Mol Gen Genet (1988) 215:65–68

MGG

Chloroplast DNA sequences integrated into an intron of a tomato nuclear gene

Eran Pichersky[1] **and Steven D. Tanksley**[2]

[1] Department of Biology, University of Michigan, Ann Arbor, MI 48109, USA
[2] Department of Plant Breeding and Biometry, Cornell University, Ithaca, NY 14853, USA

Summary. DNA sequences capable of hybridizing with chloroplast DNA have previously been reported to exist in the nuclear genome of higher plants. Here we show that the third intron of the cultivated tomato (*Lycopersicon esculentum*) nuclear gene *Cab*-7, which resides on chromosome 10 and which we recently cloned and sequenced, contains two DNA fragments derived from the coding region of the chloroplast gene *psb*G. The first fragment, 133 bp long, is located at a site 63 bp from the 3′ end of the 833 bp intron. The exact sequence of the 11 nucleotides at the 3′ end of the inserting chloroplast sequence is also found at the 5′ border of the insertion. A small (107 bp) chloroplast DNA fragment is inserted near the middle of the intron, again with the 3′ end of the inserting element (6 bp) duplicated at the 5′ border of the insertion. The second insert is a subfragment of the first insert, and is most likely directly derived from it. The *psb*G insertion sequence was found to be present in the *Cab*-7 gene of all tomato species examined but not in species from related genera (e.g. *Solanum, Petunia, Nicotiana*), suggesting that the original transposition event (chloroplast to nucleus) occurred relatively recently – since the divergence of the genus *Lycopersicon* from other genera in the family Solanaceae, but before radiation of species in that genus.

Key words: *Lycopersicon esculentum* – DNA transfer – Homologous recombination – Restriction fragment length polymorphism – Photosynthetic genes

Introduction

We have recently isolated a tomato nuclear gene encoding a chlorophyll a/b-binding (CAB) polypeptide (Pichersky et al. 1988). The gene, *Cab*-7, is a member of a nuclear gene family encoding several different types of such polypeptides. *Cab*-7 was mapped to one end of chromosome 10 (Pichersky et al. 1988). Unlike other tomato CAB genes, which either contain no introns or short ones (87–107 bp; Pichersky et al. 1985; 1987a, b), *Cab*-7 contains 4 introns ranging in size from 363–916 bp. Here we report the discovery that part of the sequence of intron III of *Cab*-7 is of chloroplast DNA origin. Although it has been demonstrated that plant nuclear genomes contain many sequences of chloroplast origin (Timmis and Scott 1983; Scott and Timmis 1984), such sequences have not previously been physically isolated and characterized.

Offprint requests to: E. Pichersky

Materials and methods

DNA sequences. The complete sequence and the chromosomal location of the tomato nuclear gene *Cab*-7 has been reported elsewhere (Pichersky et al. 1988). The complete sequence of the *Nicotiana tabacum* chloroplast DNA has been reported by Shinozaki et al. (1986).

Southern blots and probe preparation. Southern blots were performed as previously described (Bernatzky and Tanksley 1986). To prepare an intron III-specific probe, a fragment extending from the most 5′ *Dra*I site in the third intron to the most 5′ *Pst*I site in the fourth exon (see Fig. 1), was cloned into the polylinker region of pUC18 which was cut with *Pst*I and *Hin*cII. This clone was designated pD2A. For probe labelling, the insert was excised from pD2A, run in a 1% agarose gel, electroeluted and labelled with ^{32}P with Klenow enzyme and a mixture of random primers. The synthetic 52-mer probe described in the text was labelled with ^{32}P using Klenow enzyme and a 6-mer primer complementary to its end (positions 105–100 in Fig. 1). Tomato leaf DNA was isolated as described in Bernatzky and Tanksley (1986). Tomato chloroplast DNA was isolated as described by Palmer and Zamir (1982).

