

## Nucleotide sequence and chromosomal location of *Cab-7*, the tomato gene encoding the Type II chlorophyll a/b-binding polypeptide of Photosystem I

Eran Pichersky,<sup>1</sup> Steven D. Tanksley,<sup>2</sup> Birgit Piechulla,<sup>3</sup> Mark M. Stayton<sup>4</sup> and Pamela Dunsmuir<sup>4</sup>  
<sup>1</sup>Biology Department, University of Michigan, Ann Arbor, MI 48109, USA; <sup>2</sup>Department of Plant Breeding and Biometry, Cornell University, Ithaca, NY 14853, USA; <sup>3</sup>Institut für Biochemie der Pflanze der Universität Göttingen, Untere Karspule 2, 34 Göttingen, Federal Republic of Germany; <sup>4</sup>Advanced Genetic Sciences, Inc., 6701 San Pablo Avenue, Oakland, CA 94608, USA

Received 19 January 1988; accepted 7 March 1988

Stayton *et al.* [5] have recently reported the sequence of a petunia cDNA clone, LHCI-15, encoding a Photosystem I (PSI) chlorophyll a/b-binding (CAB) protein which is part of a complex known as LHCI-680 or CP24. We used the petunia cDNA clone LHCI-15 to isolate the corresponding *Lycopersicon esculentum* (tomato) genomic and cDNA clones from our genomic [6] and cDNA [2] libraries. The tomato genome contains a single gene encoding this type of CAB polypeptide. The gene was designated *Cab-7*, and it maps to the end of chromosome 10,  $1.3 \pm 1.3$  cM from the RFLP marker TG122. In Fig. 1 we present the nucleotide sequence of *Cab-7*, starting from the *Hind* III site 602 nucleotides 5' to the initiating ATG codon and terminating 360 nucleotides 3' to the last codon, a total of 4261 nucleotides.

Two "CCAAT" boxes are found in the 5' end of the gene, at positions 480-484 and at positions 517-521. A "TATA"-like sequence is found at positions 527-536, and the first nucleotide of the longest cDNA clone isolated, pCAB7-1B, is the A at position 561, 25 nucleotides downstream from the end of the TATA box. The last nucleotide of pCAB7-1B prior to the poly-A tract is the C at position 4120; another cDNA clone analyzed, pCAB7-2A, terminates at the T at position 4101. *Cab-7* contains four introns: the first intron (positions 659-2573) contains 915 bp, the second (1760-2122) 363 bp, the third (2161-2993) 833 bp, and the fourth intron (3509-3977) 377 bp.

The second intron is located in the same position as an intron in a recently identified PS II CAB gene

(E. Pichersky, unpublished); an intron occurs three codons downstream from this position in the PSI Type I CAB genes [3]. The other three introns in *Cab-7* do not occur at positions where any of the introns of previously characterized CAB genes are located. The first intron of *Cab-7* is unusual as it occurs in the middle of the region encoding the transit peptide (the first 41 or 42 codons [5]). In addition, its 5' boundary GC, is a rare variant of the consensus GT 5' splice site; as with the other such variants [4], the octamer AAGGCAAG around and including the site occurs.

The protein encoded by the tomato *Cab-7* gene is over 90% identical to the one encoded by the petunia LHCI-15 clone, but 12 of the 14 positions 45–58, at the mature N-terminus of the protein, are different, and no differences are observed in positions 157 to the end of the 270-residue protein. Comparison of the tomato CAB-7 protein and the petunia LHCI-15-encoded protein, which we designate PSI Type II CAB polypeptides, to the PSI Type I CAB polypeptide [1, 3] shows that the first two are substantially divergent from the latter. Sequence similarity is observed only within two conserved regions previously identified in comparisons between PSI and PSII CAB polypeptides [1, 5], and in these two regions the similarity between Type I and Type II PSI CAB polypeptides is only marginally greater than the similarity between any PSI CAB protein and PSII CABs. The pattern of steady-state level of *Cab-7* mRNA in the different tomato organs and under different light regimes is very similar to that of other types of toma-

