Volume 17 Number 15 1989 Nucleic Acids Research

Molecular analysis of eight U1 RNA gene candidates from tomato that could potentially be transcribed into U1 RNA sequence variants differing from each other in similar regions of secondary structure

Steffen Abel§, Tamás Kiss+ and Ferenc Solymosy*

Institute of Plant Physiology, Biological Research Center, Hungarian Academy of Sciences, 6701 Szeged, PO Box 521, Hungary

Received March 23, 1989; Revised and Accepted July 7, 1989 EMBL accession nos X14412–X14419 (incl.)

ABSTRACT

From a tomato genomic library we isolated and characterized eight U1 RNA gene candidates (U1.1 to U1.8) all of which possessed the canonical plant U-snRNA transcription signals in their 5' and 3' flanking regions and exhibited nucleotide sequence conservation in the 5' splice site recognition sequence, in the Sm antigen binding site and in Loops B, C, D as well as in Stems III and IV of their coding region. Deviations from the U1 RNA consensus sequence were mainly localized to Loop A and Stems I and II, suggesting that the putative transcripts of the tomato U1.1 - U1.8 genes would differ from each other in their capacity of binding to the U1 RNA-specific snRNP proteins.

INTRODUCTION

Five (U1, U2, U4-U6 RNAs) out of the six major, uridylate-rich, capped small nuclear RNAs (U-snRNAs) (1) are known to play, in the form of ribonucleoprotein particles (U-snRNPs) (2), an indispensable role in the splicing of pre-mRNA (3). This nuclear activity, mostly on account of alternative splicing (4), seems to be one of the factors contributing to the post-transcriptional regulation of gene expression. Whereas the precise molecular mechanisms leading to alternative splicing are not known, sequence variants of individual U-snRNAs are reasonable candidates for being modulators in this process (5).

The most straightforward way of detecting and characterizing these variant molecules consists in the isolation and sequencing of the U-snRNAs themselves (6-8). This approach, however, may not always be practicable because some U-snRNA variants that occur in minute amounts (9) in the nucleus may escape detection. The structural analysis of U-snRNA genes, on the other hand, could reveal even those putative sequence variants of individual U-snRNAs which would not easily be detected at the RNA level. There are a number of metazoan U-snRNA genes which contain point mutations in their coding region [(10) and references therein] as compared to the

© IRL Press

primary structure of the most abundant variant sequenced at the RNA level, and which could, hence, give rise to potential sequence variants of individual U-snRNA species. Although many of them [Class II to IV pseudogenes (11)], owing, first of all, to the absence of consensus promoter/enhancer elements in their flanking regions as well as to the occurrence of direct repeats therein and/or truncation of their coding regions, are definitely pseudogenes, some of them [Class I pseudogenes (11)] may well be transcribed into U-snRNA sequence variants, because, in spite of point mutations in their coding region, they do contain the consensus promoter/enhancer elements necessary for transcription *in vivo*. Therefore, at least part of the U-snRNA Class I pseudogenes could well embody *bona fide* genes and analyzing the primary and secondary structure of their coding regions might lead to a better understanding of the molecular mechanism of U-snRNA action in the splicing process.

In this paper we report the isolation and structural analysis of eight potential *bona fide* genes [Class I pseudogenes according to the nomenclature of Denison and Weiner (11)] for U1 RNA from a tomato genomic library. Our results are in line with the assumption that these genes, if transcribed, encode U1 RNA sequence variants differing from each other in the helix stability of Stems I and II and also in the nucleotide sequence of Loop A, suggesting that they would vary in their protein binding ability.

MATERIALS AND METHODS

Isolation of tomato U1 RNAs and 3'end-labeling

RNA from the U1 RNA region of a 10 % denaturing polyacrylamide gel loaded with a tomato (*Lycopersicon esculentum* Mill) nuclear RNA preparation [prepared by Method I of (12)] was extracted according to Kiss *et al.* (12) and was 3'end-labeled with [^{32}P]pCp by T4 RNA ligase (13). The labeled RNA was immunoprecipitated with anti-m_3G IgG (kindly supplied by R. Lührmann and P. Bringmann) according to Krol *et al.* (6), repurified on a polyacrylamide gel and used for hybridizations.

Screening of the genomic library

A genomic library (kindly provided by R.W. Breidenbach) constructed by inserting *Lycopersicon esculentum* Mill DNA partially digested with *Eco* RI in the Charon 4 vector was screened (approximately 50,000 plaques) by the *in situ* hybridization technique described by Benton and Davis (14), using 3'end-labeled tomato U1 RNA as a probe. Hybridizations were performed in 50% formamide, 5 x SSC, 0.1 % SDS, 1 mM EDTA, 10 mM Tris-HCl, pH 7.5, and 1 x

Denhardt's reagent [0.02 % (w/v) bovine serum albumin, 0.02 % (w/v) polyvinylpyrrolidone and 0.02 % (w/v) Ficoll 400]. Filters were washed twice in 2 x SSC for 30 min at 37 °C. After 3-4 rounds of plaque-purification, 19 recombinant phage were isolated which hybridized strongly to the tomato U1 RNA probe. The Charon 4 phage were propagated according to Maniatis et al. (15) and phage DNA was isolated as described in (16).

Molecular cloning and sequencing strategy

All techniques for manipulating DNA were carried out according to (15), unless stated otherwise. Five U1 RNA gene-containing EcoRI/EcoRI fragments (4.8-10.5 kb) from different phage (L1 to L5, see Fig. 1 for restriction maps) were cloned into the pBluescribe M13(-) vector (Stratagene) and mapped with restriction endonucleases. Positive fragments were isolated from agarose gels and further subcloned into pBluescribe M13(+/-) vectors. Single-stranded DNA was isolated using R408 helper phage as recommended in (17). Sequencing was performed essentially by the procedure commonly used for M13 phage vectors (18).

Fragments containing the genes U1.1 (PstI/HindIII, 0.6 kbp) and U1.3 (AluI/AluI, 1.4 kbp) from phage L1, and U1.8 (BglII/SalI, 1.0 kbp) from phage L5 were directly sequenced after subcloning in pBluescribe M13 (+/-) vectors. The orientation of gene U1.3 was determined by complementation test. Genes U1.5 and U1.6 contained a restriction site within the hybridizing region also present in the polylinker sequence of the vector. Therefore, gene U1.5 was sequenced using the subcloned 0.5 kbp XbaI/SphI fragment and the SphI-deletion of the 10.5 kbp EcoRI/EcoRI fragment of phage L2, and gene U1.6 using the subcloned 1.1 kbp HindIII/SalI fragment and the SalI-deletion of the 4.8 kbp EcoRI/EcoRI fragment of phage L3. Larger U1 gene-containing fragments from phage L1 (HindIII/PstI, 1.4 kbp; U1.2), phage L2 (BglII/XbaI, 2.9 kbp; U1.4) and phage L4 (HindIII/BamHI, 1.9 kbp; U1.7) were sequenced after creating overlapping sets of deletions by unidirectional size reduction using the exonuclease III approach (19). The enzymes and the protocol followed were those from Stratagene.