Results

Localization of chloroplast-derived fragments in the third intron

In Southern blots of *Eco*RI-digested DNA isolated from tomato leaf tissue, the intron III-specific probe, pD2A, hybridized to numerous fragments in addition to the fragment carrying the *Cab*-7 gene (Fig. 2, lane c). At high stringency, most of the hybridization signal could be attributed to a 3.3 kbp fragment carrying the *Cab*-7 gene as well as an additional, even more strongly hybridizing fragment of 1.0 kbp in size (Fig. 2, lane d). Since the probe has perfect homology to the sequence on the 3.3 kbp *Eco*RI fragment, the observation that this additional fragment hybridized significantly more strongly to the probe (especially at low stringency) suggested that the molar ratio of the 1.0 kbp *Eco*RI fragment to the 3.3 kbp *Eco*RI fragment present in the diploid tomato cell might be greater than one.

Since we did not observe any restriction fragment length polymorphism (RFLP) of this additional fragment in inter-

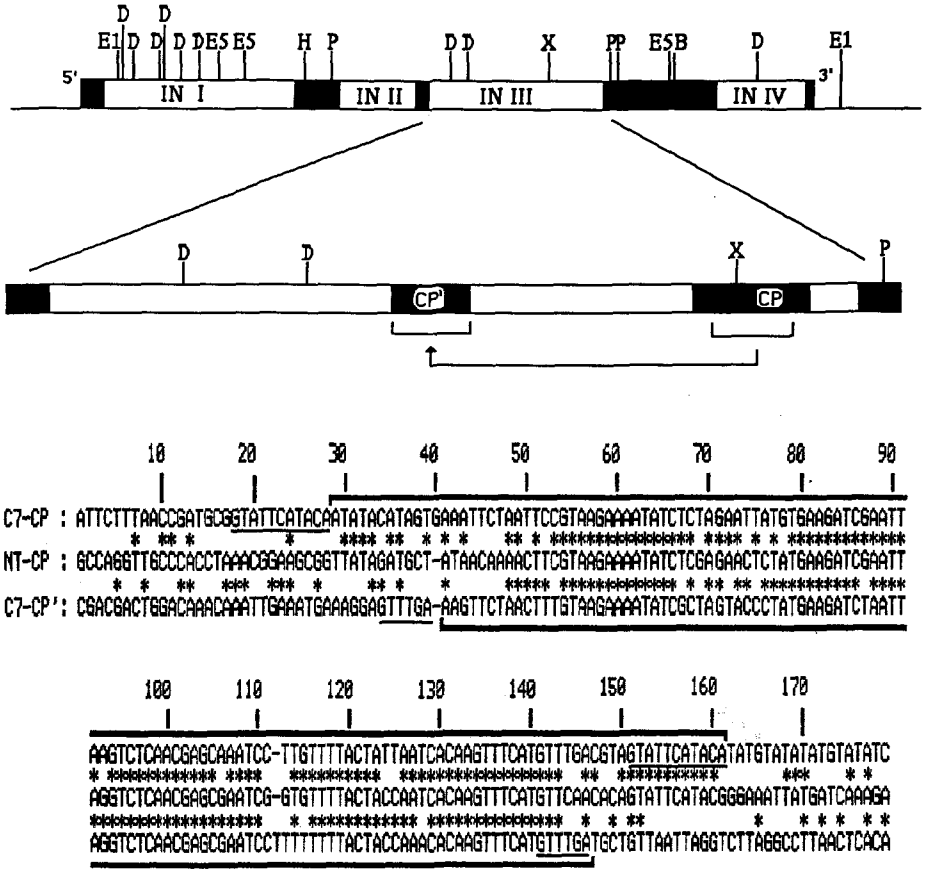

Fig. 1. Top. Schematic diagram of tomato *Cab*-7 gene. *Solid regions,* exons; *open regions,* introns. For scale, the first intron has 916 bp and the third intron has 833 bp. Sites are defined as: D, *Dra*I; E1, *Eco*RI; E5, *Eco*RV; H, *Hin*dIII; P, *Pst*I; X, *Xba*I. Middle. Enlarged view of intron III showing insertions of *psb*G gene fragments (CP, CP′). Bottom. Nucleotide sequence of part of the *Nicotiana tabacum psb*G gene (NT-CP) compared with the CP and CP′ elements in the third intron of the tomato *Cab*-7 gene (C7-CP and C7-CP′, respectively). *Brackets* indicate regions of homology between C7-CP, C7-CP′, and NT-CP. *Underlined* sequences represent direct repeats. *Asterisks* denote base identity between either C7-CP or C7-CP′ and NT-CP. The *N. tabacum* chloroplast DNA sequence shown here (NT-CP) extends from position 51833 to position 51657 in the sequence published by Shinozaki et al. (1986). The complementary strand is shown (i.e. position 51833 = position 1 in this figure, position 51657 = position 179 in this figure) because this is the sense strand for *psb*G. The upper line (C7-CP) is sequence from positions 2770–2948 in Fig. 1 of Pichersky et al. (1988). The lower line (C7-CP′) is sequence from positions 2379–2557 in Fig. 1 of Pichersky et al. (1988)

specific comparisons between *Lycopersicon esculentum* and its congeneric relative *L. pennellii* (see Fig. 2, lanes f, g for the *Eco*RI digest; additional data not shown), whereas RFLPs for the fragment carrying the *Cab*-7 gene were frequently observed (e.g. Fig. 2, lanes f, g), we hypothesized that this additional fragment was derived from the chloroplast genome. It is well established that the chloroplast genome of higher plants evolves much more slowly than plant nuclear genomes (Palmer 1985). We tested this hypothesis by using the same probe against purified chloroplast DNA digested with *Eco*RI. In this experiment, a similar sized (1.0 kbp) fragment showed strong hybridization (Fig. 2, lane e), suggesting that the 1.0 kbp *Eco*RI fragment is indeed derived from chloroplast DNA.

