Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 195–199, January 1990
Botany

Defective chlorophyll a/b-binding protein genes in the genome of a homosporous fern

E. PICHERSKY*†, D. SOLTIS‡, AND P. SOLTIS‡

*Department of Biology, University of Michigan, Ann Arbor, MI 48109; and ‡Department of Botany, Washington State University, Pullman, WA 99164

Communicated by Warren H. Wagner, October 13, 1989

ABSTRACT **The majority of homosporous ferns have a chromosome number that is severalfold greater than that of diploid seed plants. These fern species have therefore been generally considered to be of polyploid origin. Enzyme electrophoretic investigations have demonstrated, however, that within fern genera, species having the lowest chromosome numbers (n = 27–52) have the number of isozymes typical of diploid seed plants; there is no isozyme evidence for polyploidy of these plants. We have constructed a genomic DNA library from *Polystichum munitum* (n = 41), a homosporous fern, and have screened the library for sequences homologous to the chlorophyll a/b-binding (CAB) protein genes of higher plants. The majority of the sequences isolated and characterized by nucleotide sequence determination represent defective CAB genes. This result is in contrast to the situation in the genomes of diploid angiosperms, where most, and sometimes all, copies of the CAB gene family represent functional members. Several hypotheses could explain the existence of multiple defective CAB genes in *P. munitum*. (*i*) The defective CAB genes are the result of "gene silencing" following polyploidy. (*ii*) *P. munitum* has not gone through a polyploidization event, but several, and perhaps the majority, of its CAB genes have mutated to a nonfunctional state (a phenomenon not yet observed in any of the genomes of non-fern plants so far examined). (*iii*) Some defective CAB genes have been specifically amplified in the genome of *P. munitum*.**

The average haploid chromosome number for homosporous ferns, 57 (1), is so high that up to 95% of all species have been considered polyploids (2, 3). However, it has been shown (4, 5) that within fern genera, species having the lowest chromosome numbers (n = 27–52) have the number of isozymes typical of diploid seed plants (6). These original observations have now been expanded to include all the major groups of homosporous pteridophytes (reviewed in ref. 7). Therefore, these plants either have gone through repeated cycles of polyploidy and gene silencing or were initiated with high chromosome number. If these species are ancient polyploids that have "diploidized" by the accumulation of mutations in the extra copies of genes so that only one homeologous locus is now expressed, one would expect to find in these fern genomes the remnants of the mutated and silenced genes.

The chlorophyll pigment molecules, the primary receptors of light energy in the two photosystems (PSI and PSII) of eukaryotic photosynthetic organisms, are anchored in the thylakoid membranes of the chloroplast through their interaction with a set of proteins known as the chlorophyll a/b-binding proteins (CAB proteins). In many angiosperms, both dicots and monocots, the CAB proteins are encoded by a set of structurally and evolutionarily related nuclear genes, i.e., by a "gene family" (8). So far, genes encoding six types of CAB protein have been identified in angiosperm species (see relevant citations in ref. 8). Sequence comparisons revealed that these six branches of the CAB gene family arose by gene duplication and divergence prior to the radiation of the angiosperms. Within each branch of the CAB family, more recent duplications have also occurred, so that the same type of CAB polypeptide is sometimes encoded by more than one gene. Unlike the situation found in many gene families in animals, the CAB gene family in all angiosperm genomes so far examined does not contain many defective genes (a defective gene is defined here as a gene containing a structural defect that either prevents its transcription or results in no functional protein issuing from the mRNA). Indeed, to date only a single CAB gene known to be defective has been found (9) out of the many scores of CAB genes analyzed. This is also the case in another gene family extensively analyzed in higher plants, the RBCS gene family encoding the small subunits of ribulose-bisphosphate carboxylase; only a few defective RBCS genes have so far been found (10). Although other gene families in higher plants have not been as extensively analyzed as the CAB and RBCS genes families, multiple defective genes have not yet been encountered.

