Volume 236, number 1, 5-8 FEB 06146 August 1988

Hypothesis

A potential role for U1 RNA genes in gene duplication and conversion events

David S. Latchman

Department of Biology, Medawar Building, University College London, Gower Street, London WC1E 6BT, England

Received 18 May 1988

A gene encoding U1 snRNA has been identified in *Caenorhabditis elegans* by homology to the human U1 gene. The gene lies at the boundary of a duplication event also involving the small heat shock protein genes. The possible role of the U1 sequence in mediating the duplication event is discussed.

U1 RNA gene; Gene duplication

The U1 small nuclear RNA is a 164 base molecule (reviewed in [1]) which is found as part of a ribonucleoprotein complex that plays an essential role in the splicing of RNA precursors to yield mature messenger RNA [2,3]. In keeping with its functional importance U1 RNA has been identified in a range of organisms including both vertebrates such as human, rodents, *Xenopus* and chicken [4] as well as invertebrates such as *Drosophila* [5], plants [6] and yeast [7]. In all organisms studied, the U1 RNA is encoded by a multi-gene family in which the number of genes ranges from less than ten in *Drosophila* and the chicken to several thousand in *Xenopus* [8]. In humans where the situation has been intensively studied, the number of functional U1 genes is estimated at between 50 and 125 per haploid genome, these being outnumbered by approximately ten times as many pseudogenes containing U1 related sequences but unable to encode a functional RNA due to small deletions or base substitutions [9]. Such pseudogenes appear to have been generated both by DNA-mediated events resulting in the duplication of a U1 gene and a large flanking region and by RNA-mediated events resulting in duplication of the U1 coding region alone [9,10].

In an attempt to compare the U1 coding sequences in a number of species, I used a functional human U1 gene sequence [8] to search the Genbank data base. In addition to the U1 gene sequences identified previously (see for example [4,5]) this search also detected a homology of the U1 RNA to a region of DNA located adjacent to the genes encoding the small heat shock proteins of the nematode *Caenorhabditis elegans* [11] which had not previously been reported to contain a U1-like DNA sequence. Inspection of this homology (fig.1) reveals that it extends throughout the U1 coding region without extending into flanking DNA and that within the region the two sequences are 70.6% homologous. This compares well with the 73.5% homology between the functional human U1 sequence and that of *Drosophila* U1 RNA [5], the only non-vertebrate animal U1 RNA which has been completely sequenced. Most interestingly comparison of the three sequences reveals that they show highly conserved regions and that in particular bases 3-13 of the U1 se-

Correspondence address: D.S. Latchman, Department of Biology, Medawar Building, University College London, Gower Street, London WC1E 6BT, England

Published by Elsevier Science Publishers B.V. (Biomedical Division)