We next compared the nucleotide sequence of the third intron with the complete sequence of the *N. tabacum* chloroplast DNA (Shinozaki et al. 1986). This comparison (Fig. 1) revealed two segments in the third intron with strong sequence similarity to the *N. tabacum* chloroplast gene *psb*G (Steinmetz et al. 1986). *N. tabacum* and *L. esculentum* are in the same family, Solanaceae; their chloroplast DNA molecules are approximately 2%–5% divergent in sequence (J.D. Palmer, personal communication). The first segment, which we designate C7-CP, is situated near the 3′ end of the intron and consists of 133 bp with greater than 80% sequence identity with a segment situated about two-thirds into the coding region of *psb*G. The C7-CP insert is flanked by an 11 nucleotide direct repeat (Fig. 1). This sequence (with the exception of the last nucleotide) is also found in the right-hand (3′) border of the *N. tabacum psb*G sequence homologous to the C7-CP sequence. The second segment with sequence similarity to *psb*G, designated C7-CP′, contains 107 nucleotides and is situated near the middle of the intron (Fig. 1). C7-CP′ is a subfragment of C7-CP. It is also flanked by a direct repeat, of 6 bp, and again the sequence of this direct repeat is also found in the right-hand border of the C7-CP sequence showing sequence similarity to C7-CP′ (Fig. 1).

Additional repeats in the third intron

When the third intron of *Cab*-7 was probed at low stringency onto filters containing total tomato DNA, multiple hybridizing fragments were observed (Fig. 2, lane c). Possible explanations for this observation are: (1) the *psb*G insert is found in additional positions in the nuclear genome other than the *Cab*-7 gene; or (2) the third intron contains other repeated sequences, independent of the *psb*G insertions, which are found elsewhere in the genome. To distinguish between these hypotheses, a 52 bp nucleotide probe was synthesized with perfect homology to the portion of the intron containing the C7-CP element (positions 54–105, Fig. 1). When this oligomer was probed at low stringency against *Eco*RI-digested tomato leaf DNA, it hybridized only to the fragment containing the *Cab*-7 gene and to the chloroplast fragment containing the *psb*G gene (Fig. 2, lane f). These results indicate that the *psb*G insert is found in tomato only in the *Cab*-7 intron, and that this intron also contains other, independent repetitive elements.