Mapping of the 5' ends of tomato U1 RNAs

The 5' termini of U1 RNAs were determined by sequencing single-stranded sense DNA containing the coding region of the gene U1.7 and by reverse transcription of a tomato U1 RNA population, using for both reactions a 5' end-labeled oligodeoxyribonucleotide primer complementary to positions 98-139 in the tomato gene U1.6. The primer was obtained by subcloning the internal SalI/HaeIII fragment of the coding region of gene U1.6 into

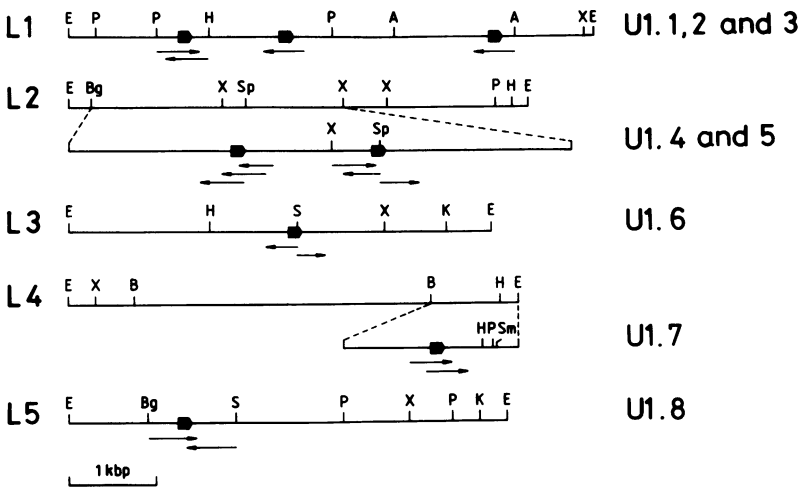


Fig. 1. Restriction maps of the 4.8-10.5 kbp EcoRI-EcoRI lambda DNA fragments containing different U1 RNA genes (numbered U1.1 through U1.8). Thin arrows indicate sequenced regions. Arrow-shaped boxes indicate coding regions in the order of designations at the right and direction of potential transcription. A, B, Bg, E, H, K, P, S, Sm, Sp and X are AluI, BamHI, BglII, EcoRI, HindIII, KpnI, PstI, SalI, SmaI, SphI and XbaI sites, respectively. The scale refers to the fragments carrying the coding regions as indicated.

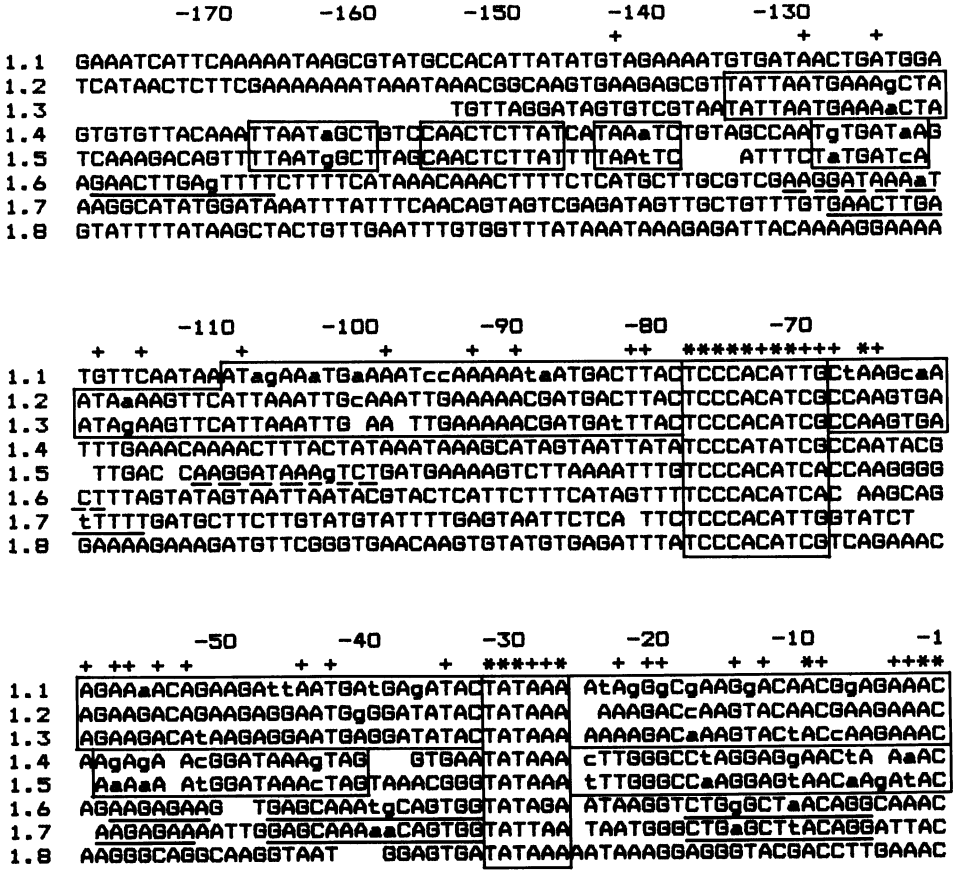
SalI/SmaI-cleaved pBluescribe M13 vector DNA, followed by 5' end-labeling of the resulting SalI/KpnI fragment and by separating the sense strand (55 nts) from the antisense strand (47 nts) on a 10 % polyacrylamide/8 M urea gel. The first 45 nucleotides of the antisense strand are complementary to the SalI/HaeIII fragment of gene U1.6.