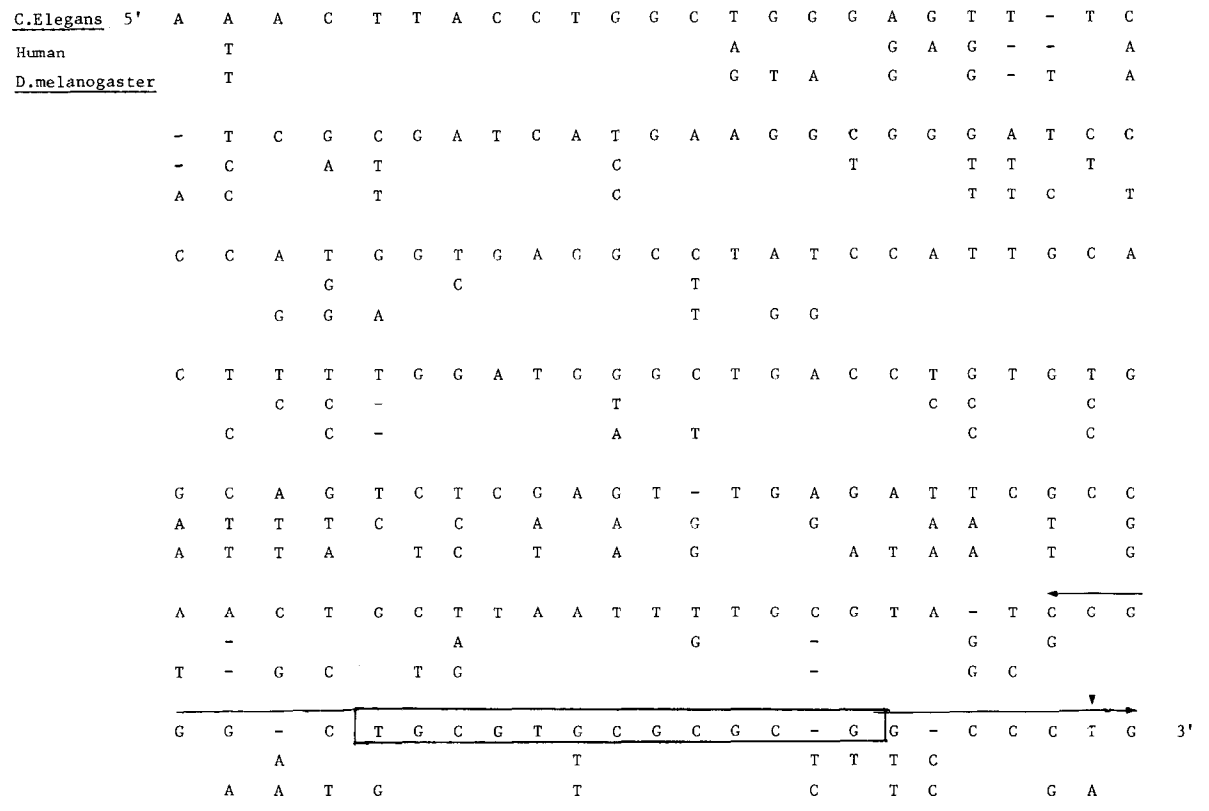

Fig.1. Comparison of the nematode U1 sequence and the sequence of functional U1 RNA genes of human and *Drosophila*. A blank indicates that the sequence is the same as that of the nematode, a dash where no base is present. The sequence in the nematode which could potentially form Z-DNA is boxed and the inverted repeats which flank it are indicated by the arrows. The single base mismatch (C for T) between the two copies of the 1.9 kb duplicated region of nematode DNA is contained within the U1 sequence and is indicated by the arrowhead.

quence are perfectly conserved in all three organisms. These bases are believed to pair with the 5′-splice acceptor site and hence play an essential function in the splicing function of U1 RNA [4]. Similarly the sequence contains at the appropriate position near the 3′-end the bases ATTTTG which conform to the consensus sequence $AT_{(4-6)}G$ required for binding of the Sm components of the spliceosome [12]. The conservation of these functionally important sequences suggests that the sequence identified may represent a functional gene encoding nematode U1. Alternatively the sequence may represent a pseudogene for U1 RNA which has arisen only recently and has not yet diverged significantly from the functional U1 sequence. Inspection of the region upstream of the U1 coding sequence reveals several potential transcriptional control sequences expected in a functional U1 RNA gene. Thus the sequence CATGTAATT which is very similar to the consensus enhancer sequence derived from several *Xenopus laevis* U1 RNA genes (Pyr ATG Pyr AAAT [13]) is found at the appropriate position (−235 base pairs) upstream of the gene within the 3′-untranslated region of the hsp 16-1 RNA whilst the sequence TCTGCATTCG which is similar to the *Xenopus* consensus sequence for 5′-end formation of the U1 RNA (TCTCCNTATG [13]) is found at the expected position 60 bases upstream from the start site. The short length of these sequences and their divergence in different U1 genes [13] do not however, permit their unequivocal identification as control elements in this case and studies involving the introduction of this region of DNA into *Xenopus* oocytes (see for example [14]) will be necessary to test whether it can direct the

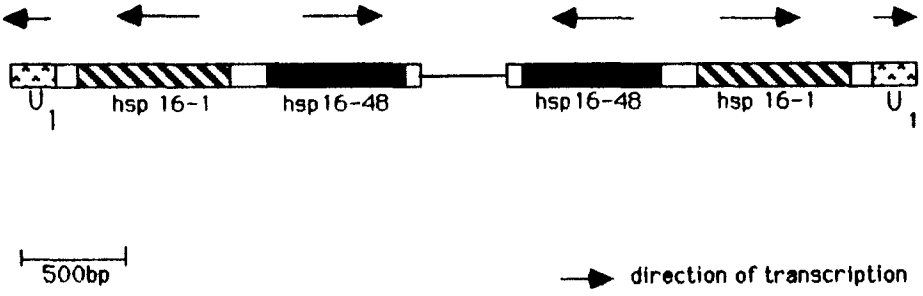

Fig.2. Location and orientation of U1 and hsp genes within the duplicated region of *C. elegans* DNA. The duplicated region is boxed, with the unique region separating the two halves of the duplication indicated as a single solid line.

synthesis of U1 RNA and conclusively prove that it represents a functional U1 gene. Whatever the case, the identification of a nematode sequence highly homologous to human U1 RNA provides further evidence for the evolutionary conservation and functional importance of this molecule.