Phylogenetic distribution of the psb*G insertion in Solanaceae*

When the synthetic oligonucleotide was probed onto a filter containing *Eco*RI-digested DNA from *L. pennellii*, it hybridized to the 1.0 kbp chloroplast fragment and another

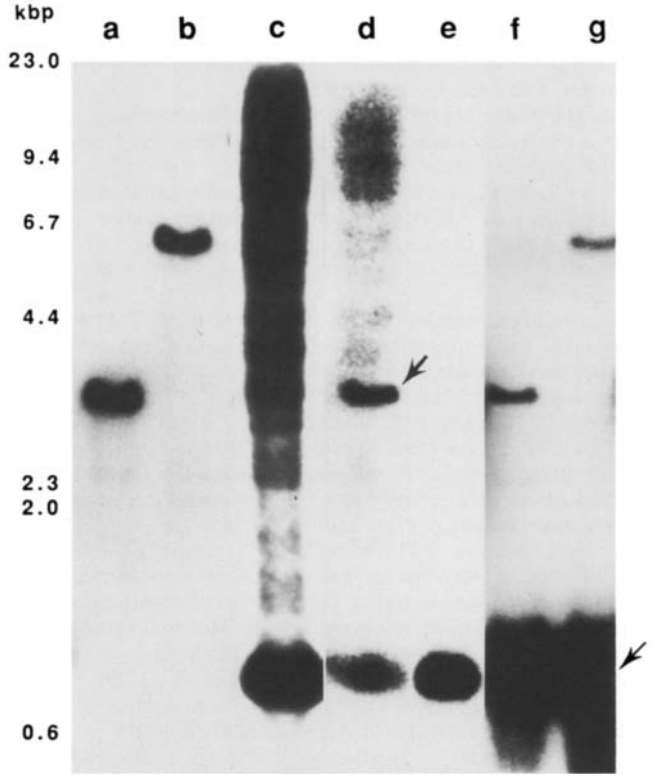

Fig. 2. Autoradiograms derived from probing genomic blots of *Eco*RI-digested tomato DNA with different probes containing sequences homologous to the *Cab*-7 gene. All lanes contained 5 µg of DNA extracted from leaf tissue except lane e which contained 0.5 µg of *Lycopersicon esculentum* chloroplast DNA (a gift from Dr. J.D. Palmer). Lanes a, b; *L. esculentum* (a) and *L. pennellii* (b) DNA probed with 0.6 kbp *Pst*I-*Eco*RI fragment derived from the 3′ end of a *Cab*-7 cDNA clone. Lanes c, d; *L. esculentum* DNA probed with pD2A, a clone containing most of the third intron of *Cab*-7. The blot in lane c was washed to low stringency (2 × SSC, 65° C) [1 × SSC = 0.15 M NaCl/0.015 M sodium citrate, pH 6.8]. The blot in lane d was washed to high stringency (0.1 × SSC, 65° C). *Arrow* in lane D points to the genomic fragment containing *Cab*-7 intron III from which the probe was derived. Lane e; Purified tomato chloroplast DNA probed with pD2A (washed at 2 × SSC, 65° C). Lanes f, g; *L. esculentum* (f) and *L. pennellii* DNA (g) probed with the 52-mer synthetic oligonucleotide homologous to the a portion of the *psb*G gene fragment integrated in *Cab*-7. This oligonucleotide has perfect homology to positions 54–105 of C7-CP. The blots were washed at 2 × SSC, 65° C. *Arrow* indicates chloroplast DNA fragment containing the *psb*G gene

6.5 kbp fragment. When the same filter was stripped of the bound probe and hybridized with a *Cab*-7 cDNA clone, only the 6.5 kbp fragment hybridized, indicating that *L. pennellii* also possesses the *psb*G insert in the *Cab*-7 gene (Fig 2, lanes b, g). Similar experiments were conducted with DNA from other *Lycopersicon* species representing all sections of the genus, as well as from *Solanum tuberosum* (potato), *Datura meteloides, Petunia hybrida* and *N. tabacum* (data not shown). In all cases the *Lycopersicon* species were found to have the *psb*G insert in the *Cab*-7 gene (however, from this kind of analysis it could not be determined if they all contained both C7-CP and C7-CP′ elements, or only one of them). Occasionally, DNA fragments of the other solanaceous species also showed hybridization with the synthetic probe; however, none of the fragments correspond to the one(s) containing the *Cab*-7 gene. It thus seems likely that the hybridizing fragments in these species represent independent insertions of some or all of the *psb*G gene into different portions of the nuclear genome. Taken together, the data indicate that the insertions of the two fragments of the *psb*G gene into the *Cab*-7 gene occurred after the divergence of the genus *Lycopersicon* from other genera, but before radiation of species in that genus.