In this report we describe the characterization of several genes encoding CAB proteins from the genome of the homosporous fern *Polystichum munitum* (n = 41, ref. 11). Only one of the four fully characterized genes encodes a potentially functional protein. Additional CAB genes isolated also appear to be defective. The results are discussed in view of CAB gene family evolution and the implications for the evolution of the fern genome.§

MATERIALS AND METHODS

Plant Material. The genomic library was constructed from a single plant of *P. munitum* (Kaulf.) Presl collected from a natural population near Saint Maries, Idaho (collection designation, *Soltis and Soltis 1845*). This plant was subsequently maintained in greenhouse culture at Washington State University.

Construction and Screening of *P. munitum* Genomic Library. Nuclei isolated by a sucrose gradient procedure (12) were then treated with hot cetyltrimethylammonium bromide and the DNA was further purified (13). Construction of the library followed Maniatis *et al.* (14). The *P. munitum* DNA (8.5 μg) was partially digested with the restriction endonuclease *Sau*3A1. After size fractionation by sucrose gradient centrifugation, fragments in the 15- to 20-kilobase range were ligated to the arms of the phage vector Charon 35. The library was screened with probes derived from tomato CAB genes. Specific probes are described in *Results*. Screening at low-stringency conditions was carried out as described (9).

Isolation and Characterization of Clones. DNA fragments from recombinant phages containing *P. munitum* DNA that

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: CAB protein, chlorophyll a/b-binding protein; PSI and PSII, photosystems I and II.

†To whom reprint requests should be addressed.

§The sequences reported in this paper have been deposited in the GenBank data base (accession no. M29584).

hybridized to tomato CAB probes were further cloned into the plasmid pBluescript (Stratagene) for nucleotide sequence determination. Sequencing was done by both the chemical (15) and the enzymatic (16) method. Most of the sequences, including all regions where deletions and insertions were found, were determined on both strands.

RESULTS

We screened 3×10^5 phages from the primary *P. munitum* genomic library (unamplified) with a probe derived from the coding region of the tomato PSII type I CAB gene *Cab-3C* (from the *Hin*dIII site in the middle of the gene to the *Xba* I site 3′ to the termination codon; ref. 9). This probe identified more than 50 positive phages. Six of these, numbered 3 through 8, were plaque-purified. Restriction mapping of the six clones revealed all six to contain different segments of the *P. munitum* genome (data not shown). We used DNA isolated from these phages in Southern blot (17) experiments with probes derived from different parts of the tomato *Cab-3C* gene. The original probe used in the isolation of these clones hybridized to all six clones, as expected. However, a probe derived from the 5′ half of the gene, (from a *Hin*dIII site 5′ to the initiating ATG codon to the *Hin*dIII site in the middle of the gene; ref. 9) hybridized only to clones 3, 4, and 7.

We further characterized, by nucleotide sequence determination, two of the clones (nos. 3 and 7) that hybridized to both the 5′ and the 3′ CAB probes and two of the clones (nos. 6 and 8) that hybridized to only the 3′ probe. The nucleotide sequences of these four clones are presented in Fig. 1. Only clone 3 contains a potentially functional CAB gene. Sequence comparisons indicate that this gene contains an intron with consensus splice sites (5′-GT . . . AG-3′) in the same position as do PSII type II CAB genes of angiosperms (18, 19) and an unspecified CAB gene from a moss (20), and it encodes a protein of 265 amino acids. The assignment of the intron was facilitated by the relatively high conservation of the amino acid