AAGCTTAAATCTCGATTGGAATTAGTTCAAAAGTTTCAGTTTAAATATATACTTGCCAAAGGATTCATAGACGGAAGTTCAAATGACCCCCACCCCCAC 100  
CCCCCGCTCCGCTCTGCTAAGGGAGTTCATTTGAACATGATATGATTAATAATATGTAATAATTTTAAATATTTTAGAAAATTTTTAAATAATTTCTAAC 200  
TTCGACATCGATGAATAACCAATGAAGAGTTGAAGAATGAAGCATGAAAAAACAAGTGAATAAGGAAAAATAAATATTTACTTTGTTAAACAAAAACA 300  
ACAAATTCATTATGATCTCATAAAAGATTCGAAAAAATAAATATACACAAATTTTACCGAATCGAATCGAGAAGTACTTAATCCAACCACTATTAGAA 400  
TGGGGACATGATGAGTGTATAGAGGGGGTATAAGAACCAATATTTGGGTTGTTGGTTGCCACATGGCAATTTAAGTAGCCAATCACATATTGACTCTTC 500  
TATCCATCAAGATAAGCCAATTTCTCATAATAAAAAATCCACAAAATCTCATTGTCCTTGGTATCTCTCATAATCACAACACAAGAGTGAAGAATTTGCCG 600  
AC ATG GCC TCA GCT TGT GCT TCA TCC ACC ATT GCT GCT GGT GCT TTT TCT TCT CCA AG GCAAGTCTTTTCGATTTCATA 680  
M A S A C A S S T I A A V A F S S P S  
TCACTAATCATCAACGATAAACTATATACGATGATCATCGAAGTATAAGTGATCATCGTTGAATTCATGTAATATGTCCAGTCACTATGTCTTTAA 780  
ATAAGTTTACGATGATACTACTTTTATCTCATTCTATATTTTAAATAAAAAATAATTATCTCAAATTTTACTACTTTTAGAAGTTTAACTAAA 880  
AGGAAGACAATTAATAACTTAATTAATTTAGCGGCTATTTAATTTTCTAATTAATCATTTTGTGTTGATTTCTAGCATGACTTTGAGAATTTATATA 980  
CTCATTAAACGGAGATGATACATATATTAAGCATATTTAATTTAGAGATATTATAGCTTAAATCAATAATAATGTTTAAATAGTTTAAAAAAAATGTAA 1080  
TCAACAGCGGATATCATTATTTCTACTATGAAAATATCTTCTTTAAATTTTGTCTTACTTTTAAATGATATGATGATCTGTTTATACAAAATAGTTAA 1180  
CTAGTATTTTTTTGAAAAGTTATTACAAGAAAAGAAAAGGTCTAAGTATATTCGAATTTTAAATCAAATTTATTGTAACGATATCTTACTTTGGAAAAG 1280  
ACCTTTTACCGTGCACATTTAATAATGTATTTAAGATATATATATGTCGGTGGGACATATAAATAATTACATAATTATAATAGTAATGTGTCCACA 1380  
TTGACACATATGCTTAAAAATACCTATTAATAGTCAGGGTAAAAGATCTTCTATAAAATTTGATATTGTAACAACAATTTTCGACCAAAAGTTAAA 1480  
ATAGTTTTTCGACCAATTTCCATAATAAATTTAGTCTGAGTATAAAAAATACGATAAATAATTTGAAAATGTTTGTTCGAG T TCG 1577  
S  
AGG AGA AAT GGA TCA ATT GTA GGA ACA ACA AAA GCT TCA TTC CTT GGA GGA AGA AGA TTG AGA GTA AGC AAA TAC 1652  
R R N G S I V G T T K A S F L G G R R L R V S K Y  
AGT ACT ACA CCT ACA GCA CGA AGT GCA ACT ACA GTG TGC GTT GCT GCA GAC CCT GAC AGA CCC CTC TGG TTC CCT 1727  
S T T P T A R S A T T V C V A A D P D R P L W F P  
GGC AGC ACC CCA CCT CCA TGG CTT GAT GGC AG GTAATTTATGATCTAATTTTGATTTTTTCGAGTTTCATATCTGAATTAGTAGATTGT 1816  
G S T P P P W L D G S  
TCGTTCTTTGGTATAATAATAAATATATCAATTAACCTTTGAGTATGACACTTAAATTTGTATAAAGTTGATCAAGTAGGTAGACACACATGTCTTACG 1916  
TGACAATGAGTATTTTACGGTCTCCTACGTTGATATGACCACATATATATACTTTGTTCAATTTTATACGTAAGTACCCTGCATTTTATTAGTAA 2016  
AGACATATAAACTACTAATTTATTTCTTTCTTTGGAATGTTGCATTTTTCATATTGCTGAACTCAATGTTAATGTTTATTTGCTTCATCGCC 2116  
TTGCG C CTT CCC GGA GAC TTC GGT TTC GAT CCT CTT GGT CTA G GTATATATTCATTTACTCGTTCTATTTCAATTTATTAGTTT 2201  
L P G D F G F D P L G L  
AAGTTGCTGTTGAATCAATTTTTTTTAAAAAGAAAAGAAATGTTTGGACATCTTTTAAATTTCTCATGTCAAGATTAACAAGATTAAAGAGTGT 2301  
