+.     ..+....+.....+. .. .  +.  .... ..+.+ ...... ..  . . ++.....  .... ...

Pel   250 --PSLERT-------------EEEIKEFLGNSTRSLRSF--FSDRWSELYLSSNPTERFTRDQKLLK--QDLSFVPYRRSENQEIVNLFKIITYLQ--KTVSIYPISSDPGCDRVPKDEL
Tob   456 GYSSMSRLFTEREKQMINHLFPEEIEEFLGNPTRSVRSF--FSDRWSELHLGSNPTERSTRDQKLLKQQDLSFVPSKRSENKEMVNIFKIITYLQ--NTVSIHPISSDPGCDMVPKDEP
Epi   458 GYSST---LFTEHEKQMINHMLPEEIEEFIGNPTRLVHSF--LSDRWSELHLGSNPTERSTRDHKLLKKQQDLSFVPSRSENKELVNILKIITYLK--NTVSIHPISSDPGCDGVLKDEP
Spi   357 GYSSMPRLFTAREKQMIIHLLPEEIEQLLENPTRSIRSF--FSDRWSELHLGSNPTERSTRDPQLLKKQQDVSFAPSRRSENKEMVNIFKIIKYLQ--NTVSIHPISSDPGCDMVPKDEL
Mar   413 SIFNFNTIFYFDSNNLLFDWLK---KNYYINNKPFLKSFLIYSSISNQFILFFKQKNSKSFNKNLVKKNSKDVITNVFSKENKIEINNFSKSIYYAFFEILSINEIDNKFVINKISLKNI
          ...  +. + ...  .... ..  + ..  .++....+  . . ....+. ...  ... +.  . .++. ...  . .. .++.+.......

Pel   349 DVDSSKKISFFKKNPFFDLLNLFHDRSRGGYTL--ERDFESEERFEERADLFTLSITEPDQVYHNS----------------------------------------------------
Tob   572 DMDSSNKISFLNKNPFFDLFHLFHDRNRGGYTL--HYDFESEERFQEMADLFTLSITEPDLVYHKGFAFSIDSCGLDQKQFLNE---ARDESKKKSLLVLPPIFYEENESFSRRIRKKWV
Epi   572 DMDSSNKISVFNKNTFIYLFHLFHDWNRVGYTLNLHHDFELEERFQEKADLFTLSITEPDLVYHKGFSFSI----------------------------------------------
Spi   473 DMDSSDKISFLNKNSFFDLFHLFHDRNRGGYAL--HHDFESEERFQEMADLFTLSITDPDLVYHRGFSFSIDSCGLDQKQFLNEVFNSRDESKKKSLLVLSPIFYEENESFYRRIRKKGV
Mar   530 NKKKQKRF-YLNKIKSSDNFRFINLWKIKNYS---SQQFVSNNSF----------------LLNPAFEILQQNYYLKKKNIL--FFKKLNE-------VFSNFFYFQYYKC--------
          ... .+.   .   .. ..+ . +.  .+. .  ..+.. +   . .........  .+ ..  . .+......+  .. +++......  .++.....

                                                                                            80 ◆ 81
Pel   413 ------------PKM------------------DLIHIQEK---GYIRKVLDRFFLMNRSARNFESGIPGAQIGNETLSHRTLMKYTVNRHFSSLKKSRKKSFDPLLFI----ERSMNRDP
Tob   687 RISCGNDLEDPKPKIVVFASNNIMEAVTQYRLIRNLIQIQYSTYGYIRNVLNRFFLMNRSDRNFEYGIQRDQIGKDTLNHRTIMKYTINQYLSNLKKSQKKWFEPLILISRTERSMNRDP
Epi   643 ---YMDQKQKMVVFASNNIMEAVNQSRFIRNMIKIQYSTYGYIRNVLHRFFLMNRSDHNLEYEIKRDQIGKDTLNHRTIIKYMINQHLSNFKKSQNKWFNPILFFSRTERSVNRNP
Spi   591 RIS--------------------------------RNVLNRFFLINRSDRSFEYGIQRDQIGNDTLNHRTIRKYMINQDFSNLKKSQKKWFDPLIFLSRTERFMNRDP
Mar   612 -----------KKLNIFLKFASLEKILKKR---------NKKFTISIKLFKKFYKNKLNENGEYKIESQILQNEK---------------ELNKKRKKNFQ---FNPNIKILS----
          ... .+.....+.....  ... . ++.+ ...   . . .++. ... .+. . .. ..  +.....  ...

Pel   497 AAYRDKWSKGGKTFQEHLEHFVSEQKNREKVQKSREKAVFDRFHQSQYSIDWSVFIDKQDFPKARLFFLFVLVRSFLYKSLPFLLSKLPLLLSKLPLLVSKVLPLLSKLLPFFFVSCGNI
Tob   807 DAYRYKWSNGSKNFQEHLEQSVSEQKSRFQV-------VFDRLRINQYSIDWSEVIDKKDLSKPLRFFLSKSL-LFLSKL------------------LFFLSNSLPFFCVSFGNI
Epi   756 DAYRYKRSNGSNNFLEHLEHFVSEQKSHFKF-----KIVFDLIRFNQYSIDWSAFIDTKDLSKPLRFFLSKL------------------LFFLSNSLPFFCVSFGNI
Spi   667 DAYRYKWFNGSKNFQEHLEHFVSEQKSRFQV-------VFDQLRINQYSIDWSEVIDKKDLSKSLRFFLSKSLRFFLSKL------------------LLFLSNSLPFFFVSFGNI
Mar   687 ------FYNSSKKNIYLQNKY-------------FFNKNLINNKLITWKKISNKLVISNSEYNKIIWNKKNMKFFSFSKNSVLDT------------FFFNKKSFNIITVIFDKL
          ... .. .+...  .  ....  .    +..           . ...  ... .. .  .   ...+..  +  ....+...

Pel   617 PIHRSEIRIYELKGPTVQPCNPLLESIGLQILHLNKLKPCLLDDHETSQKSKIV---GTI---------------------------------DNTDSYFSILFDDEENWMNPVKAFQRSSLI
Tob   897 PIHRSEIYIYELKGPNDQLCNQLLESIGLQIVHLKKWKPFLLDDHDTSQKSKFLINGGTISPFLFNK--IPKWMIDSFHTRNNRR--KSFDNPDSYFSMIFHDQDNWLNPVKPFHRSSLI
Epi   841 PIHRSEIYIYELKGPNDQLCNQFLEPIDLKIVHLKKRKPFLLGYHGTSRKLKLLITGG--RPFLFNK--IPRCMIDSFHTINNRS--KSFDNTDSYLSMIFHDKDNWLNLVKPFHRSSLI
Spi   758 PINRSEIRIYELKGPNDQLYNPLVESIGLQIVHLKKWKAFLLDDHDTFQKSKFLINGGTISPFLFNK--IPKWMIDSFHTRNNSG--KSFDNTDSYFSMTSHDQNNWLNPVKPFHRSSLI
Mar   771 --KKIQLNFQEIQ--KILNCFSLF----FNSKNIKKKTKIF---------KNSYFINENLTTTFSFNDKEFNIFFLELFISEINNDFLMRFFKKYLYYRI-YKDKEILFNPIE--NRQLLQ
          .. ....+..+...          .     ++       .    .+ .. . +. . +.++.. .+.   . ..+.+.  ..+.. . . +  ...   . +  . +..+.

Pel   704 SAFYKANRLRFLNNPHRFSFYCNKIFPFVVEKARFKNSDFLYGQFLKTFFIRKKPFSLCGEK--HAF-GERATLLPIESEVSKILIPQVSKILIPDDFPQSGDERYNLCKSFHFPTRSVP
Tob  1013 SSFYKANRLFLNNPHHFCFYWNTRFPFSVEKARINNSDFTYGQFLNILFIRNKIFSLCVDKKKHAFWG-RDTISPIES--------QVSKNIFPNDFPQSGDETYNLYKSFHFPSRSDP
Epi   955 SYFYKANRLQFLNNPHNFCYCNTRLPFYVEKAHIHNYYFTYGQFLNILFIRNKIFSLCVDKKKHAFWGGRDTISPIES-------QVSKNIFPKNFPQSGDETYNLSQPFHFPSRYDP
Spi   874 SSFYKANRLQFLNNPHHFCFYMEKARINNSDFTYRQFLNILFIHNKLFSLCVDKKNHAFLE-RDTISPIES--------QVSNIFLPNDFP---------------IRSDL
Mar   871 NFFEKTKILTFIDFLQDPELNYNNRFIFHLEKKTIKNNNLLYLRLLKIFLKDKRNFLLINEIK--SFIEKKNNLF-IKS--------QLSNVLLVKNSYKFFDNIFN----FHFLKQKEK
          . +.... + + ..... +     ..++.. ++..+    .  ..   + . .+ ...+.  ....+. ... ..   ...   . +    .. . .  ..+.+.