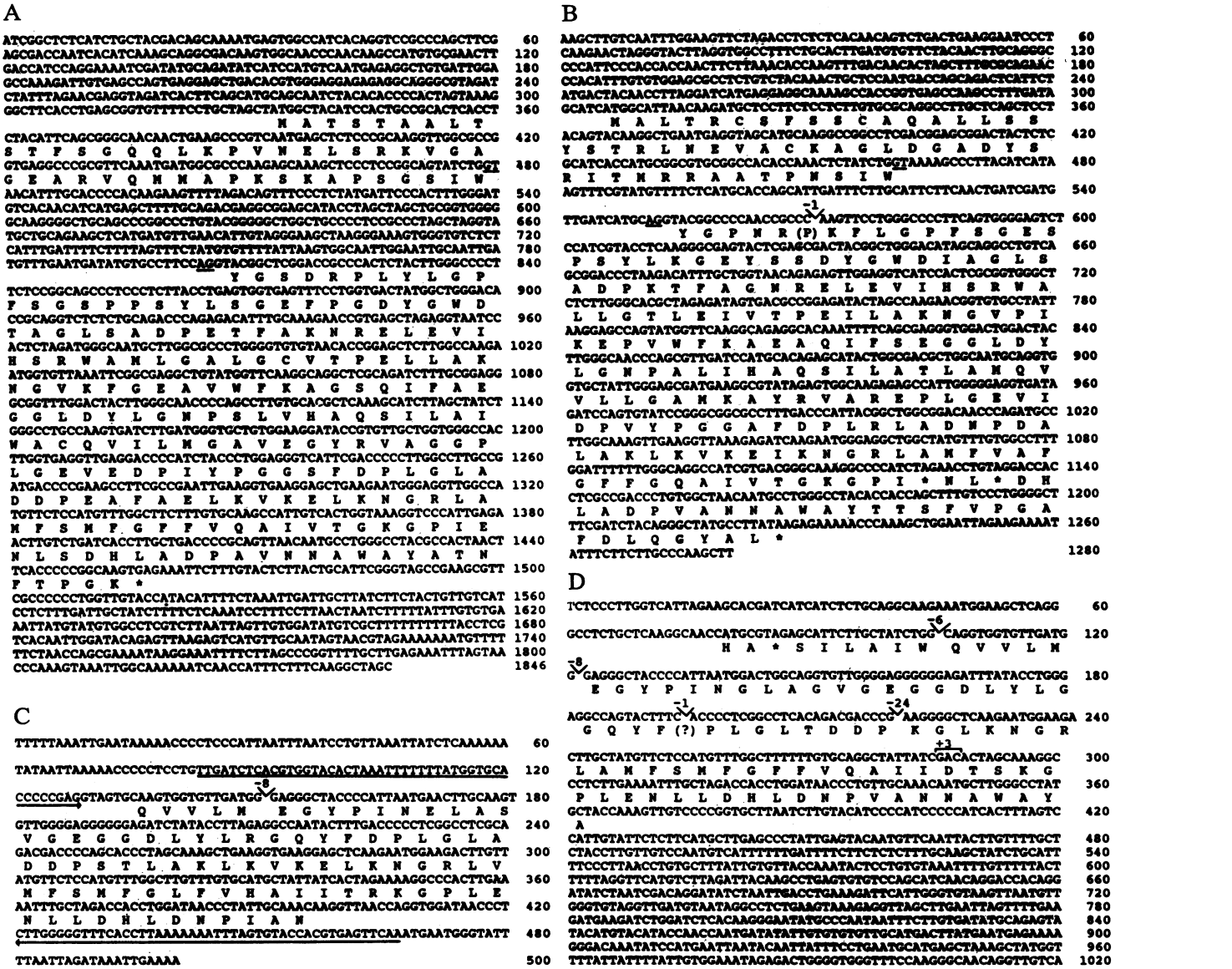

FIG. 1. Nucleotide sequences of fern CAB genes and flanking regions. (*A*) Sequence of the gene on clone 3. The amino acid sequence of the encoded polypeptide is shown (in one-letter code) below the DNA sequence. Underlined are the consensus 5′ and 3′ splice sites of the intron (GT . . . AG). (*B–D*) Sequence of the CAB gene on clones 7, 8, and 6, respectively. Positive numbers indicate sizes of insertions (in nucleotides) and negative numbers indicate sizes of deletions (all insertions and deletions are relative to both the fern CAB gene on clone 3 and the tomato CAB gene *Cab-3C;* in regions where *Cab-3C* and the fern clone 3 are not colinear, no insertions/deletions were inferred). Asterisks indicate stop codons in the middle of the coding region or directly downstream from it. Only the amino acid residues specified by the CAB-related sequences are shown. In *C*, arrows underscore positions of the inverted repeats flanking the sequence with homology to CAB sequences.