TTAATACATTTCAACATAATTTCTAGTTTCGGACCATGAATAGAAAGTTTTTTTTTACTTTTTAAAACCTCTTCCAAACGACGACTGGACAAAACAAATG 2401  
AAATGAAAGGAGTTTGAAGTTTCTAAGTTTGAAGAAAATATCGCTAGTACCCATGAAGATCTAATTAGTCTCAACGAGCGAATCCTTTTTTTACTA 2501  
CCAAACAACAAGTTTTCATGTTGATGCTGTTAATTAGGTTTAGGCCTTAACTCACAAGTAGTCCACTTTTATACCCTACTTACGTCGGTCTAGCATC 2601  
TGGTGCCTGAAACAATTTTTGATATTGAGGTTCCCAACATCCGGTGAGTCGGACACCACGCTGATACCATGTTAAATTAGGCTTAGGCATAACTCACATC 2701  
CAAAAGCTAACTCAAAGCGAGGAGGATGTCCAAACATTAATGGAGTCCACATGTTGGGCCCTTTTGTGCTTTTAAACCGATGCGGTTTCATACAATA 2801  
TACATAGTGAATTTCTAATCCGTAAGAAAATATCTCTAGAAATATGTAAGATCGAATTAAGTCTCAACGAGCAAATCCTGTTTTACTATTAATCACAC 2901  
AGTTTCATGTTTGCAGTAGTATTCATACATATGATATATGATATCTACTAATAACCAATTTTGTGATATTATATCATTTATGATACAG CA TCT 2998  
A S  
GAT CCA GAG AGC TTG AGA TGG AAC CAA CAG GCA GAA CTT GTA CAC TGC AGA TGG GCA ATG TTG GGA GCT GCA GGA 3073  
D P E S L R W N Q Q A E L V H C R W A M L G A A G  
ATT TTC ATT CCA GAA TTA CTC ACA AAA ATT GGA ATA CTT AAC ACA CCT TCA TGG TAC ACT GCT GGT GAA CAA GAA 3148  
I F I P E L L T K I G I L N T P S W Y T A G E Q E  
TAT TTC ACG GAC ACA ACA ACT CTC TTC ATC GTT GAA TTG GTA CTT ATT GGT TGG GCT GAG GGA AGA AGA TGG GCA 3223  
Y F T D T T T L F I V E L V L I G W A E G R R W A  
GAT ATC ATC AAG CCT GGA TGC GTT AAC ACG GAT CCA ATC TTC CCT AAC AAC AAG CTC ACG GGG ACA GAT GTT GGT 3298  
D I I K P P G C V N T D P I F P N N K L T G T D V G  
TAT CCT GGA GGT TTA TGG TTT GAT CCC TTG GGT TGG TCG GGT TCA CCT GCA AAG ATA AAG GAG TTG AGG ACA 3373  
Y P G G L W F D P L G W G S G S P A K I K E L R T  
AAG GAG ATT AAG AAC GGT AGA CTT GCT ATG TTA GCC GTT ATG GGT GCA TGG TTC CAA CAC ATT TAC ACT GGA ACA 3448  
K E I K N G R L A M L A V M G A W F Q H I Y T G T  
GGC CCT ATT GAT AAC CTA TTT GCT CAC CTA GCT GAT CCT GGT CAT GCC ACT ATT TTT GCT GTAAGTACAACCTCATTTTA 3528  
G P I D N L F A H L A D P G H A T I F A  
GTACTCCTCTGTTTCGTTTAAATTTGCTTACGTCCTTTTGATAACCTATTTGCTCACCAGAAAGTCTTTTAAATGTTTAAATATGTTCAAGTTAAAA 3628  
CTAAACAGACAAAATTGAGACGAGGATGATGCTGTTGTTGGTTTGAAGAAAACCTAAACAGAGAAATGTTGTTGTTTACTGTTTATATGGAGAAA 3728  
CTATATTTTACACTGTTTTGTTAACTAATAGCTATTTCTTTGAAAAAATCTGGGCTTGGTACACTGATAATCATCTTCCGTTGTTCAATTATAT 3828  
GCACATTCGCTATTCGATTTCAGTACTTGAAGTTGTTTCTACTTATATTTTTCAG GCA TTC AGT CCC AAG TGAGAAGCAGAAGCAGAACC 3922  
A F S P K  
TTTTGTTGTAATGATGTGTAAAAGTAACATTTGAAGTAGATTTGTGCAAGAAAACCTTAAATGTTTCTTTTGGCACTCAATGTAGTATACCTACATTTG 4022  
GTTTGCCCTCTGTGAAAAAGACTCAGTGTACAATAACCTTTGAGAATTCAGTACAGTCAAAAAAATCTATTTATATGTTTCGTCCTCCATTTCGCAT 4122  
TAGAGAAGATCTAATTTTACCCTAATCGTTGACTAAAAATAACATAGTGCAGGCTTGTTTTTTCTATAATTTCTATGGGTGCTCGATGTAATAATAGCAC 4222  
AAGTAAAGAATAACGTTCTCTACTACTATAGATA 4261