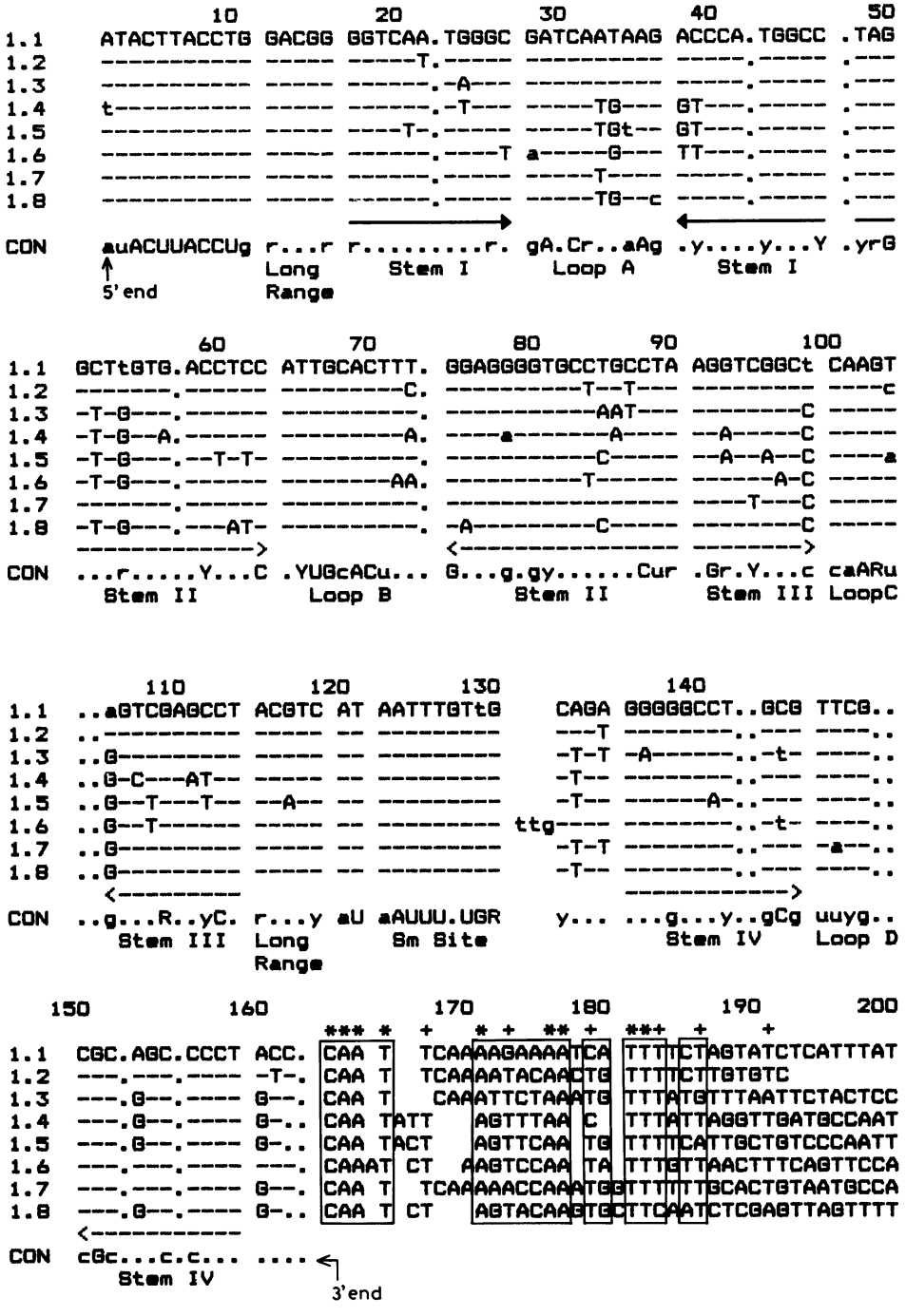
Transient expression of gene U1.1 in Orychophragmus violaceus protoplasts

Orychophragmus violaceus cell suspension cultures were maintained in MS medium (20) supplemented with 2% sucrose, 100 mg/ml inositol and 0.1 mg/ml of 2,4 D. Protoplasts were isolated from 4-5-day cultures by digestion of the cell walls for one hour at 26^{o} C in 0.1% Pectolyase Y23 (Seishin Pharmaceutical Co., Ltd., Japan), 1% Cellulase Onozuka R10(Yakult Honsha Co., Ltd., Japan), 0.4 M D-mannitol and 5 mM MES, pH 5.5. Protoplasts were filtered through a 50 µm sieve and washed with W5 medium (21). PEG-mediated transfection of protoplasts with 10 µg of pBS (+) recombinant plasmid

carrying a 0.6 kbp PstI/HindIII fragment (U1.1 locus) of L1 recombinant phage (see Fig. 1) and RNA isolation from transfected protoplasts were performed essentially as described by Vankan et al. (22) and Goodall et al. (23), respectively.

An RNA probe complementary to U1 RNA was synthesized by T3 polymerase (24) using [α-^{32}P] GTP and PstI-linearized pBS (+) recombinant plasmid harboring the tomato U1.1 gene (see above). RNase A/T1 protection was carried out as described by Vankan et al. (22). The resulting RNA produts were separated on a 6% sequencing gel.




10 20 30 40 50
1.1 ATACTTACCTG GACGG GGTCAA.TGGGC GATCAATAAG ACCCA.TGGCC .TAG
1.2 ---------- ----- -----T.----- ---------- -----.----- .---
1.3 ---------- ----- ------.-A--- ---------- -----.----- .---
1.4 t--------- ----- ------.-T--- -----TG--- GT---.----- .---
1.5 ---------- ----- ----T-.----- -----TGt-- GT---.----- .---
1.6 ---------- ----- ------.----T a-----G--- TT---.----- .---
1.7 ---------- ----- ------.----- -----T---- -----.----- .---
1.8 ---------- ----- ------.----- -----TG--c -----.----- .---
CON auACUUACCUg r...r r.....r. gA.Cr..aAg .y....y...Y .yrG
5'end Long Range Stem I Loop A Stem I
60 70 80 90 100
1.1 GCTtGTG.ACCTCC ATTGCACTTT. GGAGGGGTGCCTGCCTA AGGTCGGCt CAAGT
1.2 --------.------ ---------C. ----------T--T--- --------- ----c
1.3 -T-G---.------ ----------. ----------AAT--- --------C -----
1.4 -T-G--A.------ ---------A. ----a-------A----- --A-----C -----
1.5 -T-G---.--T-T- ----------. ----------C------ --A--A--C ----a
1.6 -T-G---.------ --------AA. ----------T------ ------A-C -----
1.7 --------.------ ----------. ---------------- ----T---C -----
1.8 -T-G---.---AT- ----------. -A---------C----- --------C -----
CON ...r.....Y...C .YUGcACu... G...g.gy......Cur .Gr.Y...c caARu
Stem II Loop B Stem II Stem III LoopC
110 120 130 140
1.1 ..aGTCGAGCCT ACGTC AT AATTTGTtG CAGA GGGGGCCT..GCG TTCG..
1.2 ..---------- ----- -- --------- ---T --------..--- ----..
1.3 ..G--------- ----- -- --------- -T-T -A------..-t- ----..
1.4 ..G-C---AT-- ----- -- --------- -T-- --------..--- ----..
1.5 ..G--T---T-- --A-- -- --------- -T-- -------A..--- ----..
1.6 ..G--T------ ----- -- --------- ttg---- --------..-t- ----..
1.7 ..G--------- ----- -- --------- -T-T --------..--- -a--..
1.8 ..G--------- ----- -- --------- -T-- --------..--- ----..
CON ..g...R..yC. r...y aU aAUUU.UGR y... ...g...y..gCg uuyg..
Stem III Long Range Sm Site Stem IV Loop D
150 160 170 180 190 200
1.1 CGC.AGC.CCCT ACC. CAA T TCAAAAGAAAATCA TTTTCTAGTATCTCATTTAT
1.2 ---.---.----- -T-. CAA T TCAAAATACAACTG TTTTCTTGTGTC
1.3 ---.G--.----- G--. CAA T CAAATTCTAAATG TTTATGTTTAATTCTACTCC
1.4 ---.G--.----- G-.. CAA TATT AGTTTAA C TTTATTAGGTTGATGCCAAT
1.5 ---.G--.----- G-.. CAA TACT AGTTCAA TG TTTTCATTGCTGTCCCAATT
1.6 ---.---.----- ---. CAAAT CT AAGTCCAA TA TTTGTTAACTTTCAGTTCCA
1.7 ---.---.----- G--. CAA T TCAAAAACCAAATGGTTTTTTGCACTGTAATGCCA
1.8 ---.G--.----- G-.. CAA T CT AGTACAAGTGCTTCAATCTCGAGTTAGTTTT
CON cGc...c.c... .... 3'end
Stem IV

RESULTS

Isolation of tomato U1 RNA genes

By screening a tomato genomic DNA library with 3' end-labeled tomato U1 RNA as a plaque-hybridization probe, six different positive EcoRI/EcoRI fragments were obtained (Fig. 1). Three of them, deriving from phage L3 to L5, contained a single and distinct region hybridizing to the probe, one fragment each from phage L1 and L2 contained more than one positive regions located close to each other (three and two tandem repeat elements, respectively), and one fragment has previously been shown to carry a U1 RNA-related pseudogene severely truncated at the 3' end of its "coding region" (25). Positive subfragments were further cloned into the pBluescribe M13 vector and sequenced either directly or after unidirectional size reduction by exonuclease III treatment. The sequences of the eight tomato U1 RNA gene candidates (U1.1 to U1.8) are compiled in Fig. 2.