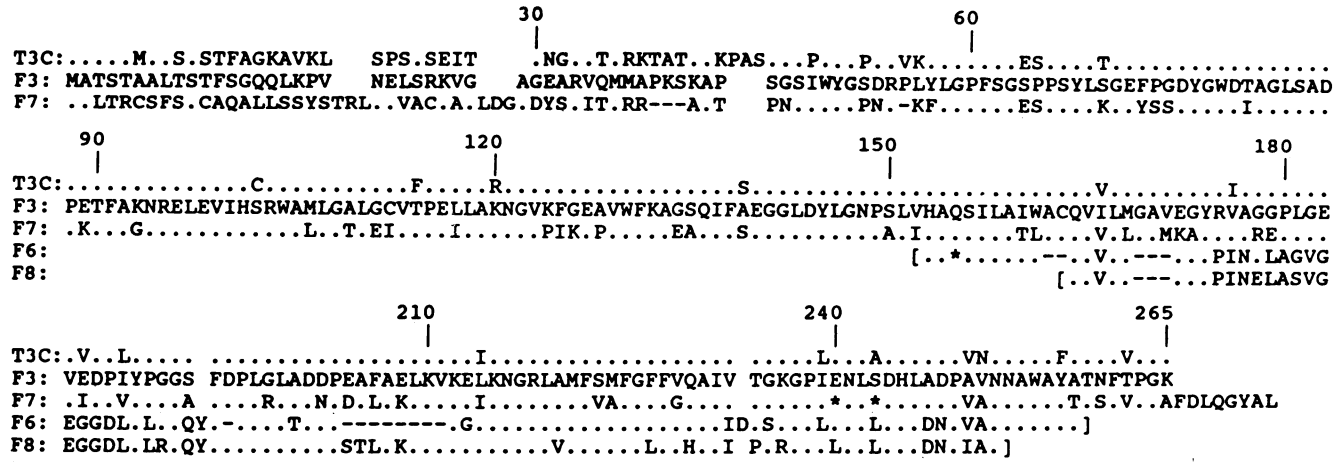

FIG. 2. Sequence comparisons of the protein-coding information of the fern CAB genes on clones 3 (F3), 7 (F7), 6 (F6), and 8 (F8) and the tomato gene *Cab-3C* (T3C). Only the F3 sequence is shown in full. Identical residues in the other proteins are indicated with dots. Brackets in the F6 and F8 sequences indicate the extent of the coding information found in these sequences. A dash represents a deletion in the defective gene(s) relative to F3 and T3C. Some gaps in the sequences have been introduced to maximize similarity without inferring direction of change (insertion or deletion).

sequence encoded by the parts of exon 1 and exon 2 flanking the intron. The part of the protein encoded by the first exon of the gene from clone 3 consists of the 36-residue transit peptide and the first 13 residues of the mature CAB protein. The amino acid sequence encoded by the first exon (with the exception of the first few codons and the last five codons) is highly variable even within angiosperms, and the fern sequence of this exon is also highly diverged from other known CAB sequences. The second exon of the fern CAB gene on clone 3 encodes a protein with >85% sequence identity to other known PSII type I and type II CAB proteins (Fig. 2, Table 1). This CAB gene does not contain any obvious structural defect (but see *Discussion*); however, we have not determined whether it is transcribed and, if it is transcribed, whether a functional protein is produced and is properly transported to and assembled in the chloroplast. This gene can thus be considered only as a "potentially functional" gene.

The CAB genes from the other three clones were revealed by nucleotide sequence determination to be clearly defective. The CAB gene on clone 7 contains both exons and the single intron in the same position as that in the CAB gene on clone 3. However, in the second exon it contains a deletion of a single nucleotide, as well as two nucleotide substitutions that result in nonsense (termination) codons (Figs. 1 and 2). The overall sequence identity of the coding region of this gene to the CAB gene on clone 3 and to angiosperm PSII CAB genes is much lower than the sequence identity of clone 3 to the angiosperm CAB sequences. The CAB genes on clones 6 and 8 each contain the central portion of the second exon only, with the first exon completely missing. Within the segment of homology, the sequences from clones 6 and 8 show additional deletions and insertions relative to known functional CAB genes. In addition, they also contain nonsense mutations (Figs. 1 and 2). Probing these clones with short probes from the extreme 5′ and 3′ coding regions of the CAB genes did not produce any additional signals, indicating that the missing parts of these genes do not reside in the immediate vicinity (≈5 kb upstream and downstream; data not shown). However, sequence determination of clones 6 and 8 did reveal a shared sequence, unrelated to CAB sequences, in the region upstream to each CAB-related sequence (data not shown).