Pel   821 LVDRALYSIADISERPLTEGQM--VNLERTSFQ---PLSDIHLSDSERKNLHQDLNFNSNMGLIHTPYSEKDLPSEKRKKRNLGRNLKKCVEKGQMFRTLLSKMYRTLLSKWNLFQTYMP
Tob  1124 FVRRAIYSIADISGTPLTEGQI--VNFERTYCQ---PLSDMNLSDSEGKNLHQYLNFNSNMGLIHTPCSEKDLSSEKRKKRSLC--LKKCVEKGQMYRTFQRDSAFSTLSKWNLFQTYMP
Epi  1067 FVRL----IANIYGTPLTEGQI--VNLGRTYCQ---PLSDMNLSDSEGKNFHQYLNFNSNMGLIHTPCSDKYLPSEKRKKRSLCINKYKCVEKGQMYRTFQRDSVAFSTLSKWNLFQTYMP
Spi   969 LVRRPISIGTPLTEGQI--VHFERTYCQ---PLSDMNLSDSEKKNLHQYLNFNSNMGFIYTPCSEKYLLSEKRKKRSLC--LKKCVEKGQMYRTFQRDSAFSTLSKWNLFQTYMP
Mar   976 NIEIILNNQNYFEKSLLKKTYLKNLNLNNSYSKFSYKIFIFQLLNILNKN--NYKTFQWISELIFYSKNLNYKIQNKIEKNNY------CYNKNISYKK-KKIKTVNFFEKNNLFQTNNS
          .  . ... .+..  .     .. .+       +.... ++ . +.+..  +  .  . . +.+. . . . .+..+.+  . .+ .. +...  + ...+++++...

Pel   936 WFLTSTGYKYLTLLFL----DLFSDLVPILSLGIRQNCVSL-FDHILGDIRDDIRQGVIRPWQILQEKWVLPQRNRIR--EISRMCLRNLTLSAERIRRNNESPLTHTHLRSPNVLEFLY
Tob  1237 WFLTSTGYKYLNLIFL----DTFSDLLPILS--SSQKFVPI-FHDIM--------HGGSISWRILQKKLCLPQWNLIS--EISSKCLHNLLLSEEMIHRNNESPLISTHLRSPNAREFLY
Epi  1178 WFLTSAGYKYINLIFL----DTFSELLSILS--SSKKFVSI-FNNIM--------HGGSISWRIINKKRCLPQWNLIS--EISSKCLHNLLSEEIIRQNNESPLISTHLRSPNVREFLY
Spi  1082 WFLTSTGYKYLNFLFL----DTFSDLLPILS--SSQKFLSI-LHDIM--------HGGSISWRILQKKLCLPPWNLIS--EISSKCLHNLLLPEEMIHRNNESPLIWTHLASPNVREFFY
Mar  1087 WFFTLEWWEYNTYILLQIIQETFFQITDVLEYFKKKKIIEKNLKFFL--------KSKKISLKTLSFHNFFLKWNLRFFNEINYK--KNYLLN----------FLWSDFNLINNCNNLY
          ++..+..  . ..  +     ..+..+.+.    ++.+ .  +.+.+.  ......  +.    ..  ....  + ++.   ...+..... .+ .+.... +   ++.

Pel  1049 STLLLLLLVAGYLVCTYLSRLSKDYGELQTELKKVKSLMIPSYTIELRKLMD--RYPPSELNSFGLKNLFLVAMEELKESLSVV------GNMLEGG--GRAYGVEAIFSN---WNLNLID
Tob  1340 SILFLLLVAGYLVRTHLLFVSRASSELQTEFEKVKSLMIPSSMIELRKLLD--RYPTSEPNSFWLKNLFLVALEQLGDSLEEIRGSASGGNML-----GPAYGVKSIRSKKKDWNINLIE
Epi  1281 SILFLLLVAGYLVRTHLLFVSRASSELQTEFRKVKSLMIPSSMIELRKLLN--RYPTPASNSFWLKNLFIVAMEQLVYSLEEIR--ASGGNLL-----GPAYGVKSICSKNKYFNINLIE
Spi  1185 SILFLLFVAGYLVRTHLLFVSRASSELQTEFERVKSLMIPSYMIELRKLLD--RYPTSEPNSFWLKNLFLVALEQLGDSLEEIRGSASGDNMLLGGGPGPAYGFKSIRSKKKYLNINLID
Mar  1186 WVIFSLVIFIFLY-------------YQKIFSIIGSDCFHLWKNFEIIQYLTDRS-----RSLYFTKLTR-----RNKTALNKTENLL------------------SYFFQNLTH
          .. ..+          ...+.  ...+..+.+ .+.+.+.    .+     ..   +   ...++.++. .   .....++  + ..+.

Pel  1156 ISDLISLIPNPIDRITFSINTRHLSHTSKEIYSFIRKRERVYGAWIDDKIESLLSTSVAID--DCDRGNLLQFSTLTTTEKGVDQILLSLTQS---SKNASGSQMIEQPGEMYLRHVVD
Tob  1453 IIDLI---PNPINRITFSRNTRHLSHTSKEIYSLIRKRKNVGDWIDDKIESWVANSDSID--DEEREFLVQFSTL--TTENRIDQILLSLTHSDRLSKNDSGYQMIEQPGAIYLRYLVD
Epi  1391 --DLI---PNPINRIIFSRNMRHLSHTSKEIYSLIRKRKNVNGDWIDDIIESWVANSDSID--DEEREFLVQFSAL--TTEKRIYQILLSLTHSDHLSKNDSGYKMIEQPGAIYLRYLVD
Spi  1303 ILDLISLIPNPINRTNRHLSHTSKEIYSLIRKRKNVNGDWIDDKIESWVANSDSID--DEEREFLVQFSTL--TTENRIDQILLSLTHSDHLSKNDSGYQLIEQPGAIYLRYLVD
Mar  1261 Y----------ITNIKF-------------YLLTKK------NLKKWLINNKTLDLSRRKRKLLVQ--SL--ITHNKIQNYGFELNSNKQFFTSYFGYQITNQQGLLYFQYLAQ
          . ... .+    ...+ .     .++.+..+.+ .+.+.   .+ .. ...  .  .  .   ...+. .+....+  + ...+ .+... +. ...  ++.
```

Fig. 1.

```
Pel 1271 LQKKYLMGYEFNTSSLAERRIFLAHYQTMTYSKTSCGVNAFHFPSHEKPFSFRLDLSPPRGILVIGSIGTGRSYLIKSLAKNTHFPLITLEMEARSSWPFFQHLDEIYGDQLEYVYDDTS
Tob 1566 IHKKHLMNYEFNPSCLAERRIFLAHYQTITYSQTSCGENSFHFPSHGKPFSLRLALSPSRGILVIGSIGTGRSYLVKYLATNSYVPFITVFLNK-----FLDN------KPKGFLLDEI-
Epi 1502 IHKKYLLNYECNTSCLVERRVFLAHSQTITYSQTSRGTNTLHFPSQGKPFSISLALSPSKGILVIGSIGTGRSFLVKYLATNSYVPFITVFLNK-----FLDN------KPKGFL-----
Spi 1419 IHKKYLMNYEFNTSCLAERRVFLAHYQTITYSQTSCGANSFHFPSHGKPFSLRLALSPSRGILVIGSIGTGRSYLVKYLATNSYVPFITVFLNK-----FLDN------KPKGSL-----
Mar 1342 FFQKNLINNSLD---LANKWIVFSFWHKIFSSQKLRQTNNIELGFQNIPVPLQFGLSYSKGILLIGPIETGRSYLIKNLAAESYVPLFKISINK-----LLYN------KP---------
         .+ +    .. .. +    ...  ...  .... ...+ .. .+    ...   +.  . ++.  +++.++.+.+++++ + + ++  .  .
```

```
Pel 1391 LSVEVEEEEDTSWGIEEWSLPDTREDDEIEDQAEMDTRRDLDGIEYTHAIQDMIDLGISVDAQLNFLPESILINEIEGHDVDELDEAETELWELWMEEWDRHLLGISLQFALVRAMTPCI
Tob 1674 ------------------------DIDDSDDIDDSDNLDASD---------DIDRDLDTELK------LLTRMNGLTMDMMPEIDR------------FYITLQFELAKAMSPCI
Epi 1606 ------------------------VDDNDDNDSSDDIYASD---------DINSDLDTELE------LITMMNALTMDMMMLELDR------------FFTTLQLELAKAMSPWI
Spi 1523 ------------------------IDASD---------DIDRDLDTELE------LLTMMNALTMDMMMPEIDQ------------FSITLQFELAKAMSPCI
Mar 1439 ------------------------DVITE---------SWMNILIESL------------RRLNLTLDFAKKMSPCI
                                                                                              +.    .+ + +
```