5' and 3' termini of tomato U1 RNA

Owing to the occurrence in tomato nuclei of several variants of U1 RNA [cf. Lane C of Fig. 1 in (8)] we were not able to directly determine either the 5' or the 3' termini of the tomato U1 RNAs. Therefore, we turned to an indirect approach in both cases.

The 5' end of the tomato U1 RNAs was determined (Fig. 3) by primer extension, using a 45-nt internal SalI/HaeIII fragment of the coding region

Fig. 2. Sequences of U1 RNA genes of tomato. Numbering refers to gene U1.1. Other sequences are aligned with U1.1 allowing gaps to improve aligment in the noncoding regions and to separate secondary structural elements of the putative transcripts in the coding regions. The coding regions are aligned with the recently published consensus sequence of U1 RNAs (29). Dots in the consensus sequence stand for any nucleotide, and those in the aligned sequences of the coding regions denote deletions with respect to the consensus sequence. Hyphens indicate nucleotide residues which are identical in the coding regions with those of U1.1. Capital letters designate matches with, and lower case letters indicate deviations from, the consensus sequence. Gene U1.6 contains a trinucleotide insertion between positions 130 and 131 of the standard numbering. In the non-coding regions asterisks and crosses indicate identical nucleotides in all or in at least six out of the eight sequences, respectively. In the 5' flanking region the TATA-homology-like sequences and the motives around position -70 in all genes as well as conserved sequences between genes contained in the same recombinant lambda phage (U1.1, U1.2, U1.3 and U1.4, U1.5, respectively) are boxed. Sequences conserved between genes U1.5 and U1.6 as well as U1.6 and U1.7 are indicated by broken and continuous lines, respectively; mismatches therein are indicated by lower case letters. In the 3' flanking region, aligned for maximum sequence similarity, sequence blocks from which the consensus sequence in Fig. 6 has been constructed, are boxed.

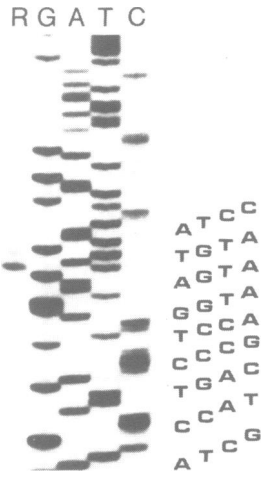


Fig. 3. Mapping of the 5' ends of a tomato U1 RNA population by primer extension assay. Reverse transcription of tomato U1 RNAs (Lane R) and sequencing of U1.7 single-stranded sense DNA (Lanes G, A, T and C) were performed as described in Materials and methods.

of gene U1.6 (cf. Fig.2) as a primer. This fragment was chosen because (i) it is partially complementary to a stretch (from nt 98 to nt 139) of the coding region of all the eight tomato U1 RNA genes and (ii), owing to the presence of two restriction sites at a reasonable distance and location within the coding region, it seemed to be the most suitable candidate for subcloning. The data (Fig. 3) indicated that the first transcribed nucleotide corresponded to an A residue (cf. Fig. 2). This result, together with the immunoprecipitability of the RNA with anti-m_3G antibody (cf. Materials and methods) and with the highly conserved nature of the 5' end in all other U1 RNA species so far examined (7) strongly suggests that the 5' end of tomato U1 RNA sequence variants, with the exception of the hypothetical transcript of the U1.4 gene (cf. Fig. 2), is m_3GAUACUU. The appearance in the sequencing gel (Fig. 3, Lane R) of a strong band corresponding to an A residue in position 3 is most probably due to the fact that both the A and U residues following the cap structure are methylated in tomato U1 RNA, similarly to all other U1 RNAs sequenced at the RNA level (7) and this leads to the falling off of reverse transcriptase.

The 3' end of the tomato U1 RNAs was determined by digesting a 3' end-labeled, hybrid-selected tomato U1 RNA population with RNase T2 and analyzing the ribonucleolytic products by thin layer chromatography. The exclusive

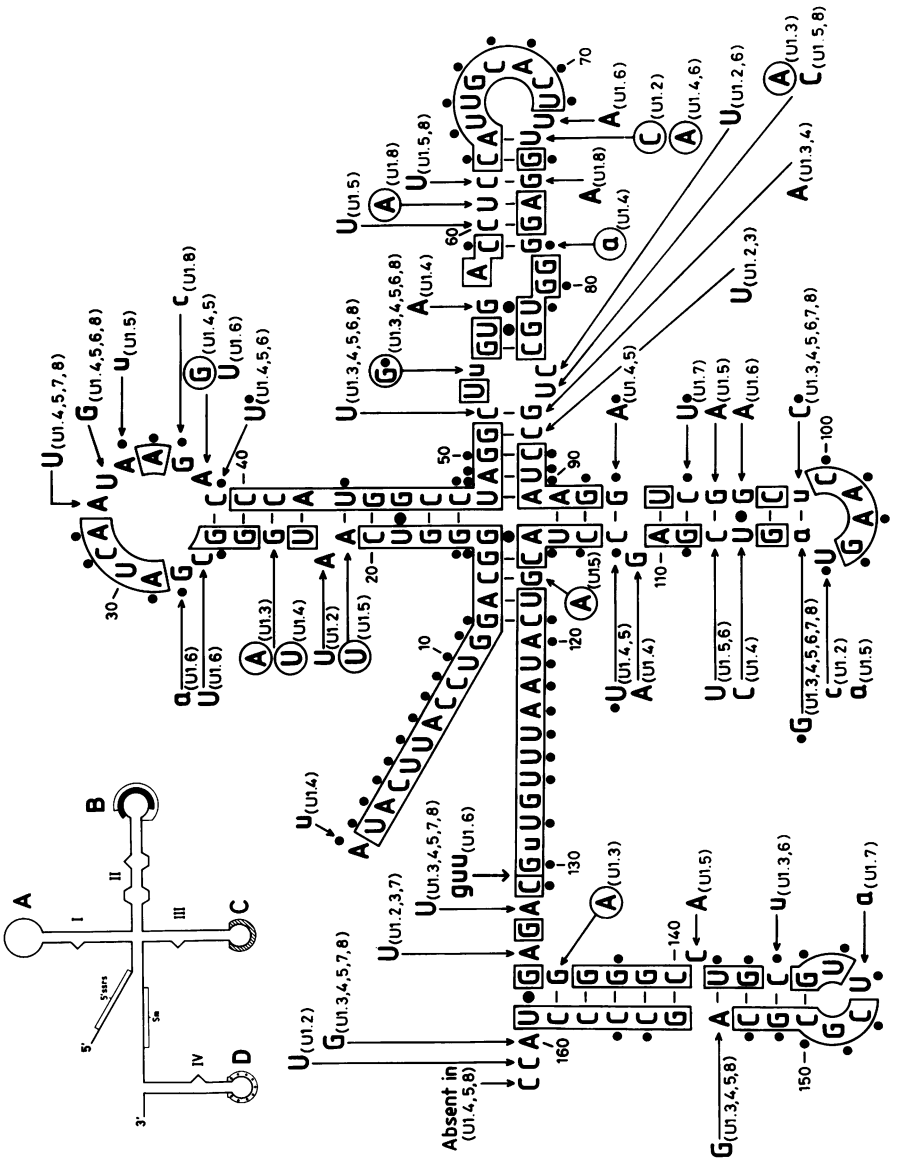
formation of labeled 2',3' cyclic CMP (not shown) revealed that virtually all of the tomato U1 RNA species are terminated by a cytidine residue. The 3' end (cf. Fig. 2) was deduced from the estimated length of tomato U1 RNAs (about 162-163 nts as determined by polyacrylamide gel electrophoresis under denaturing conditions), from a comparison with the 3' termini of U1 RNAs of pea (6), common bean (26), soybean (27), and broad bean [our unpublished results, sequence shown in (26)] as well as from a comparison with conserved downstream sequence elements in plant U-snRNA genes (22, 28).