Table 1. Sequence identity (percent) of the protein segment encoded by the second exon of CAB genes of fern, moss, and tomato

	F3	M1	I-T3C
M1	87.0		
I-T3C	90.3	87.5	
II-T5	87.0	84.3	88.0

F3, fern CAB gene on clone 3; M1, CAB gene of a moss (20); I-T3C, tomato PSII type I CAB gene *Cab-3C* (9); II-T5, tomato PSII type II CAB gene *Cab-5* (21).

DISCUSSION

The CAB gene from clone 3 has an intact structure typical of other known functional CAB genes from moss (20) and angiosperms (8). The protein it encodes shows high sequence similarity to the previously reported moss CAB sequence and to PSII type I and II CAB sequences of angiosperms (Table 1). However, it does not contain the highly conserved sequence Met-Arg-Arg or Met-Arg-Lys at the processing site of the transit peptide but instead contains the sequence Met-Met-Ala (Fig. 2). Thus, although a protein with high sequence identity to other known CAB proteins may result from the expression of this gene, further work is necessary to show whether indeed this gene is transcribed and, if so, whether a functional protein is produced and is properly transported to and assembled in the chloroplast.

In angiosperms, the PSII type I genes contain no introns (9) but the type II genes do contain a single intron (18, 19). This intron is present in the moss CAB gene and in the two fern sequences described here that contain sequences homologous to both the 5′ and the 3′ coding region of CAB genes (i.e., the CAB genes from clones 3 and 7). Thus, a CAB gene containing no introns has not yet been identified in any species other than angiosperms. It is likely that the PSII type I CAB genes of angiosperms represent a lineage that has lost this intron. The angiosperm type I and type II CAB genes (represented by tomato) encode proteins that are only 12% divergent from each other in the region compared in Table 1. In the same region, the protein encoded by the fern CAB gene from clone 3 is slightly more similar to the tomato PSII type I protein (9.7% divergence) than to the tomato PSII type II protein (13% divergence), suggesting that the angiosperm PSII type I and type II CAB gene lineages diverged slightly before, or perhaps immediately after, the separation of the fern and angiosperm lineages. Other types of angiosperm CAB genes encode proteins that show much greater divergence (60–70%) (1), and the splits among these lineages are therefore likely to have occurred prior to the divergence of lineages leading to angiosperms and ferns. Genes for these additional CAB lineages have not been isolated from ferns.

The sequences found on clones 6, 7, and 8 represent CAB-related sequences. However, structural defects (insertions, deletions, nonsense mutations) clearly render these CAB genes nonfunctional. In each case, the portion of the gene that has survived also shows a high rate of coding sequence divergence compared with the known functional CAB genes (Fig. 2). The large proportion of nonfunctional PSII CAB genes in *P. munitum* is in marked contrast to the situation in angiosperms, where a detailed investigation of several species has revealed few or no clearly defective CAB genes. For example, the tomato genome contains eight functional type I and two functional type II CAB genes and only a single truncated type I gene (9).

A possible trivial explanation for these observations is that artifacts in the cloning procedures have produced DNA sequences that do not reflect the situation *in vivo* in the *P. munitum* genome. Two observations argue against this possibility. First, artifacts in cloning usually involve large deletions and insertions (22), but not short deletions and insertions and nucleotide substitutions, and all these changes have been found in the cloned genes. Another observation that suggests that the cloned sequences are accurate representations of sequences in the genome is the similarity between the defective CAB genes from clones 6 and 8. For example, both contain an eight-nucleotide deletion in the exact same position (Fig. 1), and the sequences of the encoded proteins are very similar to each other but very different from other known CAB sequences (Fig. 2). Clones 6 and 8 are independent isolates from the primary library construction (unamplified) and it is, therefore, very unlikely that the many small changes that are shared by both sequences have occurred independently during the construction and handling of the library.