```
                                              81 ♦ 82
Pel 1511 LWIPNIHDVDLE------DRT-TLA-GLM-NSLSGDSEGRSTRKTLVIASTHLPKKVDPALIASNRFNTCIKVRRLLSTQERERFFTLSYTRGFHFEKEIFDTNGFRSIT----TQDLAA
Tob 1738 IWIPNIHDLDVN------ESN-DLALGLLVNHLSRDCERCSTRNILVIASTHIPQKVDPALIAPNKLNTCIKIRRLLLPQQRKHFFTLSYTRGFHLEKKMFHTNGFGSITMGSNARDLVA
Epi 1669 IWIPNIHDLDVN------ESN-YLSFGLLVNHLS---ERCSTNNIIVIASTHIPKKVDPALLAPNKLNTCIKIRRLLIPQQRKHFCTLSYTRGFHLENKIFHTNGFGSITMGSSARDLVA
Spi 1574 IWIPNIHDLDVN------ESN-YLSLGLFVNYLSRDCERGSTRNILVIASTHIPQKVDPALIAPNQLNTCIKIRRLRIPQQRKHFFTLSYTRGFHLEKKMFHTNGFGSITMGSNVRDLVA
Mar 1471 IWIQNIHQLNVNRLTQNVESDPTFLLGILLKYFQTDFSKTKKNNIIVIGSTHLPKKVDPALISPNRLDKIINVRLFNISQRKKQFPLLLKKKNFQLKENLFFLNEFGSRTMGYNLRDLSA
         ++.+++.  .         .  +.  ...  + .   ++.+++ + +++++++ . +   ...+. +..    +   + .+.....+.   + .+.+ +.+++.   ++ +
```

```
                                         82 ♦ 83
Pel 1618 LTNEALSISITQKKSIIDTNTIRFALHRQTWAVESRSISILDH---GILFYQTGRALAQNLFLSQGLIDPISV--YIKKNWCN-DDGYSYLYRWYLELGTS---MKRLTILLYLLSCFAG
Tob 1851 LTNEVLSISITQKKSIIDTNTIRSALHRQTWDLRSQVRSVQDH---GILFYQIGRAVAQNVLLSNCPIDPISI--YMKKKSC--NEGDSYLYKWYFELGTS---MKRLTILLYLLSCSAG
Epi 1779 LTNEALSISITQNKSILDTNTIRSALHRQTWDLRSGVRSFQDN---GILSYQIGRAITQNVLLSNCPIDPISI--YMKKKSCTCNGGDYYFYKWYFGLGTS---MKKLTILLYLLSCSAG
Spi 1687 FINEALSISITQKKSIIDTNTIRSALHRQTWDLRSQVRSVQDH---GILFYQIGRAVAQNVLLSNCPIDPIST--YMKKKSC--NEGDSYLYKWYFELGTS---MKKLTILLYLLSCSAG
Mar 1591 LTNEVLLISITKNRSFIDTDTLKLAFHRQIFGLTYTNNKLNFDRIFKIVIYKVGKTIIQNILIKSSSMNLLNIGNFLWKKN------FYYLSKWYLEPSIDESIIKELTILTHILACLAG
         ++ +.+++. .+. ++.+.. +.+++..     .  .  .+. +.. ++   ....  ++ .... .+ ++++...+.+ ++
```

```
Pel 1729 SVAQDLWSPPGPDDKAEMFSYGLVENDSHLAQGLLEL---EGALVASG-TEKLSSRFDNDQV--TLLF---GRTKLDRMQDGFCSIFEQ-------------EGEEAAPGLEKPLVT
Tob 1961 SVAQDLWSLSGPDEKNGITSYGLVENDSDLVHGLLEV---EGALVGSSRTEKDCSQFDNDRV--TLLLRPEPRNPLDMMQNGSCSILDQRFLYEKYESEFEEGEGEGALDPQ---EDLFN
Epi 1891 SVAQDLWSLPGPAEKNGITSYGLVENDSDLVRGLLEV---EGALVGSSRTEKDCSPFDNDRVIFTLILRPEPGNPLDIIKKGSCSIFDHRFIYEKYESEFEEGYGEGALDPQQIEEDLFN
Spi 1797 SVAQDLWSLPGPDEKNGITSYGLVENDSYLVHGLLEV---EGALVGSSRIEKACSQ--NDRV--TLFLRPELRNPLDMMQNGSCSILDQRFLYEKYESELEEG--EGALDPQQIEEDLFN
Mar 1705 TAARDSWFLLEKKAESLLPIDKLVENDFTLAFSILESFFSEFPWLEICQTNVVNSK--KNKI--IEFS---TKNSMNIMQNGIFAIANKKFIYTQNHLQYKSSLSQ-QISFNKKKNYEFK
         ..+.+.+.  .. .  ...+++++. + ..++   +.... .. +....  ..  .. + ..+   .+ +....... ...  ...          .  .
```

```
Pel 1824 HIVSAPRIWNPWLI---LFDWIDIEDGSRKEEEAELQDEEAELQDEGAGRKDEEAELQEFQDLEDPQDQEGGLQELQGLQKEEAYLYQKALELQAQEDELQKYGPGFLESGTIMKKYPTR
Tob 2073 HIVWAPRIWRPWGF---LFDCIERPNELGFPYWSRSFRGKRIIYDEE--------------------------------------------------------DELQENDSEFLQSGTM--QYQT-
Epi 2008 HIVWAPRIWRPWGF---IFYCIERPNELGFPYWSRSFRGKRIVYDKD--------------------------------------------------------EEGELQENDSELLKSGTV--QYQT-
Spi 1908 HIVWAPRIWNPWGF---LFDCIERPNELGFPYWARSFRGKRTIYDKE--------------------------------------------------------DELQENDSEFLQSGTM--QYQT-
Mar 1817 NTSWSPRFWRLSFFRSNLFDWIKRPNDFEFSY--KFGFTKKKEYLFSANLQKKNNYGQFIEKKKKEQLLYERILPRIRRRNVQELESQFEEILLEEQFEILGFFRLSE-------QYPM-
         ...  .++.+ ...  + +         .            . +..  +++.   .          +++    +.+++  ..
```

```
Pel 1941 NRFRFFTEQGFFQASQFIWYPADSLFFLLKEQALGFVFSQPEFFADEEISK---------------------------------------------------------------EGLLALT
Tob 2137 -RDRSSKEQGLFRISQFIWDPADPLFFLFKDQPPGSVFSHRELFADEEMSK---------------------------------------------------------------GLLT
Epi 2074 -RDRSSKEQGLLKINQFIWDPADPLFFLLKDQPPGSVFSHRRFFADEEMSK---------------------------------------------------------------GLLT
Spi 1972 -RDRSSKEQGFFRISQFIWDPADPLFFLFKDQPFVSVFSHREFFADEEISK---------------------------------------------------------------GLLT
Mar 1927 -EYQLYNKPRLFIGKRILWDPIG-LFFQIRH----FVFSRREFFVDEEMLRRLYVTYGARRERERSRSSQKIKQFFLCRGYNKDLISKLSIRWWSQLPINEKKNIDTLKRIEHISIQLKR
         .. . ...    ...+ +.. +++. ...   +++    +.+++ ..
```