Primary and secondary structure of the putative transcripts of the tomato U1 RNA genes

As compared to gene U1.1, the coding regions of tomato U1 RNA genes differ in 7 (U1.2) to 23 (U1.5) positions (Fig. 2) displaying sequence similarities of 85-96 % to each other. The sequence similarities between the coding regions of multicopy tandem repeat units, U1.1 to U1.3 and U1.4/U1.5, are 90-96 % and 88 %, respectively. The similarity between the coding regions of tomato U1 RNA genes and those of bean and soybean U1 RNAs is 80-85 %, that to human U1 RNA about 60 %. The sequence of all eight U1 RNA variants agrees fairly well with the U1 RNA consensus sequence recently established by Guthrie and Patterson (29). The regions of experimentally proven functional significance (5' splice site recognition sequence, Sm antigen binding site) are conserved (for deviations, see below).

The putative transcripts of tomato U1 RNA genes can be folded into the experimentally determined secondary structure of chicken, rat, and human U1 RNAs (30) and those proposed for Drosophila (31), common bean (26), soybean (27) and Chlorella (32) U1 RNAs (Fig. 4). This phylogenetically highly conserved secondary structure consists of four stem/loops, a single-stranded 5' region, and a single-stranded region separating stems III and IV and containing the Sm antigen-binding site. These two single-stranded regions, as well as the sequences of loops II-IV are most conserved in all U1 RNAs (29, 32).

With respect to sequence conservation (29) the nucleotides in the putative transcripts of tomato U1.1 to U1.8 genes fall into eight groups: (i) those which are conserved (with transition permitted) both in the U1 RNA consensus sequence and in all tomato U1 RNA variants (e.g. U_2 through G_{12}, U_{65} through U_{71}, C_{119} through U_{126}, C_{148} through C_{152}, etc.); (ii) those which are variable in the U1 RNA consensus sequence but are conserved (with no transition permitted) in all tomato U1 RNA variants (e.g. G_{18} to C_{20}, C_{40} to A_{42}, G_{44} to C_{46}, C_{157} to U_{159}, etc.); (iii) those which are conserved (with


A
B
C
D
I
II
III
IV
5'
3'
5'sss
3a
Absent in (U1.4,5,8)

no transition permitted) in the U1 RNA consensus sequence but diverge from the consensus in all (U_{129}) or some (e.g. U_1 in U1.4, A_{28} in U1.6, U_{35} in U1.5, C_{37} in U1.8, U_{99} and A_{105} in U1.1 and U1.2, etc.) of the tomato U1 RNA variants, without altering the secondary structure of the molecule; (iv) those which are variable both in the U1 RNA consensus sequence and in the tomato U1 RNA variants (e.g. nt positions 22, 33, 34, 52, 57, 60, 62, etc.) and do not alter the secondary structure of the molecule; (v) those which are conserved (with no transition permitted) in the U1 RNA consensus sequence but diverge from the consensus in some of the tomato U1 RNA variants (U_{54} in U1.1, U1.2 and U1.7 and A_{78} in U1.4) and alter the secondary structure of the molecule, (vi) those which are variable both in the U1 RNA consensus sequence and in some of the tomato U1 RNA variants (positions 21, 24, 38, 61, 73, 85, 117 and 136) and do alter the secondary structure of the molecule. In addition, (vii) insertion between positions 130 and 131 of a UUG sequence occurs in the putative transcript of U1.6 and (viii) C_{162} seems to be either present or absent, depending on the U1 RNA variant.

Structure of the non-coding regions in the tomato U1 RNA genes

The 5' upstream non-coding regions of the eight tomato U1 RNA genes exhibit sequence similarities to different extents, mostly concentrated in blocks of identity (Fig.2). As shown in Fig. 5 the 5' flanking sequences in all eight U1 genes contain two highly conserved sequence elements also present in all bona fide plant U-snRNA genes studied so far (22, 28): (i) a TATA homology-like motif positioned 22 to 27 nucleotides upstream of the cap site, and (ii) a sequence element located 32 to 36 nucleotides upstream of the TATA motif in the -70/-80 region which has previously been shown to be required for the expression of Arabidopsis U2 RNA genes (28). The degree of overall sequence conservation in the 5' upstream region (-1 to -80) is significantly

Fig. 4. Secondary structure of tomato U1 snRNA. The sequence of a hypothetical non-capped transcript of gene U1.1 is shown. Nucleotides that are invariant at identical positions in all eight (U1.1 to U1.8) tomato genes are boxed. Nucleotide changes with respect to the U1.1 RNA sequence are marked by arrows together with the gene(s) (in parentheses) in which they occur. Nucleotide substitutions affecting secondary structure are encircled. Evolutionarily highly conserved nucleotides in the consensus sequence (cf. Fig. 2.) are marked by dots and nucleotides which deviate from those in the consensus sequence are in lower case letters. Stems (roman numerals) and loops (capital letters) as well as regions of functional importance are identified in the inset: 5' ssrs, 5' splice site recognition sequence; Sm, Sm antigen binding site; ▬ and ▨, loop sequences complementary to U2 RNA loop sequences (32); ▭ and ▭, loop sequences similar to U2 RNA loop sequences (32).

```
                        "-70 Box"                  TATA Element
                                                                  -1
U1.1          ..... CTCCCACATTG ............... TATAAAAT .........C
U1.2          ..... CTCCCACATCG ............... TATAAAAA .........C
U1.3          ..... CTCCCACATCG ............... TATAAAAA .........C
U1.4          ..... ATCCCATATCG ............... TATAAACT .........C
U1.5          ..... GTCCCACATCA ............... TATAAATT .........C
U1.6          ..... TTCCCACATCA ............... TATAGAAT .........C
U1.7          ..... CTCCCACATTG ............... TATTAATA .........C
U1.8          ..... ATCCCACATCG ............... TATAAAAA .........C

                                                        **
                                                        AA
CONSENSUS (A) ..... NTCCCAcATYR ............... TATaaA-- .........C
                                                        YT
                                  (32-36 nt)              (22-27 nt)

                                                        **
                                                        TA
CONSENSUS (B) ..... RTCCCACATcG ............... TAtAaa-- .........C
                                                        RY
                                  (33-34 nt)              (23-27 nt)
```