Several scenarios could be invoked to explain the existence of multiple defective CAB genes in the genome of *P. munitum*. First, some defective CAB genes may have been specifically amplified, by unspecified mechanisms, in the genome of *P. munitum* during evolution. Unusually large numbers of copies of a specific gene in some species have been observed previously. For example, the genome of petunia contains >200 copies of actin genes, whereas other plants examined have at most 10–15 copies (23). It is not known, however, whether any of the petunia actin genes are defective. In mammalian genomes, multiple nonfunctional "pseudogenes" are often created through reverse transcription of mRNA and integration of the copy DNA into the chromosome (24). The defective CAB genes whose sequences we present here are unlikely to have been amplified by an RNA intermediate, since one of them (on clone 7) contains an intron, and the other two (on clones 6 and 8) are missing both 5′ and 3′ ends of the coding region and show no evidence of a poly(A) tail (Fig. 1). The similarity between clones 6 and 8 is intriguing. They may indeed represent an amplified sequence in the genome, or they may simply be allelic (with extensive deletions on both homologous chromosomes at this locus).

A second scenario assumes that the *P. munitum* lineage has never undergone a polyploidization event, and that starting with a normal complement of PSII CAB genes, several, and perhaps the majority, of the members of the PSII CAB gene family in the genome of this plant have mutated to a nonfunctional state. As discussed in the Introduction, this phenomenon has not been observed in the CAB gene family in other plant genomes or in other angiosperm gene families.

A third scenario is based on the assumption that *P. munitum* was, sometime during its evolution, a polyploid. Then its present genome may have undergone the process of diploidization, which, among other things, also involved mutations to nonfunctionality in extra copies of genes (4, 5). Gene silencing in plant polyploids has not been systematically studied on the molecular level, but some angiosperm plants of known recent polyploid origin, such as hexaploid wheat (*Triticum aestivum*, $n = 21$), have been shown to maintain multiple functional copies of genes encoding many, but not all, isozymes (25). *P. munitum* and other ferns with similar chromosome numbers ($n = 27$–52) do not show multiple copies of functional genes encoding most isozymes (4, 5). If these ferns are in fact descended from polyploid lineages, as suggested by their chromosome number, it must be assumed that the polyploidization events in the history of these lineages are much more ancient than the polyploidization event that gave rise to wheat. Alternatively, gene silencing in fern polyploids may occur much faster than in angiosperm polyploids.

For several reasons, some of which depend on the specific gene system examined (CAB) and some of which do not, it is not possible with the present data to distinguish among the three competing hypotheses listed above. In screening the *P. munitum* DNA library, we identified more than 50 positive phages, and the first 6 that we further characterized were all different, suggesting that the number of CAB genes in *P. munitum* is considerably greater than 6. However, even if the number of CAB genes in *P. munitum* were much higher than that encountered in diploid angiosperms (at present, the known range is 3–17; refs. 19, 26, and 27), it would still not be possible to distinguish between specific gene amplification and ancient polyploidy. On the other hand, because PSII CAB gene number varies widely even within diploid angiosperms, a relatively low copy number (6 or a few more) cannot rule out the ancient polyploidy model either. Although discriminating among the hypotheses we present here is further complicated by the fact that we have examined a gene family with a variable number of genes in different organisms, in principle the same difficulties will also arise when single-copy gene systems are examined. This is so because if multiple copies (defective or functional) are found, *ad hoc* explanations can again be invoked—for example, specific gene amplification.

Thus, in practice, gene silencing and, by implication, the ancient polyploid origin of a fern species with $n = 27$–52 (or of any other polyploid) can be demonstrated only when the presence of multiple defective genes is shown to be the case not just for one or a few gene systems (each system may have its own peculiarities) but for *most* nuclear gene systems, including gene families and the genes usually found in single copy in diploid seed plants. The accumulation of these kinds of data is therefore of the utmost importance in the study of genome evolution in general and in particular in polyploid plants, both in ferns and in other types of plant groups, and the molecular techniques (e.g., gene cloning and sequencing) to carry out such investigations are now available.

Although it is thus premature to conclude that our results indicate an ancient polyploid origin for *P. munitum*, it is intriguing that this study, which to our knowledge is the first molecular examination of a fern nuclear gene system, detected multiple defective genes. If defective copies are found in most other gene systems in ferns, it would be of great interest to examine the effect such nonfunctional genes have on the expression of the fern genome, and, in consequence, on the cellular and developmental processes of the plant.