```
                                                                       83 ♦ 84
Pel 1999 VQRL-PFSTTCHWFIKKRQER--HFEFLIHRERWLRTNSSLS--NGFFCSKTQTLFESYQYLSNLFLSNGRLVDQMTKTLLRKRWIFPDEMKIQIGFMYTGEGFPITSPERYVAMK*
Tob 2191 SQTDPPTSIYKRWFIKNTQEK--HFELLINRQRWLRTNSSLS--NGSFRS--NTLSESYQYLSNLFLSNGTLLDQMTKTLLRKRWLFPDEMK--IGFM*
Epi 2128 SQKDPPTSIYKRWFIKNTQEQ--HFELLINRQRWLRTKSSLSKSNGSFRS--NTLFESYQYLSTLFLSNGTLFDKMTKTLLIKRWLFPDEMQ--M*
Spi 2026 SQMNPPISIFQRWFIKNTQEK--HFELLINRQRWLSTNSSLS--NGSFRS--NTLSESYQYLSNLFLSNGTLLDQMTKTLLRKRWLFPDEMK--IGFMQEEKDFPFLSRKDMWP*
Mar 2041 PQIFTPVYLYQRWLIENSPEKFFRFELLTHRKKWLKINSLLL--NDSFIY--TTLLEIYEYLLHFFIANKKLLNQMTKILLKKGWLFENEIE--TIINETRQ*
         + .+ .   +.+.  .+  .++ +. + .++.. +.+.  +. + .   ++ +.+.++. .+..+. + . +++.++ +.+ +..+.  .. ..  .
```

Fig. 1. Multiple alignment of the amino-acid sequences of ORF2280 homologs from *Pelargonium* × *hortorum* (ORF2109; this study), tobacco (ORF2280), *Epifagus virginiana* (ORF2216), spinach (ORF2131) and *Marchantia polymorpha* (ORF2136). *Plus signs* mark positions where all five sequences are identical (permitting gaps in one or two of the dicots). *Dots* mark positions where no differences (other than gaps in one or two species) are seen among the dicots. Blocks of four or more residues absolutely conserved in all five species are indicated by *boldface*. Sequences that are directly repeated within a particular protein are *underlined*. *Diamonds* indicate the positions of the five *Bam*HI sites and the one *Eae*I site that define the five ORF2280 probes (see Figs. 2 and 5)

scattered throughout the N-terminus of the protein, preventing complete deletion of the region. From Fig. 2 the part of the protein whose sequence is most likely to be functionally important is that between positions 1 750 and 2 200 in the multiple alignment (between, approximately, the conserved peptides TGRS and LVEND). The extreme C-terminus is also quite well-conserved (Figs. 1 and 2).

Different rates of sequence evolution

Pairwise amino-acid sequence identities between the five ORF2280 homologs are presented in Fig. 3 (top). High identity (81–91%) is evident between tobacco and *Epifagus* (see also Fig. 2, bottom), tobacco and spinach, and *Epifagus* and spinach. In contrast, amino-acid sequence similarity between the four angiosperms and *Marchantia* is marginal over much of the protein (averaging 27%), although several conserved domains approaching 50% identity can be found (Fig. 2).

The level of sequence conservation between the *Marchantia* ORF and the angiosperm ORFs (estimated time of divergence = 350–400 million years) is anomalously low relative to both the high conservation of this ORF within the angiosperms (tobacco-spinach divergence = approximately 100 million years) and typical levels (70–80% amino-acid identity) of divergence between *Marchantia* and angiosperm chloroplast gene products (Shimada and Sugiura 1991). A major reason for the first part of this anomaly is probably the gene's differential location with respect to the IR, a region whose apparent mutation rate is four-times lower than that of single-copy sequences (Wolfe et al. 1987). Although the molecular mechanisms responsible for maintaining the low frequency of mutations in IR sequences are not understood, the presence of two identical copies of this region in most angiosperm cpDNAs could imply the operation of some sort of gene conversion or copy-correction process (Palmer 1991; Birky and Walsh 1992). One would expect an accelerated rate of evolution for ORF2280 in *Marchantia* (where the

Fig. 2. Divergence of ORF2280 homolog sequences. Upper: schematic representation of a multiple alignment of ORF2280 homologs (Fig. 1), showing the positions of all gaps greater than or equal to five amino-acid residues. Approximate locations of deletions (*upright open triangles*) and insertions (*inverted shaded triangles*), postulated on the basis of filter hybridizations (Downie and Palmer, 1992 b, 1994), relative to the tobacco ORF and the five tobacco hybridization probes (shown in the topmost line) are as follows: a = 300–400-bp deletion in 12 families; b = 200-bp deletion in Nyctaginaceae and Phytolaccaceae; c = 500-bp insertion in Caprifoliaceae, Dipsacaceae, and Valerianaceae; d = 500-bp deletion in *Pereskia* (Cactaceae), *Portulaca* (Portulacaceae), Bignoniaceae, Convolvulaceae, and Solanaceae; and e = 500-bp insertion in Rosaceae. Lower: graphs of amino-acids sequence identity in tobacco versus *Epifagus* and tobacco versus *Marchantia* ORF comparisons, calculated using a sliding window of 100 non-gapped positions. Because of computer limitations, each point plotted represents the mean of five adjacent 100-residue windows. The sequences and approximate positions of the only blocks of greater than or equal to four residues absolutely conserved in all five species are indicated (cf. Fig. 1)

	Tobacco	Epifagus	Spinach	Marchantia
Pelargonium	68.5%	62.3%	66.2%	24.2%
Tobacco		84.3%	90.8%	27.8%
Epifagus			81.0%	27.5%
Spinach				28.3%

	ORF2280	rbcL	petD	rpl16	rps11
Tobacco	0%	0%	0%	0%	0%
Epifagus	16%	N/A	N/A	19%	17%
Spinach	9%	7%	3%	10%	10%
Pelargonium	32%	6%	2%	13%	29%
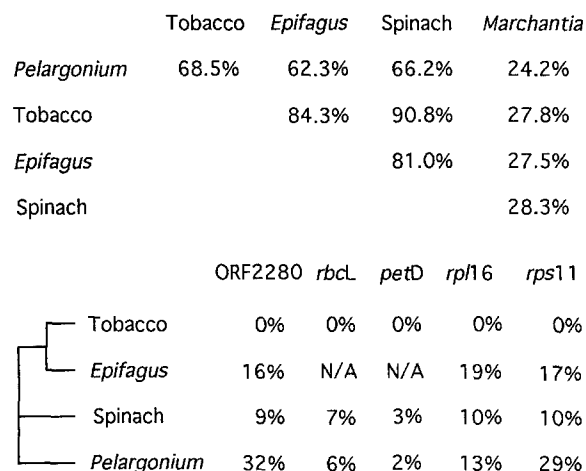
Fig. 3. Upper: pairwise amino-acid sequence identity (%) between ORF2280 homologs. Lower: phylogenetic evidence for accelerated evolution of ORF2109 in *Pelargonium* and ORF2216 in *Epifagus*. The *numbers* indicate the percentage amino-acid divergence, relative to tobacco, of five plastid genes. Genes *rbcL* and *petD* are not found in the severely-reduced plastid genome of the nongreen parasite, *Epifagus* (Wolfe et al. 1992; N/A = non applicable). The phylogeny is based on traditional data (i.e., Cronquist 1981) and molecular data (Chase et al. 1993) and shows that tobacco and *Epifagus* are closer relatives to each other than either is to spinach or *Pelargonium* and that the three major lineages illustrated are approximately equally unrelated

gene is single copy) compared to most angiosperms, where the gene is in the IR. Since ORF2280 appears to be primitively single copy in land plants (Palmer and Stein 1986; Palmer 1991; Raubeson and Jansen 1990), it is not surprising that the *Marchantia* ORF is so divergent from those of angiosperms. Another manifestation of the combination of a low mutation rate in the IR as a whole but a high rate of acceptance of mutations in ORF2280 is its remarkably greater level of amino-acid divergence (9% for tobacco-spinach) compared to nucleotide divergence (13%).

Virtually all classifications of angiosperms (Takhtajan 1980; Cronquist 1981; Thorne 1992) and all recent molecular phylogenies (Chase et al. 1993; C. dePamphilis, personal communication) place tobacco and *Epifagus* [both members of Cronquist's (1981) subclass Asteridae] as closer relatives to each other than either is to spinach (Caryophyllidae) or *Pelargonium* (Rosidae). The three subclasses are approximately equally unrelated, again according to both conventional and molecular data, and relationships among the four taxa are illustrated in Fig. 3 (bottom). Evidence that the markedly higher amino-acid divergence for the *Pelargonium*-tobacco comparison (32% as compared to only 9% for the equally unrelated spinach-tobacco comparison) is due to increased rates of molecular evolution in *Pelargonium* ORF2109, rather than to a decreased rate in tobacco (and spinach), comes from relative-rate