Fig. 5. Derivation of a consensus (A) sequence for the 5' flanking region of the tomato U1 RNA gene candidates U1.1 to U1.8. The "-70 box" (28) and the TATA element (28) are indicated. Invariant residues are denoted by capital letters; N indicates any nucleotide; R and Y stand for an invariant purine and pyrimidine, respectively. A residue which is identical in all compared sequences except one is designated by a lower case letter. At the positions marked by an asterisk (*) the nucleotide residue in the numerator is as abundant as, or more abundant than, in the denominator. The spacings between the "-70 box" and the TATA element and between this latter and the cap site are indicated by the number of residues in parenthesis. For comparison, the consensus (B) and spacings obtained by Vankan et al. (22) for the 5' flanking regions of U-snRNA genes from Arabidopsis thaliana (22, 28), common bean (26) and soybean (27) are given in the bottom line. The symbols in consensus B are the same as in consensus A.

higher between genes organized as tandem arrays (i.e. U1.1 to U1.3, 73-91 %; and U1.4/U1.5, 67 %) than between those (U1.6, U1.7, U1.8; 53-58 %) isolated as single fragments or between genes from different EcoRI/EcoRI fragments (38-62 %). Further upstream (up to -135), pronounced sequence similarity exists only between genes U1.1, U1.2, and U1.3, which might be a reflection of successive duplication events. However, in the -80 to -180 region several identical or similar short sequence elements are shared by pairs of genes (Fig. 2), but no sequences common to all the genes seem to be present in this region.

In the 3' non-coding region (Figs 2 and 6) the plant U-snRNA gene-specific consensus sequence reported in (22) was present in, and could be extended considerably for, all the eight tomato U1 genes.

```
U1.1              CAA T....AAGAAAA.CA TTT.CT.....
U1.2              CAA T....AATACAA.TG TTT.CT.....
U1.3              CAA T ...ATTCTAA.TG TTT.TG.....
U1.4              CAA T... AGTTTAA C  TTT.TT.....
U1.5              CAA T... AGTTCAA TG TTT.CA.....
U1.6              CAA.T.. .AGTCCAA TA TTT.TT.....
U1.7              CAA T....AAACCAA.TG.TTT.TT.....
U1.8              CAA T..  AGTACAA.TG.TTC.AT.....

                           ***            *
                           TYY            T
CONSENSUS (A)     CAA.T....Ar---AA.Yr.TTt.y-.....
                       (2-4  RAA           R
                        nt)

                             *
                             A
CONSENSUS (B)     CA.......AGtN-Aa....
                    (4-9 nt)   T
```

Fig. 6. Derivation of a consensus (A) sequence for the 3' flanking region of the tomato U1 RNA gene candidates U1.1 to U1.8. The residues CA abut on the 3' ends of the coding regions. For comparison, the consensus (B) and spacing obtained by Vankan et al. (22) for the 3' flanking region of U-snRNA genes from Arabidopsis thaliana (22, 28) common bean (26) and soybean (27) are given in the bottom line. Nucleotide residues that are not considered in the construction of the consensus sequence are marked by dots. For additional explanation of the symbols used see the legend to Fig. 5.

Detection of the expression of gene U1.1 in a heterologous system

To examine whether the regulatory signals present in the non-coding regions of a representative tomato U1 RNA gene are functional, the transient expression of gene U1.1 in a heterologous system was investigated. Fig. 7 shows that the tomato U1.1 gene introduced into O. violaceus protoplasts was faithfully transcribed as tested by an RNase A/T1 protection assay. Owing to the high overall nucleotide sequence conservation of U1 RNA species in all organisms tested so far, some transcript(s) of the O. violaceus resident U1 gene(s) yielded a negligible protection to the U1.1-specific probe which thus escaped degradation by RNase A and RNase T1 (Fig. 7, Lane C). The specificity of the RNase A/T1 protection assay, however, is duely reflected by the pronounced difference mainly in relative mobility but also in intensity of labeling between the U1 band in Lane B and that in Lane C of Fig. 7.

DISCUSSION

The results presented in this paper indicate that in the tomato genome there are at least five loci which could encode eight (U1.1 to U1.8) distinct U1 RNA variants. Although we do not have direct evidence to show that all the

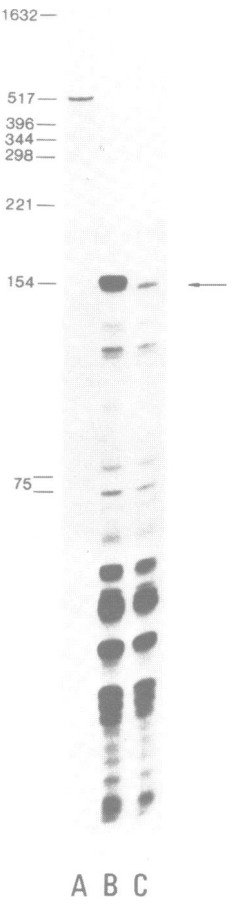


Fig. 7. Expression of the tomato U1.1 gene in transfected Orychophragmus protoplasts. Lane A. Undigested RNA probe. Lane B. RNase A/T1 mapping of RNA isolated from protoplasts transfected with recombinant pBS (+) plasmid carrying the 0.6 kbp PstI/HindIII fragment of L1 recombinant lambda phage. Lane C. RNase A/T1 protection analysis of RNAs from mock-transfected protoplasts. The positions of molecular markers (5' end-labeled Hinf I digest of pBR_{322}) are indicated on the left and the position of U1 RNA is marked by an arrow.

above loci represent bona fide genes rather than pseudogenes, the following considerations argue for their being true genes transcribed into U1 RNA variants:

(i) RNA molecules encoded by U1.1 were present (Fig. 7) in O. violaceus protoplasts transfected with a recombinant plasmid carrying the U1.1 locus of recombinant phage L1.

(ii) Both the spacing and the sequences of the transcription signals in both the 5' and 3' flanking regions of tomato U1.1 to U1.8 genes fit well (Figs 5 and 6) into the consensus structure deduced (22, 28) for experimentally established bona fide U2 and U5 RNA genes of Arabidopsis thaliana. We note that the consensus sequence of the 3' flanking region of plant U-snRNA genes could be extended downstream of the putative transcription termination or RNA processing signal [cf. (22), and Figs 2 and 6]. Since both a C residue at the 3' end of the 5' flanking region and a CA residue at the 5' end of the 3' flanking region are invariant in all plant U-snRNA gene consensus sequences constructed to date (Figs 5 and 6), these were taken as markers to indicate the length of the putative transcripts of the tomato U1.1 to U1.8 genes.