The U1 RNA sequence identified in the nematode lies in a region of DNA with an unusual structure (fig.2). As described by Russnack and Candido [11] this region consists of a perfect inverted duplication (with only one base mismatch) of a 1.9 kb unit containing the genes for two small heat shock proteins. The two portions of the duplicated region are separated by 416 bp of unique DNA.

The U1 RNA sequence lies at the extreme end of the duplicated region, the sequence homologous to human U1 RNA being immediately followed by the short sequence AAAATA which terminates the duplication. The region of U1 DNA immediately adjacent to this sequence (boxed in fig.1) is of particular interest in that as pointed out by Russnack and Candido [11] it contains a 12 bp region of alternating pyrimidine and purine residues capable of forming Z-DNA, flanked by a six base pair inverted repeat structure. This unusual sequence located at the end of the duplication together with the evidence of U1 gene duplications in other organisms [9,10] suggests the possibility that the U1 sequence may have played some role in the duplication event. In this regard it is of interest that a study of four different human U1 DNA sequences [10] showed that they had virtually identical sequences up to at least 2.5 kb upstream of the gene but diverged dramatically in sequence only 50 bases downstream of the coding sequence, suggesting that DNA-mediated gene duplication events involving U1 duplicate upstream but not downstream sequence. This situation clearly parallels exactly that described here with the upstream sequences in the nematode case containing two functional genes for the heat shock proteins. It is unclear however, whether the U1 DNA acts as an initiator of the duplication event or serves as the terminus of an event initiated by other sequences. In either case, the Z-DNA sequence at the end of nematode U1 may function as suggested by Russnack and Candido [11] as an initiator of gene conversion events which maintain sequence identity between the two copies of the duplication, a similar role for Z-DNA elements having been proposed in other situations [15,16].

Whatever the precise role of the U1 DNA in the complex events occurring in this region of *C. elegans* DNA, the location and nature of the U1 sequences provide strong evidence that in addition to the role of its RNA product in splicing, the U1 DNA plays an important role in the fluidity of the eukaryotic genome.

REFERENCES

[1] Busch, H., Reddy, R., Rothblum, L. and Choi, Y.C. (1982) Annu. Rev. Biochem. 51, 617–654.
[2] Lerner, M.R., Boyle, J.A., Mount, S.M., Wolin, S.L. and Steitz, J.A. (1980) Nature 283, 220–224.
[3] Yang, V.W., Lerner, M.R., Steitz, J.A. and Flint, S.J. (1981) Proc. Natl. Acad. Sci. USA 78, 1371–1375.
[4] Branlant, C., Krol, A., Abel, J.P., Lazar, E., Gallinaro, H., Jacob, M., Sri-Widada, J. and Jeanteur, P. (1980) Nucleic Acids Res. 8, 4143–4154.
[5] Mount, S.M. and Steitz, J.A. (1981) Nucleic Acids Res. 9, 6351–6368.

[6] Van Santen, V.L. and Spritz, R.A. (1987) Proc. Natl. Acad. Sci. USA 84, 9094-9098.
[7] Kretzner, L., Rymond, B.C. and Rosbash, M. (1987) Cell 50, 593-602.
[8] Lund, E. and Dahlberg, J.E. (1984) J. Biol. Chem. 259, 2013-2021.
[9] Denison, R.A. and Weiner, A. (1982) Mol. Cell. Biol. 2, 815-828.
[10] Manser, T.F. and Gesteland, R.F. (1982) Cell 29, 257-264.
[11] Russnack, R.H. and Candido, P.M. (1985) Mol. Cell. Biol. 5, 1268-1278.
[12] Branlant, C., Krol, A., Ebel, J.-P., Gallinaro, H., Lazar, E. and Jacob, M. (1981) Nucleic Acids Res. 9, 841-858.
[13] Krol, A., Lund, E. and Dahlberg, J.E. (1985) EMBO J. 4, 1529-1535.
[14] Murphy, J.T., Burgess, R.R., Dahlberg, J.E. and Lund, E. (1982) Cell 29, 265-274.
[15] Flanagan, J.G., LeFranc, M.P. and Rabbits, T.H. (1984) Cell 36, 681-688.
[16] Kilpatrick, M.W., Klysik, J., Singleton, C.K., Zarling, D.A., Jovin, T.M., Hanau, L.H., Erlanger, B.F. and Wells, R.D. (1984) J. Biol. Chem. 259, 7268-7274.