analyses using sequences from four other chloroplast genes (Fig. 3): *rbcL, petD, rpl16,* and *rps11* (the first two genes are not present in the severely-reduced plastid genome of the nonphotosynthetic plant *Epifagus,* hence comparisons here cannot be made). Amino-acid divergences for spinach-tobacco and *Pelargonium*-tobacco are approximately the same for *rbcL, petD,* and *rpl16.* Consequently, the high divergence of ORF2109 in the *Pelargonium*-tobacco comparison relative to most other pairwise comparisons shows that the rate of molecular evolution of this protein in *Pelargonium* is accelerated compared to ORF2280 [*rps11* also shows evidence of accelerated evolution in *Pelargonium,* although this is a far shorter protein (138 amino acids)]. The high rate of sequence evolution in *Pelargonium* ORF2109 is also apparent from a phylogenetic analysis (PAUP; D. Swofford, Illinois Natural History Survey) of the five land plant ORF2280 homolog sequences. In this analysis, the numbers of amino-acid or nucleotide substitutions assigned to the *Pelargonium* lineage are substantially higher than those for tobacco, spinach, or *Epifagus* (data not shown). This unusual sequence divergence in *Pelargonium* ORF2109, combined with its somewhat greater than usual amount of length mutation (Figs. 1 and 2), explains the relatively weak hybridization of tobacco ORF2280 probes to *Pelargonium* cpDNA which was the original motivation for sequencing the *Pelargonium* gene (see Introduction).

Although ORF2109 is located within the IR in cpDNA of *Pelargonium,* its specific location is unusual and it is conceivable that in an early *Pelargonium* lineage the gene was single copy, which would explain, at least in part, its accelerated evolution. In virtually all angiosperms, such as tobacco (Fig. 4), the other angiosperms mapped in Fig. 5, and over 100 other angiosperms (S. Downie and J. Palmer, unpublished data), the *trnI*-ORF2280-*trnL* gene cluster is situated between the *rpl23* operon and *ndhB.* However, as one of many gene rearrangements in *Pelargonium* cpDNA (Palmer et al. 1987), the ORF2280-*trnL* cluster has been "transposed" to a novel location between duplicated genes for *rpoA* and away from *rpl23* and *trnI* (Fig. 4). The highly-expanded size and rearranged gene order of the *Pelargonium* IR (Palmer et al. 1987), in conjunction with the absence of the IR in other members of Geraniaceae (Price et al. 1990; Price and Palmer 1993), raise the possibility that the IR was lost early in Gerania-

ceae evolution and then regained in *Pelargonium.* More data are needed to test this conjecture, but it would provide some explanation for the otherwise mystifyingly large divergence of ORF2109 in *Pelargonium.*

The higher identity between spinach-tobacco rather than between *Epifagus*-tobacco (Fig. 3), despite the latter pair being more-closely related phylogenetically, suggests that *Epifagus* ORF2216 is also evolving at an accelerated rate relative to the tobacco protein. This is in keeping with the accelerated evolution observed for virtually all genes that remain in the shrunken plastid genome of the parasite *Epifagus* (Wolfe et al. 1992 a, c).

Distribution in other angiosperms

Southern-blot hybridizations were used to survey for length variation in ORF2280 homologs from 279 species (representing 111 families; Table 1), including both dicots and monocots and representing angiosperm lineages that diverged from a common ancestor in the order of 150–200 million years ago (Wolfe et al. 1989). All but 18 of the 279 examined cpDNAs hybridized strongly to all five ORF2280-specific probes used (Table 2 and below). In addition, the construction of restriction-site maps of the entire IR for many of these taxa revealed that the position of their ORF2280 homologs is conserved within the IR and that their restriction fragments align readily with tobacco (Downie and Palmer 1992 b, 1994).

A 2949-bp *Bam*HI fragment from tobacco ORF2280 (probe 81 in Figs. 2 and 5) was used as a hybridization probe to illustrate the conservation of two *Bam*HI restriction sites, and the size of its intervening fragment, across a diversity of angiosperm species (Fig. 6). These taxa represent nine subclasses and 19 families of angiosperms, including both monocots and dicots. With the exception of six taxa (in lanes 3, 4, 13, 17 ,18, and 20), whose cpDNAs hybridized weakly or not at all, or exhibited variation in length, all cpDNAs hybridized to a conserved 2.9-kb *Bam*HI fragment. One end of this probe is located within the totally-conserved motif KVDPAL (Fig. 1), which helps explain why this restriction site is virtually always present. This probe failed to hybridize to cpDNA from *Passiflora helleri* (lane 13); hybridized weakly to cpDNAs from *Triticum aestivum* (lane 4), *Sarcocaulon vanderie-*

Pelargonium x hortorum
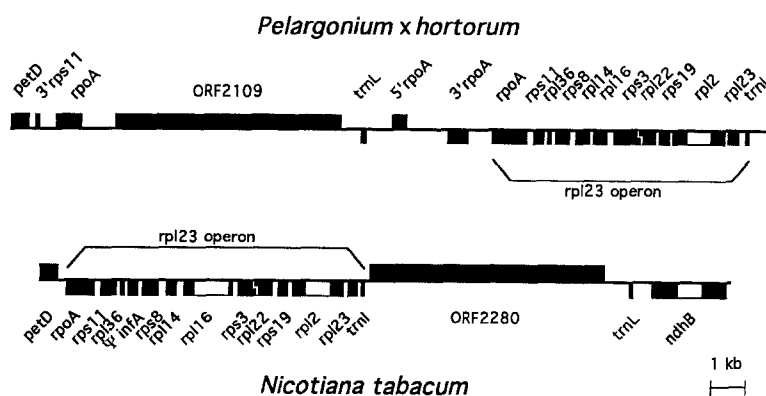
Fig. 4. Organization of chloroplast genes in the vicinity of ORF2280 homologs in *N. tabacum* (bottom) and virtually all other angiosperms investigated (S. Downie and J. Palmer, unpublished data) and *Pelargonium* × *hortorum* (top). In *Pelargonium* cpDNA, ORF2109 plus *trnL* have been "transposed" to a novel location between duplicated genes for *rpoA.* Genes transcribed from left to right are shown above the two lines. The *Nicotiana* map is based on Shinozaki et al. (1986); *Pelargonium* data are based on the hybridization experiments reported in this study and on sequencing much of the region shown (P. Calie et al. unpublished data). The pseudogene *infA* in tobacco is marked "ψ"

Passiflora

BamHI	2.8		.4	.9	.8	4.0
EcoRV	2.6		2.6			3.3
Bgl II	5.2	1.8		.2	.8	2.7
HindIII	4.8					

Sarcocaulon

BamHI	5.5		.4	2.2	2.4
HindIII	12		5.1		
SacI	9.4		7.0		
BstXI	11			6.0	

Zea

BamHI	5.1	1.5	4.9
EcoRV	9.4	.2	3.0
Bgl II	4.3	.4	1.5	1.7
HindIII	4.5		2.5
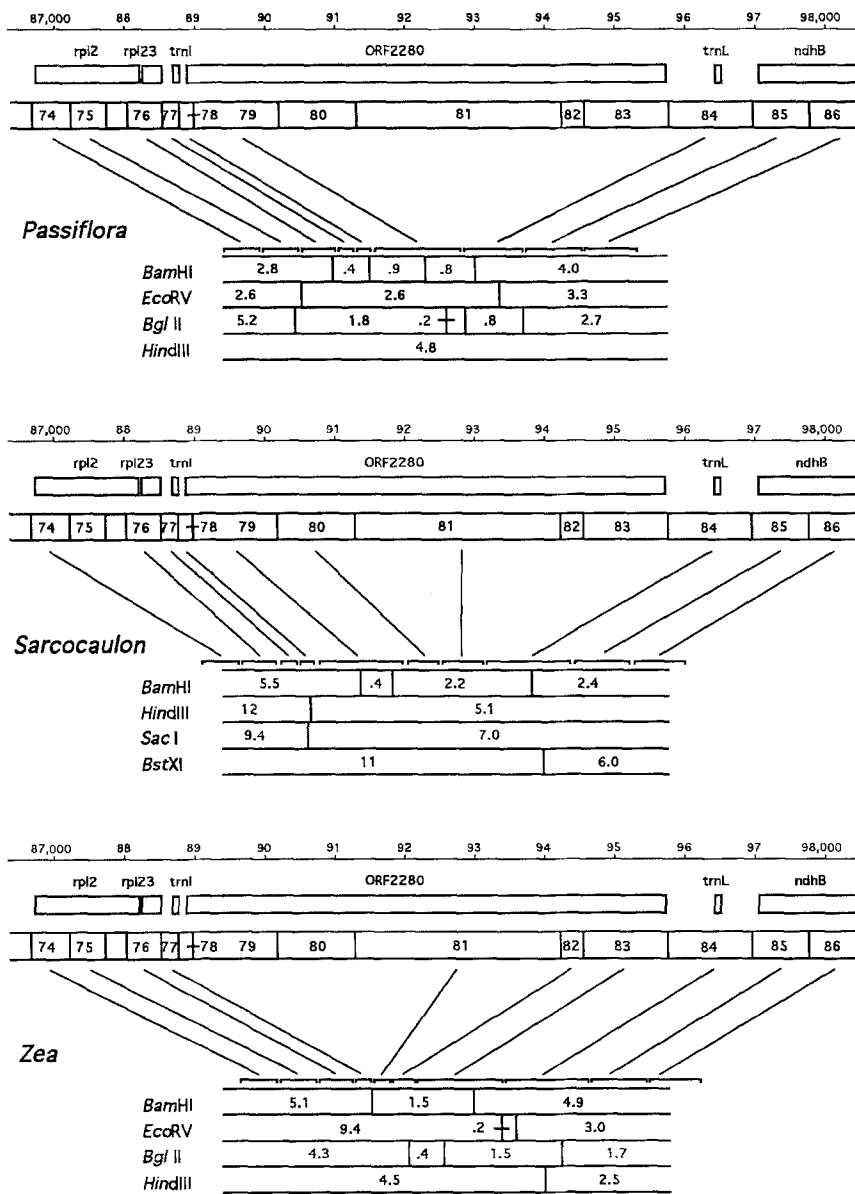
Fig. 5. Restriction site and gene maps in the vicinity of ORF2280 for cpDNAs of *Passiflora helleri*, *Sarcocaulon vanderietiae*, and *Zea mays* showing major deletions in their ORF2280 homolog sequences. Cloned restriction fragments used as hybridization probes are numbered from 74 to 86. Coordinates and restriction sites for the five ORF2280 probes from tobacco cpDNA (Shinozaki et al. 1986) are as follows: *79* = 88992 (*Bam*HI) –90181 (*Bam*HI); *80* = 90182 (*Bam*HI) –91293 (*Bam*HI); *81* = 91294 (*Bam*HI) –94242 (*Bam*HI); *82* = 94243 (*Bam*HI) –94562 (*Bam*HI); *83* = 94563 (*Bam*HI) –95764 (*Eae*I). The *square brackets* above the restriction maps indicate the sets of fragments to which each probe hybridized

tiae (lane 17), *Erodium chamaedryoides* (lane 18), and *Campanula garganica* (lane 20); and hybridized to a considerably smaller *Bam*HI fragment in *Zea mays* (lane 3); the implications and significance of these hybridization results are discussed below. *Polygonum persicaria* (lane 10) cpDNA was underloaded relative to most other cpDNAs; it hybridized weakly to probe 81 and to a 2506-bp *Bam*HI fragment from tobacco containing portions of chloroplast genes *psaB* and *psaA* (coordinates 40920–43426 in Shinozaki et al. 1986). The latter probe was used as a positive control, as it hybridized to all 21 cpDNAs.

Among most other angiosperm cpDNAs, conservation in restriction sites and fragment lengths was also revealed using the other enzymes and the four other probes specific for tobacco ORF2280, but not to the extent exhibited by the 2949-bp probe (data not shown). Restriction site mapping studies of cpDNA IR sequences for 99 species of Asteridae and related taxa (Downie and Palmer 1992 b) and 24 species of Caryophyllidae (Downie and Palmer 1994),

using four or ten restriction enzymes, respectively, detected five regions showing length variation of greater than 200 bp in ORF2280 homologs. The approximate location and extent of these length variants are illustrated in Fig. 2. These length variants, representing three deletions and two insertions, ranged between 200 and 500 bp in size and were found in representatives of 17 families (Table 1). Two of these events, deletions a and d, are estimated to have occurred several times independently (minimally nine and four times, respectively; Downie and Palmer 1992 b, 1994). Deletions in ORF2280 homologs from five families of Hamamelidae and Rosidae are also known, but their precise sizes have not been reported (Manos et al. 1993). This conservation in restriction enzyme recognition sites and fragment sizes in ORF2280 homolog sequences, indeed in the whole cpDNA IR region, has permitted the use of these characters to infer phylogeny within subclasses of angiosperms (e.g., Downie and Palmer 1992 b, 1994; Manos et al. 1993).