(iii) Both in animals [cf. (7)] and plants (6, 8) sequence variants of individual U-snRNA species have been detected at the RNA level and in Arabidopsis thaliana at the gene level (28). Expression of these U2 RNA genes in Arabidopsis has been given experimental support (28). Recently, Lund (9) reported the isolation and sequencing of a number of U1 RNA variants from human cultured cells. Minor variants of the human U1 RNA comprised between 5 % and 15 % of the total U1 RNA population in these cells, and nucleotide replacements found in some of them led to a change in secondary structure (9).

(iv) If U1.1 to U1.8 were pseudogenes, sequence variation in the coding region would be expected to be random. In fact, however, single-stranded regions proven or thought to be of functional importance in a sequence-dependent manner are identical in most putative transcripts of tomato U1.1 to U1.8 (Fig. 4): the 5' splice site recognition sequence (ACUUACCUG, from position 3 to 11) shown to be involved in base pairing interaction with the 5' splice site of pre-mRNA (33) and the Sm antigen-binding site (AAUUUGUUG, at positions 122 to 130) shown to bind to the core proteins (D to G) of U-snRNPs in vertebrates [cf. (2)], are invariable in all putative transcripts of tomato U1.1 to U1.8. The deviation $G_{129} \rightarrow U_{129}$ in the Sm antigen-binding site from the U1 RNA consensus (Fig. 2) was found also in pea (6) and broad bean [Kiss and Solymosy unpublished results, cited in (26)] U1 RNAs partially sequenced at the RNA level. The evolutionarily highly conserved sequences in Loops B, C and D, thought to be involved in base pairing interaction with U2 RNA (32), are also identical, with the exception of a mismatch in Loop C of U1.2 ($U_{104} \rightarrow C_{104}$) and of U1.5 ($U_{104} \rightarrow A_{104}$) as well as of one in Loop D of U1.7 ($U_{147} \rightarrow A_{147}$). In

addition, Stems III and IV are invariable, too, in their helical structure, with the exception of the latter in U1.3 ($G_{136} \rightarrow A_{136}$). The universal structural characteristic of all U1 RNAs sequenced so far, the "closure structure" (34) termed also "long range" (29) is also invariable in all tomato U1 variants exept U1.5, suggesting that U1.5 might be a pseudogene. Nucleotide sequence variation among the putative tomato U1 RNA transcripts appears mainly in Stem I, Loop A and Stem II of their secondary structure (Fig. 4). Stem I, Loop A, and part of Stem II were shown by antibody-mediated nuclease protection to be associated with one (70K) of the three (C, A and 70K) U1 snRNP-specific polypeptides in human U1 snRNP (35) and studies on the structure of U1 snRNP by in vitro assembly experiments coupled with immunoprecipitation and a protein sequestering assay indicated that the 70K and A proteins interact strongly with Stem I and Loop A and weakly with Stem II-Loop B (36). Hamm et al. (37) have recently suggested, by analyzing the assembly of U1 RNAs from common bean, soybean and a set of mutant Xenopus U1 RNAs into U1 snRNP in Xenopus egg extracts that protein C apparently recognizes a specific nucleotide sequence rather than a secondary structural element in Loop A of U1 RNA. Sequence variation in Loop A of the putative transcript of U1.1 to U1.8, together with the sequences of Loop A of the U1 RNAs that do and do not assemble into U1 snRNPs, respectively, is presented in Fig. 8. It can be seen that there is ample variation in the primary structure of Loop A in U1.1 to U1.8. So far only double point mutants (bean U1 RNA and delta A3 of Xenopus U1 RNA) have been available (37) for studying the nucleotide sequence-specificity of Xenopus U1 snRNP assembly. Inclusion of T7 or SP6 RNA polymerase transcripts of the coding regions of tomato U1.1 to U1.8 into protein-binding assays using U1 RNP-specific proteins of plant origin (38) would certainly enlarge the scope of our knowledge about the role of Loop A of U1 RNA in the assembly of plant U1 snRNP. Major differences in the secondary structures of Stem II of the putative transcripts of genes U1.1 to U1.8 (Fig. 4) may act as modulating factors in the binding of U1 snRNP-specific proteins to U1 RNA. Stem II of U1 RNA is assumed to be involved in interaction with the above proteins (35, 36).

In conclusion, we have no a priori reason to reject the idea that at least some of the tomato U1 RNA gene candidates we analyzed in this paper are transcribed in the nucleus. Although by definition (full length coding region with scattered substitutions, extensive flanking homology with U1 gene family) they should belong to Class I pseudogenes [a terminology introduced by Denison and Weiner (11) for human U1 RNA genes], recent data

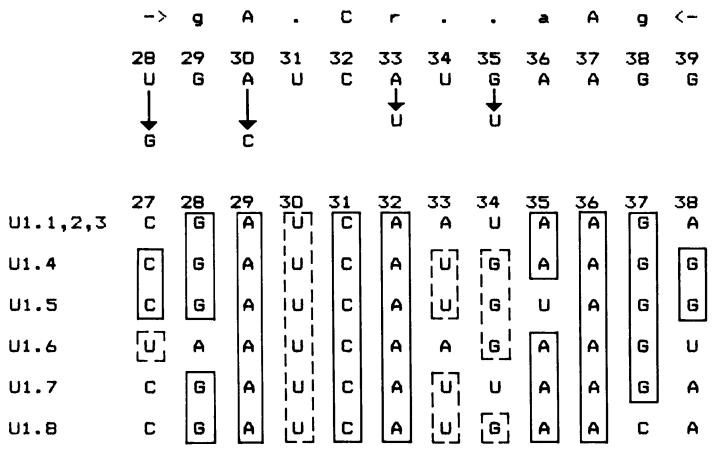


Fig. 8. Sequence variation in Loop A of U1 RNAs. In the top line the consensus sequence for this region of U1 RNA, as reported by Guthrie and Patterson (29), is shown. The inverted arrows mark positions which contain complementary nucleotides capable of base pairing. The second line represents the sequence of Loop A of Xenopus U1 RNA which is able to assemble into U1 snRNP (37).The arrows pointing downwards identify nucleotide replacements in bean (short arrow) and Xenopus (long arrow) double-mutant U1 RNAs which are unable to assemble into U1 snRNP (37). The sequences of the putative tomato U1 RNA variants, as indicated, are tabulated further below: residues which conform to the consensus are boxed by heavy lines and those which are variable in the consensus, but invariant in comparison with the Xenopus Loop A sequence, are delineated by dotted lines. The deviation of the numbering of residues in the Xenopus sequence from that in the tomato sequence is due to the inclusion of the m_3G cap in the former (37).

obtained by characterizing some Class I pseudogenes tend to attenuate the borderline between Class I pseudogenes and bona fide genes (5, 39, 40) and our tomato U1.1 gene candidate was faithfully transcribed in a heterologous system. Although the possibility cannot be ruled out that some of our tomato U1 RNA gene candidates are not expressed, or, even if they are, their transcripts are not stable (e.g. U1.5) and, hence, are not functional, it is tempting to speculate that most U-snRNAs possess a number of sequence variants which are encoded by "Class I pseudogenes". Such sequence variants might have a decisive role in differential splicing of unique pre-mRNAs, but would be difficult to detect at the RNA level because of their minute amounts in the nucleus. The occurrence of U-snRNA sequence variants at reduced concentrations is expected to be case in all cells of an organism if they participate in the splicing of a minor mRNA population, and in certain cells of the organism if their expression is tissue-specific or under

developmental control [for well-documented examples see (41, 42, 43) and for a recent review see (44)]. Differentiating plant cells, owing to their totipotence, may serve as an excellent experimental object for studying developmentally controlled expression of U-snRNA genes and the role of sequence variants of U-snRNAs in the post-transcriptional regulation of gene expression.

