# A pseudogene for human U4 RNA with a remarkable structure

**Karin Hammarström, Gunnar Westin, and Ulf Pettersson***

Department of Medical Genetics, University of Uppsala, The Biomedical Center, Box 589, S-751 23 Uppsala, Sweden

*Communicated by U.Pettersson*
*Received on 21 June 1982*

**The human DNA library of Lawn *et al.* (1978) was screened for sequences complementary to the small nuclear (sn) RNA U4. Several positive clones were identified by screening 100 000 recombinants, indicating that U4 sequences like other snRNA sequences are dispersed in the human genome. One recombinant was characterized in detail by subcloning a *Bgl*II fragment 1.9 kilobases (kb) long in the pBR322 plasmid. The subcloned fragment was partially sequenced and the results revealed a pseudogene for U4 RNA. The pseudogene was found to have a remarkable structure; it contains a sequence that, except in two positions, matches the first 68 nucleotides of the human U4 RNA sequence and the pseudogene is, moreover, flanked by perfect direct repeats 20 bp long. The results support the model of van Arsdell *et al.* (1981) suggesting that certain snRNA pseudogenes arise by reverse transcription of the RNA followed by integration of the cDNA copy at a new chromosomal locus.**

*Key words:* snRNA/U4 RNA/pseudogene/direct repeat/DNA sequence/mobile genetic element

## Introduction

Structural information concerning the mammalian genome is currently being collected at a remarkable pace. Dozens of human genes have now been isolated by molecular cloning and their structure is in many cases precisely known at the molecular level. One of the most unexpected findings that has come out of these studies is that mammalian genomes, in addition to *bona fide* genes, often contain many related copies of the gene, so-called pseudogenes. Pseudogenes are in most cases believed not to be functional since they contain mutations that render them inactive. The way in which the pseudogenes arise is still a matter of speculation. The expected mechanism would be gene duplication followed by mutational changes due to lack of selection pressure. The properties of many pseudogenes suggest, however, that other mechanisms might be involved in their creation. It has, for instance, been observed that some pseudogenes unlike their functional counterparts lack intervening sequences (Nishioka *et al.*, 1980; Vanin *et al.*, 1980; Hollis *et al.*, 1982; Wilde *et al.*, 1982). One possible mechanism that could generate such pseudogenes is reverse transcription followed by integration of the cDNA copy. We and others (Westin *et al.*, 1981; Monstein *et al.*, 1982; Denison *et al.*, 1981; van Arsdell *et al.*, 1981; Manser and Gesteland, 1981; Hayashi, 1981; Ohshima *et al.*, 1981) have recently initiated a study of genes for mammalian snRNA. The results show that sequences complementary to the snRNAs U1, U2, U3, and U6 are present at many different chromosomal sites. Sequence studies have, moreover, revealed that most copies represent pseudogenes since they carry several mutational changes (Denison *et al.*, 1981; van Arsdell *et al.*, 1981; Hayashi, 1981; Ohshima *et al.*, 1981; Manser and Gesteland, 1981; Westin *et al.*, 1981). An analysis of the sequence flanking U2 pseudogenes (Westin *et al.*, 1981) suggests, moreover, that there is little DNA sequence homology outside the pseudogene itself which supports the hypothesis that they have arisen as the result of transposition rather than by duplication of large chromosomal segments. Van Arsdell *et al.* (1981) have recently reported that U1, U2, and U3 pseudogenes sometimes are surrounded by direct sequence repeats, and on the basis of their findings they proposed that the pseudogenes represent integrated cDNA copies of the corresponding snRNA. In the present communication we describe the structure of a pseudogene for U4 RNA which gives additional support for the cDNA integration model, originally proposed by van Arsdell *et al.* (1981).

## Results

### *Isolation of recombinants containing U4 sequences*

The human DNA library of Lawn *et al.* (1978) was screened according to Benton and Davis (1977), using $^{32}$P-labelled U4 RNA from HeLa cells as a probe. Several plaques giving positive hybridization were encountered while screening 100 000 recombinants. Three of the positive recombinants which gave particularly strong hybridization signals were plaque purified and grown in large scale for DNA extraction. DNA from these three recombinants, designated U4/1, U4/4, and U4/5, was cleaved with different sets of restriction endonucleases and the resulting fragments were hybridized with $^{32}$P-labeled U4 RNA after separation and transfer to a nitrocellulose membrane. The cleavage patterns and the hybridization results showed that the three recombinants were different and thus must represent distinct chromosomal loci (data not shown). Cleavage of the U4/5 recombinant with *Bgl*II and subsequent hybridization revealed a fragment ~1.9 kb long containing U4-related sequences. This fragment was subcloned in the *Bam*HI cleavage site of the pBR322 plasmid, and one subclone designated pU4/5 was identified by colony hybridization according to Grunstein and Hogness (1975). This clone was grown in large scale and plasmid DNA was extracted for further characterization.