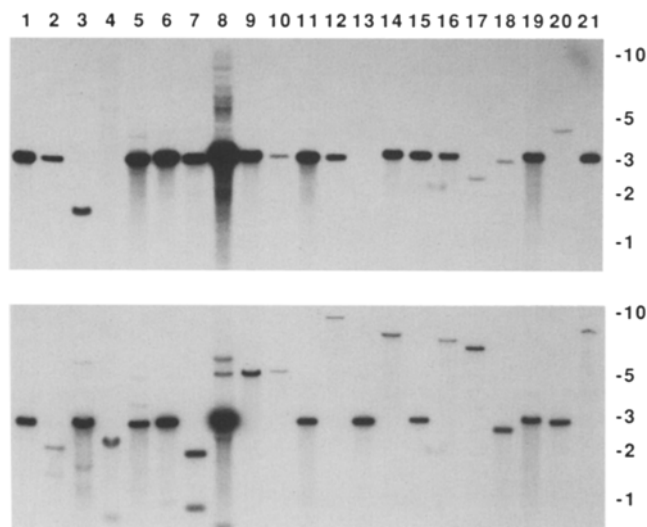
Fig. 6. Conservation or absence of a 2.9-kb ORF2280 *Bam*HI fragment from angiosperm cpDNAs. CpDNA fragments produced by digestion of 21 angiosperm cpDNAs with *Bam*HI were electrophoresed in a 1.0% agarose gel, blotted, and hybridized with either a 2.949-kb *Bam*HI fragment from tobacco internal to ORF2280 (probe 81; top panel) or a 2.506-kb *Bam*HI fragment from tobacco specific for portions of chloroplast genes *psaB* and *psaA* (coordinates 40920–43426 in Shinozaki et al. 1986; bottom panel). The identification of taxa in each lane is as follows: *1, Erythronium albidum* (Liliaceae); *2, Sagittaria latifolia* (Alismataceae); *3, Zea mays* (Poaceae); *4, Triticum aestivum* (Poaceae); *5, Aristolochia durior* (Aristolochiaceae); *6, Podophyllum peltatum* (Berberidaceae); *7, Ulmus americana* (Ulmaceae); *8, Chenopodium murale* (Chenopodiaceae); *9, Agrostemma githago* (Caryophyllaceae); *10, Polygonum persicaria* (Polygonaceae); *11, Brassica juncea* (Brassicaceae); *12, Viola* sp. (Violaceae); *13, Passiflora helleri* (Passifloraceae); *14, Apium graveolens* (Apiaceae); *15, Fuchsia hybrida* (Onagraceae); *16, Hydrangea* sp. (Hydrangeaceae); *17, Sarcocaulon vanderiatiae* (Geraniaceae); *18, Erodium chamaedryoides* (Geraniaceae); *19, Clytostoma callistegioides* (Bignoniaceae); *20, Campanula garganica* (Campanulaceae); and *21, Barnadesia caryophylla* (Asteraceae). Numbers at right of panels indicate fragment sizes in kb

Major deletions within the gene, as judged by the failure to obtain hybridization signals under standard stringency conditions using probes that span 1.2–5.6 kb of tobacco ORF2280, were detected in representatives from four families: *Passiflora helleri* (Passifloraceae); three genera (five species) of Geraniaceae; two genera (four species) of Campanulaceae; and eight species of Poaceae (Table 2). Restriction-mapping studies using four restriction enzymes (Fig. 5) confirmed these deletions in *Passiflora*, *Sarcocaulon* (Geraniaceae), and *Zea* (Poaceae). The deletion within the ORF2280 homolog of *Passiflora* appears largest, with almost 6 kb of sequence missing with respect to the comparable region in tobacco, while those in *Sarcocaulon* and *Zea* are approximately 4–5 kb in size. The apparent deletions in Campanulaceae cpDNAs could not be corroborated by the mapping studies owing to extensive structural rearrangement within this region (Downie and Palmer 1992 a). Thus, in this case, we cannot rule out the possibility that these genomes contain normal-sized ORF2280 homologs that are, however, abnormally divergent in sequence.

The four groups of plants exhibiting major deletions in ORF2280 homolog sequences (i.e., those taxa belonging to angiosperm families Passifloraceae, Geraniaceae, Campanulaceae, and Poaceae) represent a diverse array of plants that have been assigned to four subclasses of angiosperms. These groups are distantly related to one another in systems of classification based largely on morphology (e.g., Takhtajan 1980; Cronquist 1981; Thorne 1992) and in phylogenies based on *rbcL* sequences (Chase et al. 1993). Moreover, ORF2280 mapped as if it were intact in each of the closest relatives (as implied by *rbcL* phylogenies; Chase et al. 1993) for these four groups that were included in the survey (i.e., ORF2280 is intact in *Pelargonium*, the closest relative examined of *Sarcocaulon*, *Erodium* and *Geranium*; in *Euphorbia*, the closest relative of *Passiflora*; in *Campanula ramosa* and *Platycodon grandiflorus* (Campanulaceae), the closest relatives of the four Campanulaceae species; and in Commelinaceae, the closest relative of Poaceae).

On the basis of these data, we infer that major deletions within the gene have occurred at least four times independently during the evolution of angiosperms. A fifth putative independent deletion may have occurred in the plastid DNA (ptDNA) of the holoparasitic *Cuscuta reflexa*, where a region of residual ORF2280 sequence, recognized as ORF740, has recently been reported (Bömmer et al. 1993). In *Cuscuta*, ORF740 is adjacent to an apparent major deletion of approximately 6.5 kb corresponding, in tobacco, to genes *rpl2*, *rpl23* and *trnI*, and a region encoding 1540 residues of the 5' end of ORF2280 (Bömmer et al. 1993). The reported deletion, however, occurs precisely at the tobacco IR-large single copy (LSC) junction and could also be explained by a contraction of the IR. If the *Cuscuta* IR-LSC junction were to fall within ORF2280, the region containing genes *rpl2*, *rpl23* and *trnI*, along with the 5' end of ORF2280, could be found in the LSC region at the other end of the IR and, thus, would be overlooked if only one IR-LSC junction was sequenced. Although Bömmer et al (1993) present Southern blots of tobacco and *Cuscuta* total cellular DNA hybridized against two probes from tobacco specific for genes *rpl2-rpl23* and the 5' end of ORF2280, they fail to provide a positive control showing the presence of significant levels of *Cuscuta* ptDNA on these blots. Without a complete restriction map for *Cuscuta* ptDNA, or more rigorous Southern hybridizations, we are somewhat skeptical about the loss of this region from *Cuscuta* ptDNA. *Cuscuta* is often allied with Convolvulaceae either as a member of that family or else treated as a separate family, Cuscutaceae (Asteridae; Cronquist 1981). The three genera of Convolvulaceae examined in this investigation (*Calystegia*, *Convolvulus*, and *Ipomoea*) all possess intact ORF2280 homologs.

Rice pseudogene

Several deletions in the rice chloroplast genome, totalling about 6 kb in length, coincide with the region occupied by ORF2280 in tobacco. Three short ORFs (ORF28, ORF64 and ORF249) were previously reported to occupy this same region in rice (Hiratsuka et al. 1989; Shimada and

376

83402
|
Rice GATCCAATTCTATTGAGTCTGACTCATAGTGATCATTTCTCTTTAGCAAAGAATGACCTTGGTTATCAAAGGATTGAACAACCGGGATCCGTTTAC
Tob GATCAAATTCTATTGAGTCTGACTCATAGTGATCGT------TTATCAAAGAATGACTCTGGTTATCAAATGATTGAACAACCGGGAGCAATTTAC
        D Q I L L S L T H S D R   L S K N D S G Y Q M I E Q P G A I Y
|
93469
                                          ↓ ORF28 start
                                          fM N Y E F N R S S L A E R R I F L A H
TTATGATACCTAGTTGACATTGATAACAAGGATCTAATGAATTATGAGTTTAATAGATCCTCTTTAGCAGAAAGACGTATATTCCTTGCTCAT------
TTACGATACTTAGTTGACATTCATAAAAAGCATCTAATGAATTATGAGTTCAATCCATCCTGTTTAGCAGAAAGACGGATATTCCTTGCTCATTATCAG
L R Y L V D I H K K H L M N Y E F N P S C L A E R R I F L A H Y Q

                     ↓ ORF28 stop
         h l f p n l v w g *
-----CACTTATTCCCAAACCTCGTATGGGGCTAATCGTTTTCATTTACCATCTCAGGGAAAATCCTTTTCGTTCCGCTTAGCCCTATC--------G
ACAATCACTTATTCACAAACTTCGTGTGGGGAAAATAGTTTTCATTTCCCATCTCATGGAAAACCCTTTTCGCTCCGCTTAGCCTTATCCCCCTCTAGG
T I T Y S Q T S C G E N S F H F P S H G K P F S L R L A L S P S R
    ↓ ORF64 start   83765
      fM I G S I G T G R S Y L V K Y L M K N S Y F P F I K V|
GGTATTTTAGTGATAGGTTCTATAGGAACTGGACGATCCTATTTGGTCAAATACCTAATGAAAAATTCCTATTTTCCTTTCATTAAGGTA
GGTATTTTAGTGATAGGTTCTATAGGAACTGGACGATCCTATTTGGTCAAATACCTAGCGACAAACTCCTATGTTCCTTTCATTACGGTA (434 bp)
G I L V I G S I G T G R S Y L V K Y L A T N S Y V P F I T V|
  83766   93846
     |R G L L I P Q E K N e s t f s f f h i l g v f t w k r q c s
   CGAGGGCTTCTTATTCCACAAGAAAAGAACGAAAGCACCTTTTCATTCTTTCATATACTGGGGGTTTTTACTTGGAAAAGACAATGTTC
CATTAAGATACGAAGGCTTCTTCTTCCACAACAA-----CGAAAGCACTTTTTCACTCTTTCATATACTAGGGGATTTCACTTGGAAAAGAAAATGTTC