ACKNOWLEDGEMENTS

We thank Dr. R.W. Breidenbach for kindly supplying us with a tomato genomic library as well as Professor R. Lührmann and Dr. P. Bringmann for their kind gift of anti-$m_3^{2,2,7}$G IgG. Thanks are due also to Ms. G. Kaliczka for drawing the figures, Mr. B. Dusha for preparing the photographs and Ms. M. Doktor for typing the manuscript. This work was supported by Grants OTKA 564/1986, OKKFT(Tt)/1986, OTKA 174/1988 to F.S. from the Hungarian Academy of Sciences and by a grant to F.S. from the Hungarian Ministry of Industry.

*To whom correspondence should be addressed

+Present address: Friedrich Miescher-Institut, PO Box 2543, CH-4002, Basel, Switzerland

§Permanent address: Martin Luther Universität Halle-Wittenberg, Sektion Biowissenschaften, Pflanzenbiochemische Abteilung, 4020 Halle (Saale), Neuwerk 1, GDR

REFERENCES

(1) Reddy, R. and Busch, H. (1988) in Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles, Birnstiel, M.L. Ed., Springer Verlag, Berlin, pp. 1-37.
(2) Lührmann, R. (1988) in Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles, Birnstiel, M.L. Ed., Springer Verlag, Berlin, pp. 71-99.
(3) Steitz, J.A., Black, D.L., Gerke, V., Parker, K.A., Krämer, A., Frendewey, D. and Keller, W. (1988) in Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles, Birnstiel, M.L. Ed., Springer Verlag, Berlin, pp. 115-154.
(4) Breitbart, R.E. Andreadis, A. and Nadal-Ginard, B. (1987) Annu. Rev. Biochem. 56, 467-495.
(5) Patton, J.G. and Wieben, E.D. (1986) J. Cell Biol. 103, A 184.
(6) Krol, A., Ebel, J.P., Rinke, J. and Lührmann, R. (1983) Nucleic Acids Res. 11, 8583-8594.
(7) Reddy, R. (1988) Nucleic Acids Res. 16, r71-r85.
(8) Kiss, T., Jakab, G., Antal, M., Pálfi, Z., Hegyi, H., Kis, M. and Solymosy, F. (1988) Nucleic Acids Res. 16, 5407-5426.
(9) Lund, E. (1988) Nucleic Acids Res. 16, 5813-5826.
(10) Dahlberg, J.E. and Lund, E. (1988) in Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles, Birnstiel, M.L. Ed., Springer Verlag, Berlin, pp. 38-70.
(11) Denison, R.A. and Weiner, A.M. (1982) Mol. Cell. Biol. 2, 815-828.
(12) Kiss, T., Tóth, M. and Solymosy, F. (1985) Eur. J. Biochem. 152, 259-266.

(13) England, T.E., Bruce, A.G. and Uhlenbeck, O.C. (1980) Methods in Enzymology 65, 65-74.
(14) Benton, W.D. and Davis, R.W. (1977) Science 196, 180-182.
(15) Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor, New York.
(16) Helms, C., Graham, M.Y., Dutchik, J.E. and Olson, M.V. (1985) DNA 4, 39-49.
(17) Russel, M., Kidd, S. and Kelley, M.R. (1986) Gene 45, 333-338.
(18) Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
(19) Henikoff, S. (1987) Methods in Enzymology 155, 156-165.
(20) Murashige, T. and Skoog, F. (1962) Physiol. Plantarum,15, 473-497.
(21) Negrutiu, I., Shillito, R.D., Potrykus, I., Biasini, G. and Sala, F. (1987) Plant Mol. Biol. 8, 363-373.
(22) Vankan, P., Edoh, D. and Filipowicz, W. (1988) Nucleic Acids Res. 16, 10425-10440.
(23) Goodall, G., Wiebauer, K. and Filipowicz, W. (1989) Methods in Enzymol., in press.
(24) Melton, D.A., Krieg, P.A., Rebagliatti, M.R., Maniatis, T., Zinn, K. and Green, M.R. (1984) Nucleic Acids Res. 12, 7035-7056.
(25) Kiss, T., Abel, S. and Solymosy, F. (1989) Plant Mol. Biol. 12, 709-712.
(26) van Santen, V.L. and Spritz, R.A. (1987) Proc. Natl. Acad. Sci. USA 84, 9094-9098.
(27) van Santen, V.L., Swain, W. and Spritz, R.A. (1988) Nucleic Acids Res. 16, 4176.
(28) Vankan, P. and Filipowicz, W. (1988) EMBO J. 7, 791-799.
(29) Guthrie, C. and Patterson, B. (1988) Annu. Rev. Genet. 22, 387-419.
(30) Branlant, C., Krol, A., Ebel, J.P., Gallinaro, H., Lazar, E. and Jacob, M. (1981) Nucleic Acids Res. 9, 841-858.
(31) Mount, S.M. and Steitz, J.A. (1981) Nucleic Acids Res. 9, 6351-6368.
(32) Kiss, T., Antal, M., Hegyi, H. and Solymosy, F. (1988) Nucleic Acids Res. 16, 2734.
(33) Zhuang, Y. and Weiner, A.M. (1986) Cell 46, 827-835.
(34) Kretzner, L., Rymond, B.C. and Rosbash, M. ((1987) Cell 50, 593-602.
(35) Patton, J.R. and Pederson, T. (1988) Proc. Natl. Acad. Sci. USA 85, 747-751.
(36) Hamm, J., Kazmaier, M. and Mattaj, I.W. (1987) EMBO J. 6, 3479-3485.
(37) Hamm, J., van Santen, V.L., Spritz, R.A. and Mattaj, I.W. (1988) Mol. Cell. Biol. 8, 4787-4791.
(38) Pálfi, Z., Bach, M., Solymosy, F. and Lührmann, R. (1989) Nucleic Acids Res. 17, 1445-1458.
(39) Hoffman, M.L., Korf, G.M., McNamara, K.J. and Stumph, W.E. (1986) Mol. Cell. Biol. 6, 3910-3919.
(40) Hammarström, K., Bark, C., Westin, G., Zabielski, J. and Pettersson, U. (1986) Chemica Scripta, 12B, 165-170.
(41) Forbes, D.J., Kirschner, M.W., Caput, D., Dahlberg, J.E. and Lund, E. (1984) Cell 38, 681-689.
(42) Lund, E., Kahan, B. and Dahlberg, J.E. (1985) Science 229, 1271-1274.
(43) Lund, E. and Dahlberg, J.E. (1987) Genes Develop. 1, 39-46.
(44) Mattaj, I.W. and Hamm, J. (1989) Development 105, 183-189.