Fig. 1. Nucleotide sequence of the *Cab-7* gene and amino acid sequence of the encoded polypeptide. The presumed "CCAAT" and "TATA" boxes, the first nucleotide of cDNA clone pCAB7-1B, and the last nucleotides in the cDNA clones pCAB7-1B and pCAB7-2A are underlined.

to CABs, and the level of expression of *Cab-7* is similar to that of the Type I PSI CAB genes.

### Acknowledgements

This work was supported by a USDA grant BARD G-I-1100-86 and a University of Michigan's Rackham Faculty Research Grant to EP, and an NSF grant DCB-8517773 to SDT.

### References

- Hoffman NE, Pichersky E, Malik VS, Castresana C, Ko K, Darr SC, Cashmore AR: The nucleotide sequence of a tomato cDNA clone encoding a photosystem I protein with homology to photosystem II chlorophyll a/b-binding polypeptides. *Proc Natl Acad Sci USA* 84: 8844–8848 (1987).
- Pichersky E, Hoffman NE, Malik VS, Bernatzky R, Tanksley SD, Szabo L, Cahsmore AR: 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 (1987).
- Pichersky E, Hoffman NE, Bernatzky R, Piechulla B, Tanksley SD, Cashmore AR: 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 (1987).
- Shapiro MB, Senapathy P: RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucl Acids Res* 15: 7155–7174 (1987).
- Stayton MM, Brosio P, Dunsmuir P: Characterization of a full-length petunia cDNA encoding a polypeptide of the light harvesting complex associated with photosystem I. *Plant Mol Biol* 10: 127–137 (1987).
- Sugita M, Manzara T, Pichersky E, Cahsmore AR, Gruissem W: Genomic organization, sequence analysis and expression of all five genes encoding the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase from tomato. *Mol Genet* 209: 247–256 (1987).