I K I |R R L L L P Q Q   R K H F F T L S Y T R G F H L E K K M F
    94291   ↓ ORF64 stop   83929
                                                                |
    i l k d s g p *
CATACTAAAGGATTCGGGTCCATAACCACGAGTTCCAGTGCACTAGATCTTGTAGCACTTAGCAACGAGGCCCTA
CATACTAACGGATTCGGGTCCATAACCATGGGTTCCAATGCACGAGATCTTGTAGCACTTACCAATGAGGTCCTA
H T N G F G S I T M G S N A R D L V A L T N E V L|
                                                 94449

Fig. 7. Comparison between rice (top) and tobacco (bottom) sequences illustrating high nucleotide and amino-acid similarity in a portion of the region occupied by tobacco ORF2280. The rice sequence is from Hiratsuka et al. (1989) and represents coordinates 83402–83929; the tobacco sequence is from Shinozaki et al. (1986) and represents coordinates 93469–94449. The region between nucleotides 93846 and 94281 in tobacco has been omitted; this 434-bp region is flanked by a 10-bp imperfect direct repeat (underlined). Only one copy of this repeat is present in rice cpDNA. Four smaller gaps ranging in size from 5 to 11 bp are necessary to increase sequence identity between rice and tobacco

Sugiura 1991), but only the first two are homologous to ORF2280. Comparisons between tobacco and rice sequences, however, reveal greater identity than reported previously. In rice, a region of 528 bp (coordinates 83402–83929 in Hiratsuka et al. 1989) bears 90% DNA identity to tobacco ORF2280 (Fig. 7). This recognition of conserved sequence results from the insertion of five gaps, the largest, of 444 bp, being between rice coordinates 83765 and 83766. In the vicinity of the 444-bp gap in rice, the tobacco sequence contains two copies of a small imperfect direct repeat, whereas rice (coordinates 83756–83765) has only one copy of this repeat and has lost all the DNA between them. The other four gaps are much smaller, ranging in size from 5 to 11 bp. No substantial sequence identity was found between ORF249 and any part of tobacco ORF2280. The previously-recognized (Hiratsuka et al. 1989; Shimada and Sugiura 1991) ORFs 28 and 64 in this region are almost certainly fortuitous and nonfunctional: note their short length, the presence of internal frame shifts (of 11 bp in ORF28, and of 5 bp in ORF64), and of an immense (relative to the size of the ORF) insertion/deletion of 444 bp in ORF64 (Fig. 7).

Although rice was not included in our survey to uncover structural variation in ORF2280 homologs, it is expected (based on the available sequence data) that rice cpDNA should not hybridize to tobacco probes 79, 80, and 83 and should hybridize only weakly to probe 81 (Fig. 2). However, since probe 83 did hybridize to all eight species of Poaceae examined (Table 2), the rice pseudogene is probably smaller than that present in the other grass species.

Possible gene transfer

The cyanobacterial affinities of many nuclear genes that encode chloroplast proteins suggests that these genes were transferred from the chloroplast to the nucleus at some stage after the endosymbiotic origin of the chloroplast (Palmer 1991). Several known gene-content differences among cpDNAs of land plants and green algae suggest that this process continues to occur, with a few of these gene transfer events having been well-characterized. For example, the gene *tufA* is found in most green algal cpDNAs examined but is absent from all land plant plastid DNAs, implying that the gene was transferred from the chloroplast to the nucleus during green algal evolution, prior to the origin of land plants (Baldauf and Palmer 1990; Baldauf et al. 1990). A more recent case of gene transfer involves *rpl22*, which is plastid-localized in most land plants but nuclear-localized in legumes (Gantt et al. 1991).

The conservation of ORF2280 homolog sequences among most examined land plants – including *Epifagus* and *Pelargonium*, two species that exhibit rapidly evolving plastid genomes and have sustained numerous gene deletions – implies that the gene must be evolving under constraint, presumably because it is performing some necessary function. However, in at least four independent lineages of angiosperms, major deletions have occured within the gene. These deletions remove about 4–6 kb of this approximately 6.5-kb gene in members of Passifloraceae, Geraniaceae, Cuscutaceae, and Poaceae, while the Campanulaceae deletion is less-clearly defined and requires further mapping. Such massive deletions indicate that only

Table 2. Angiosperms exhibiting major length variation in the chloroplast gene ORF2280. The failure or success of obtaining hybridization signals, using five probes (numbered 79–83; see Fig. 5) that together comprise virtually the entire *N. tabacum* ORF2280 gene, is indicated as "–" or "+", respectively; wk = weak hybridization signals. Putative deletions in *P. helleri*, *S. vanderietiae*, and *Z. mays* have been confirmed by the construction of restriction-site maps for each of these taxa (see Fig. 5). For Campanulaceae and Poaceae only one representative taxon is listed; the other three Campanulaceae species examined are *Campanula kemulariae*, *C. lactiflora*, and *Jasione montana*, and the other seven Poaceae species examined are *Agropyron repens*, *Avena sativa*, *Bambusa* sp., *Danthonia spicata*, *Dendrocalamus* sp., *Secale cereale*, and *T. aestivum*

Species	Probe				
	79	80	81	82	83
Passifloraceae					
Passiflora helleri	+	–	–	–	–
Geraniaceae					
Erodium ciconium	+	+	wk	–	–
Geranium grandiflorum	–	+	+	+	+
Geranium nervosum	–	+	+	+	+
Sarcocaulon vanderietiae	+	+	wk	–	–
Sarcocaulon crassicaule	+	+	wk	–	–
Campanulaceae (four species; all show same hybridization pattern)					
Campanula garganica et al.	–	–	wk	+	wk
Poaceae (eight species; all show same hybridization pattern)					
Zea mays et al.	–	–	wk	+	+

pseudogenes of ORF2280 are likely to remain in these plastid genomes (this is clear in the case of the sequenced rice genome), and therefore it is likely that a functional copy of ORF2280 exists in the nucleus in these plants. The relocation of gene function from the plastid to the nucleus occurs both by gene transfer (see above) and gene substitution, in which a preexisting, primordially nuclear gene takes over the function of a lost chloroplast gene (Palmer 1991). Once a chloroplast gene becomes redundant due to gene transfer or substitution, the deletion processes and pressures that maintain the compact organization of most plastid genomes (Palmer 1991) would be expected to lead to the gene's rapid disintegration. In some cases, as with the ribosomal protein and tRNA pseudogenes of certain angiosperms (Zurawski and Clegg 1987; Wolfe et al. 1992 c), these processes may be fairly subtle and detectable only by DNA sequencing. In other cases, as reported here, major deletions can occur. However, as with the *rpl22* gene losses from pea and *Epifagus*, and the *infA* losses from *Pelargonium* and tobacco (Wolfe et al. 1992 c), we are faced with an enigma of multiple losses of a gene in independent angiosperm lineages. These may be indicative of ancient gene transfers to the nucleus (in which case the maintenance of the gene in the remaining lineages of angiosperms is difficult to explain), or alternatively may indicate independent gene transfers or activations of a nuclear gene (which is an unparsimonious explanation involving multiple occurrences of a rare and difficult event). A third possibility is that ORF2280 encodes an enzyme that functions in a biochemical pathway that has simply been lost in some plant lineages. A search for sequences homolo-

gous to ORF2280 in the nucleus is in order, but will be difficult due to the relatively-rapid evolution of its gene product (Figs. 1 and 2).

References

Baldauf SL, Palmer JD (1990) Evolutionary transfer of the chloroplast *tufA* gene to the nucleus. Nature 344: 262–265