We thank B. E. Froman for assistance and advice in the construction of the genomic library, Dr. L. D. Gottlieb for providing his lab facilities and advice in the library construction, D. Stein for providing us with some DNA samples, Dr. A. Zilberstein for disclosing her unpublished results (which may be similar to ours) and Dr. L. D. Gottlieb, C. H. Haufler, and W. H. Wagner for critical review of the manuscript. This work was supported by U.S. Department of Agriculture Competitive Research Grant Office Grant 88-37262-3878

(to E.P.) and National Science Foundation Grants BSR 8620444 (to P.S.) and BSR 8607054 (to L. D. Gottlieb).

1. Klekowski, E. J. & Baker, H. G. (1966) *Science* **153,** 305–307.
2. Grant, V. (1981) *Plant Speciation* (Columbia Univ. Press, New York).
3. Wagner, W. H. & Wagner, F. S. (1980) in *Polyploidy: Biological Relevance*, ed. Lewis, W. H. (Plenum, New York), pp. 199–214.
4. Haufler, C. H. & Soltis, D. E. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 4389–4393.
5. Haufler, C. H. (1987) *Am. J. Bot.* **74,** 953–966.
6. Gottlieb, L. D. (1982) *Science* **216,** 373–380.
7. Soltis, D. E. & Soltis, P. S. (1989) in *Isozymes and Plant Biology*, eds. Soltis, D. E. & Soltis, P. S. (Dioscorides, Portland, OR), in press.
8. Pichersky, E., Brock, T. G., Nguyen, D., Hoffman, N. E., Piechulla, B., Tanksley, S. D. & Green, B. R. (1989) *Plant Mol. Biol.* **12,** 257–270.
9. Pichersky, E., Bernatzky, R., Tanksley, S. D., Breidenbach, R. B., Kausch, A. P. & Cashmore, A. R. (1985) *Gene* **40,** 247–258.
10. Dean, C., Pichersky, E. & Dunsmuir, P. (1989) *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **40,** 415–439.
11. Love, A., Love, D. & Pichi Sermolli, R. E. G. (1977) *Cytotaxonomical Atlas of the Pteridophyta* (Cramer, Vaduz, Liechtenstein).
12. Palmer, J. D. (1986) *Methods Enzymol.* **118,** 167–186.
13. Saghai-Maroof, M. A., Soliman, K. M., Jorgensen, R. A. & Allard, R. W. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 8014–8018.
14. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning* (Cold Spring Harbor Lab., Cold Spring Harbor, NY).
15. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560.
16. Sanger, F., Nicklen, S. & Coulsen, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
17. Southern, E. M. (1975) *J. Mol. Biol.* **98,** 503–517.
18. Karlin-Neumann, G. A., Kohorn, B. D., Thornber, J. P. & Tobin, E. M. (1985) *J. Mol. Appl. Genet.* **3,** 45–61.
19. Stayton, M. M., Black, M., Bedbrook, J. & Dunsmuir, P. (1986) *Nucleic Acids Res.* **14,** 9781–9796.
20. Long, Z., Shue-Yuan, W. & Nelson, N. (1989) *Gene* **76,** 299–312.
21. Pichersky, E., Hoffman, N. E., Malik, V. S., Bernatzky, R., Tanksley, S. D., Szabo, L. & Cashmore, A. R. (1987) *Plant Mol. Biol.* **9,** 109–120.
22. Federoff, N. (1983) *Plant Mol. Biol. Rep.* **1,** 27–29.
23. Baird, W. V. & Meagher, R. B. (1987) *EMBO J.* **6,** 3223–3231.
24. Wagner, M. (1986) *Trends Genet.* **2,** 134–137.
25. Hart, G. E. (1983) *Isozymes: Curr. Top. Biol. Med. Res.* **10,** 365–380.
26. Dunsmuir, P. (1985) *Nucleic Acids Res.* **13,** 2503–2518.
27. Leutwiler, L. S., Meyerowitz, E. M. & Tobin, E. M. (1986) *Nucleic Acids Res.* **14,** 4051–4076.