Discussion

Relationship between C7-CP and C7-CP′

Since the sequence of C7-CP′ is wholly contained within C7-CP, it seems likely that the first is derived from the latter. To strengthen this hypothesis, it is necessary to show that both elements have had a common history after leaving the chloroplast genome. There are ten positions within the region shared by the two in which they contain the same nucleotide not found in the NT-CP sequence. Undoubtedly the nucleotides in some of these positions in the chloroplast *psb*G genes of *Lycopersicon* species are identical to those of C7-CP and C7-CP′, but in some of the ten positions the nucleotide found in C7-CP and C7-CP′ most likely were not present in the ancestral *Lycopersicon psb*G gene. This is so because the T at position 47, and the T at position 144 or the G at position 145 (or both) create stop codons within the open reading frame of the *psb*G sequence (position numbers refer to Fig. 1). The TGA trinucleotide at position 144–146 is especially indicative because it is part of the 6 nucleotides direct repeat flanking C7-CP′. If the direct repeat was present in C7-CP′ from its inception, as is likely, then the stop codon which is part of this 6 bp sequence was not created by substitution events which occurred independently in the total of 3 copies of this sequence found in C7-CP and C7-CP′. Another intriguing observation is that C7-CP has an insertion of one nucleotide at the 5′ border of the homology with C7-CP′ (Fig. 1).

Novel footprints at insertion sites

Features of the *psb*G insertions do not match those of known sequence insertion mechanisms in plants. Transposons, which are well documented in plants, generate direct duplications of *target* DNA (Doring and Starlinger 1986). The C7-CP and C7-CP′ insertions are flanked by direct repeats, but in these two cases the copy of the sequence which is duplicated is part of the *donor* sequence (in both cases, on the 3′ end). C7-CP and C7-CP′ also differ from transposons in that the length of the direct repeat is variable. C7-CP is flanked by an 11 bp repeat and C7-CP′ by a 6 bp repeat. While different transposon families may generate different lengths of direct repeats, the length of the repeat is normally constant for a given family (Doring and Starlinger 1986).

Analysis of the site of integration

Since all tomato species appear to contain the *psb*G element in the *Cab*-7 gene, it is not possible to determine the original sequence of the intron before the insertion events, since the sequence of the corresponding intron in other solanaceous species is likely to have diverged considerably. Thus we cannot rule out the possibility that a sequence identical

to the duplicated 3′ end of each insertion was originally present in the target DNA. If such a sequence did exist in the target site, a mechanism involving homologous recombination might be invoked to explain the insertions. The insertion of C7-CP occurred at the 5′ end of a stretch of alternating purine and pyrimidine residues, and if C7-CP′ indeed originated from C7-CP, then the 3′ end of the inserting element was within a few nucleotides from the same stretch of alternating purine and pyrimidine bases. Alternating purine and pyrimidine residues have been shown to constitute "hot spots" where gene conversion among fetal globin genes initiates (Slighton et al. 1980). Presumably, such stretches form Z-DNA helix which easily unwinds, thus making single-stranded DNA available for heteroduplex formation (Rich et al. 1984). A model of homologous recombination to explain the presence of *psb*G sequences in *Cab*-7 implies that the initial chloroplast DNA insertion occurred because a fragment from the *psb*G was present in the nucleus and happened to fit the sequence in the third intron of *Cab*-7 which became available for heteroduplex formation as the intron DNA was unwinding at the alternating purine-pyrimidine stretch. The susceptibility of the third intron for sequence invasion would also explain the presence of other repetitive elements in that same intron.