Baldauf SL, Manhart JR, Palmer JD (1990) Different fates of the chloroplast *tufA* gene following its transfer to the nucleus in green algae. Proc Natl Acad Sci USA 87: 5317–5321

Birky CM Jr, Walsh JB (1992) Biased gene conversion, copy number, and apparent mutation rate differences within chloroplast and bacterial genomes. Genetics 130: 677–683

Blasko K, Kaplan SA, Higgins KG, Wolfson R, Sears BB (1988) Variation in copy number of a 24-base pair tandem repeat in the chloroplast DNA of *Oenothera hookeri* strain Johansen. Curr Genet 14: 287–292

Bömmer D, Haberhausen G, Zetsche K (1993) A large deletion in the plastid DNA of the holoparasitic flowering plant *Cuscuta reflexa* concerning two ribosomal proteins (*rpl2*, *rpl23*), one transfer RNA (*trnI*) and an ORF2280 homologue. Curr Genet 24: 171–176

Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, Mishler BD, Duvall MR, Price RA, Hills HG, Qiu Y-L, Kron KA, Rettig JH, Conti E, Palmer JD, Manhart JR, Sytsma KJ, Michaels HJ, Kress WJ, Karol KG, Clark WD, Hedrén M, Gaut BS, Jansen RK, Kim K-J, Wimpee CF, Furnier GR, Strauss SH, Xiang Q-Y, Plunkett GM, Soltis PS, Swensen SM, Williams SE, Gadek PA, Quinn CJ, Eguiarte LE, Golenberg E, Learn GH Jr., Graham SW, Barrett SCH, Dayanandan S, Albert VA (1993) Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. Ann Missouri Bot Gard 80: 528–580

Cronquist A (1981) An integrated system of classification of flowering plants. Columbia University Press, New York

Downie SR, Palmer JD (1992 a) Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. In: Soltis PS, Soltis DE, Doyle JJ (eds) Molecular systematics of plants. Chapman and Hall, New York

Downie SR, Palmer JD (1992 b) Restriction site mapping of the chloroplast DNA inverted repeat: a molecular phylogeny of the Asteridae. Ann Missouri Bot Gard 79: 266–283

Downie SR, Palmer JD (1994) A chloroplast DNA phylogeny of the Caryophyllales based on structural and inverted repeat restriction site variation. Syst. Bot 19(2): (in press)

Doyle JJ, Doyle JL (1987) A rapid DNA isolation for small quantities of fresh tissue. Phytochem Bull 19: 11–15

Gantt JS, Baldauf SL, Calie PJ, Weeden NF, Palmer JD (1991) Transfer of *rpl22* to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron. EMBO J 10: 3073–3078

Glick RE, Sears BB (1993) Large unidentified open reading frame in plastid DNA (ORF2280) is expressed in chloroplasts. Plant Mol Biol 21: 99–108

Herdenberger F, Weil J-H, Steinmetz A (1988) Organization and nucleotide sequence of the broad bean chloroplast genes *trnL-UAG*, *ndhF* and two unidentified open reading frames. Curr Genet 14: 609–615

378

Higgins DG, Sharp PM (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73: 237–244

Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, Kondo C, Honji Y, Sun C-R, Meng B-Y, Li Y-O, 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 tRNAs accounts for a major plastid DNA inversion during the evolution of the cereals. Mol Gen Genet 217: 185–194

Igloi GL, Meinke A, Döry I, Kössel H (1990) Nucleotide sequence of the maize chloroplast $rpoB/C_1/C_2$ operon: comparison between the derived protein primary structures from various organisms with respect to functional domains. Mol Gen Genet 221: 379–394

Manos PS, Nixon KC, Doyle JJ (1993) Cladistic analysis of restriction site variation within the chloroplast DNA inverted repeat region of selected Hamamelididae. Syst Bot 18: 551–562

Nimzyk R, Schöndorf T, Hachtel W (1993) In-frame length mutations associated with short tandem repeats are located in unassigned open reading frames of Oenothera chloroplast DNA. Curr Genet 23: 265–270

Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, Sano S, Umesono K, Shiki Y, Takeuchi M, Chang Z, Aota S, Inokuchi H, Ozeki H (1986) Chloroplast gene organization deduced from complete sequence of liverwort Marchantia polymorpha chloroplast DNA. Nature 322: 572–574

Palmer JD (1986) Isolation and structural analysis of chloroplast DNA. Methods Enzymol 118: 167–186

Palmer JD (1991) Plastid chromosomes: structure and evolution. In: Bogorad L, Vasil IK (eds), The molecular biology of plastids, vol 7A. Academic Press, San Diego, California, pp 5–53

Palmer JD, Stein DB (1986) Conservation of chloroplast genome structure among vascular plants. Curr Genet 10: 823–833

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

Palmer JD, Jansen RK, Michaels HJ, Chase MW, Manhart JR (1988) Chloroplast DNA variation and plant phylogeny. Ann Missouri Bot Gard 75: 1180–1206

Price RA, Palmer JD (1993) Phylogenetic relationships of the Geraniaceae and Geraniales from rbcL sequence comparisons. Ann Missouri Bot Grad 80: 661–671

Price RA, Calie PJ, Downie SR, Logsdon JM Jr, Palmer JD (1990) Chloroplast DNA variation in the Geraniaceae. In: Vorster P (ed) Proc Inter Geraniaceae symposium. Stellenbosch, South Africa

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

Richards CM, Hinman SB, Boyer CD, Hardison RC (1991) Survey of plastid RNA abundance during tomato fruit ripening: the amounts of RNA from the ORF2280 region increase in chromoplasts. Plant Mol Biol 17: 1179–1188

Shimada H, Sugiura M (1991) Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes. Nucleic Acids Res 19: 983–995

Shimada H, Fukuta M, Ishikawa M, Sugiura M (1990) Rice chloroplast RNA polymerase genes: the absence of an intron in rpoC1

and the presence of an extra sequence in rpoC2. Mol Gen Genet 221: 395–402

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, Todoh N, Shimada H, Sugiura M (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J 5: 2043–2047

Smith AG, Wilson RM, Kaethner TM, Willey DL, Gray JC (1991) Pea chloroplast genes encoding a 4-kDA polypeptide of photosystem I and a putative enzyme of C_1 metabolism. Curr Genet 19: 403–410

Sugiura M, Shinozaki K, Zaita N, Kususa M, Kumano M (1986) Clone bank of the tobacco (Nicotiana tabacum) chloroplast genome as a set of overlapping restriction endonuclease fragments: mapping of eleven ribosomal protein genes. Plant Sci 44: 211–216

Takhtajan AL (1980) Outline of the classification of flowering plants (Magnoliophyta). Bot Rev 46: 225–359

Thorne RF (1992) Classification and geography of the flowering plants. Bot Rev 58: 225–348

Weglöhner W, Subramanian AR (1991) A heptapeptide repeat contributes to the unusual length of chloroplast ribosomal protein S18. FEBS Lett 279: 193–197

Wolfe KH (1994) Similarity of plastid ORF2280 to the CDC48 family: a proteolytic ATPase?. Curr Genet 25: 379–383

Wolfe KH, Li W-H, Sharp PM (1987) Rates of nucleotide substitutions vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci USA 84: 9054–9058

Wolfe KH, Gouy M, Yang Y-W, Sharp PM, Li W-H (1989) Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc Natl Acad Sci USA 86: 6201–6205

Wolfe KH, Morden CW, Palmer JD (1991) Ins and outs of plastid genome evolution. Curr Opin Genet Dev 1: 523–529

Wolfe KH, Morden CW, Palmer JD (1992 a) Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc Natl Acad Sci USA 89: 10648–10652

Wolfe KH, Morden CW, Palmer JD (1992 b) Small single-copy region of plastid DNA in the non-photosynthetic angiosperm Epifagus virginiana contains only two genes. Differences among dicots, monocots and bryophytes in gene organization at a non-bioenergetic locus. J Mol Biol 223: 95–104

Wolfe KH, Morden CW, Ems SE, Palmer JD (1992 c) Rapid evolution of the plastid translational apparatus in a nonphotosynthetic plant: loss or accelerated sequence evolution of tRNAs and ribosomal protein genes. J Mol Evol 35: 304–317

Zhou D-X, Massenet O, Quigley F, Marion MJ, Moneger F, Huber P, Mache R (1988) Characterization of a large inversion in the spinach chloroplast genome relative to Marchantia: a possible transposon-mediated origin. Curr Genet 13: 433–439

Zurawski G, Clegg MT (1987) Evolution of higher-plant chloroplast DNA-encoded genes: implications for structure-function and phylogenetic studies. Annu Rev Plant Physiol 38: 391–418