It has been previously demonstrated by the Southern blotting technique that many chloroplast DNA sequences are present in the nuclear genome (Timmis and Scott 1983; Scott and Timmis 1984). The analysis of the sequences around the integration sites we examined in this study suggests the possibility that the insertions occurred through some mechanism involving homologous recombination, although alternative mechanisms cannot be ruled out. Since, according to the endosymbiotic hypothesis, a massive movement of chloroplast DNA to the nucleus must have occurred (Weeden 1981) and such a process is still ongoing (Timmis and Scott 1983; Scott and Timmis 1984), as well as DNA movement from chloroplast to mitochondria (Stern and Lonsdale 1982), the mechanisms involved in such a movement is of great interest.

Acknowledgements. We thank Dr. J.D. Palmer for his gift of *L. esculentum* chloroplast DNA and for providing the computer tapes with the *N. tabacum* chloroplast DNA sequence. This work was supported by a USDA grant and a University of Michigan's Rackham Faculty Research Grant to EP, and an NSF grant to SDT.

References

Bernatzky R, Tanksley SD (1986) Majority of random cDNA clones correspond to single loci in the tomato genome. Mol Gen Genet 203:8–14

Doring HP, Starlinger P (1986) Molecular genetics of transposable elements in plants. Annu Rev Genet 20:175–200

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

Palmer JD, Zamir D (1982) Chloroplast DNA evolution and phylogenetic relationships in *Lycopersicon*. Proc Natl Acad Sci USA 79:5006–5010

Pichersky E, Bernatzky R, Tanksley SD, Breidenbach RB, Kausch AP, Cashmore AR (1985) Molecular characterization and genetic mapping of two clusters of genes encoding chlorophyll a/b-binding proteins in *Lycopersicon esculentum* (tomato). Gene 40:247–258

Pichersky E, Hoffman NE, Malik VS, Bernatzky R, Tanksley SD, Szabo L, Cashmore AR (1987a) The tomato *Cab*-4 and *Cab*-5 genes encode a second type of CAB polypeptides localized in photosystem II. Plant Mol Biol 9:109–120

Pichersky E, Hoffman NE, Bernatzky R, Piechulla B, Tanksley SD, Cashmore AR (1987b) Molecular characterization and genetic mapping of DNA sequences encoding the Type I chlorophyll a/b-binding polypeptide of photosystem I in *Lycopersicon esculentum* (tomato). Plant Mol Biol 9:205–216

Pichersky E, Tanksley SD, Piechulla B, Stayton M, Dunsmuir P (1988) Nucleotide sequence and chromosomal location of *Cab*-7, the tomato gene encoding the Type II chlorophyll a/b-binding polypeptide of Photosystem I. Plant Mol Biol 11:69–71

Rich A, Nordheim A, Wang AJ-H (1984) The chemistry and biology of left-handed Z-DNA. Annu Rev Biochem 53:791–846

Scott NS, Timmis JS (1984) Homologies between nuclear and plastid DNA in spinach. Theor Appl Genet 67:279–288

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Cunwongse J, Obakata J, Yamaguchi-Shinozaki K, Ohto C, Torazawa K, Meng BY, Sugita M, Deno H, Kamogashira T, Yamada K, Kusuda J, Takaiwa F, Kato A, Tohdon N, Shimoda H, Sugiura M (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J 5:2043–2049

Slighton JL, Blechl AE, Smithies O (1980) Human fetal ^{G}gamma and ^{A}gamma globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicate genes. Cell 21:627–638

Steinmetz AA, Castroviejo M, Sayre RT, Bogorad L (1986) Protein PSII-G. J Biol Chem 261:2485–2488

Stern D, Lonsdale DM (1982) Mitochondrial and chloroplast genomes of maize have a 12-kilobase DNA sequence in common. Nature 299:698–702

Timmis JN, Scott NS (1983) Spinach nuclear and chloroplast DNAs have homologous sequences. Nature 305:65–67

Weeden NF (1981) Genetic and biochemical implications of the endosymbiotic origin of the chloroplast. J Mol Evol 17:133–139

Communicated by J. Schell

Received June 13, 1988