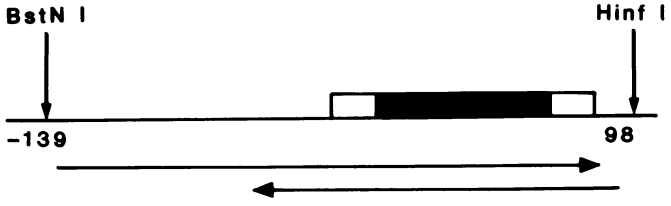

**Fig. 1.** A diagram that illustrates our sequencing strategy. A 1.9-kb *Bgl*II fragment was used for sequencing after cloning in the *Bam*HI site of the pBR322 vector. The pseudogene is illustrated by the thick line and the flanking direct repeats by unfilled boxes.

*To whom reprint requests should be sent.

© IRL Press Limited, Oxford, England. 0261-4189/82/0106-0737$2.00/0.

```
U4/5 5'-GAGACTATCTCCAATAAATAAATGAATTAATTAATTAAAATAATTTTAAATAAAGAAGACAGCATTTTTAACCTGAATTGAGGAAAT


                                                              2        10        20        30
         TAAACAGATATCAGGAAGAAATGTAAGACGAAACAGAAAATGCCTGCAGTTTA GCTTTGCGCAGTGGCAGTATCGTAGCCAATGAG
U4A RNA                               ····              NpppAmXCψUUGCGCAGUGGCAGUAUCGUAGCCAAUGAG

          * 40      * 50       60        70        80        90       100       110       120
         GTCTATCCGAGGAGCGATTATTGCTAATTGA AAAAGAAAATGCCTGCAGTTTAGAAGATGGCTG
         GUUUAUCCGAGGCGCGAUUAUUGCUAAUUGAmAAACUUψUCCCAAψACCCCGCCGUGACGACUUGCAAUAUAGUCGGCAUUGGCAAU


               130       140
         UUUUGACAGUCUCUACGGAGACUGGOH-3'
```

Fig. 2. The established sequence. The human U4 sequence as described by Krol *et al.* (1981) is also shown. Two point mutations in the pseudogene are indicated by asterisks. The flanking direct repeats are underlined. The dots show that the direct repeats will include 25 nucleotides if one mis-match is taken into account.

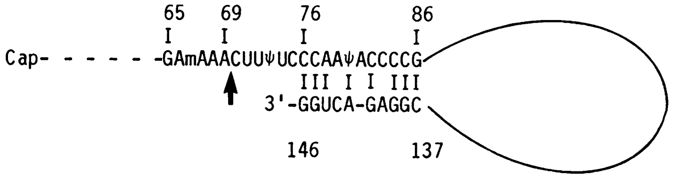

Fig. 3. A hypothetical model which illustrates how self-priming could occur due to intramolecular base pairing. The arrow indicates the 3′ end of the pseudogene in clone U4/5. This model is only valid for the U4A species since the other U4 variants have different 3′ termini.

### *Sequence analysis of subclone pU4/5*

A sequence 237 nucleotides long was established according to the strategy depicted in Figure 1 and the established sequence is shown in Figure 2. It is apparent from the results that the clone contains a severely truncated pseudogene, which is related to U4. The sequence of the pseudogene is colinear with the first 68 nucleotides of the U4 sequence but contains two mutations as compared with the U4 RNA sequence established by Krol *et al.* (1981). The truncated pseudogene is moreover flanked by perfect direct repeats 20 bp long. If one mis-matched base pair is taken into account, the repeats will include 25 nucleotides (Figure 2).

## Discussion

The mammalian genome appears to contain numerous pseudogenes for small nuclear RNAs and we describe, in the present communication, the first example of a pseudogene related to U4 RNA. The clone U4/5 represents a pseudogene by several criteria; it is severely truncated, only containing the first 68 nucleotides of the 142–146 nucleotides long U4 sequence. There are furthermore two point mutations, one of which is a transversion among the 68 nucleotides present in the pseudogene (Figure 2).

The most interesting finding in the present study is the observation that the truncated U4 pseudogene in clone U4/5 is flanked by direct repeats 20 bp long. The repeat at the 5′ side overlaps with the first two nucleotides in the truncated U4 sequence. This overlap could, however, be a coincidence if the repeat at the 3′ end of the pseudogene happens to be followed by the same two nucleotides that constitute the 5′ end of U4 RNA. The length of the actual repeats would in the latter case be reduced to 18 nucleotides.

The presence of flanking direct repeats raises the interesting question as to the mechanism by which pseudogenes are created. Previous studies indicate that snRNA pseudogenes are widely scattered in the mammalian genome (Denison *et al.*, 1981; Manser and Gesteland 1981; Westin *et al.*, 1981). Studies of the flanking sequences indicate, however, that they are unlikely to be the result of duplication events involving large chromosomal segments since sequences which precede and follow the pseudogenes in most cases share little sequence homology (Westin *et al.*, 1981; Ohshima *et al.*, 1981; Westin *et al.*, in preparation). Direct sequence repeats have previously been found to flank integrated proviruses (for a review, see Temin 1980) as well as mobile genetic elements and are often regarded as remnants of transposition events (for a review, see Calos and Miller, 1980). An interesting model for the generation of pseudogenes flanked by direct repeats was recently proposed by van Arsdell *et al.* (1981). The hypothesis postulates that the RNA sequence is first copied into a DNA sequence before becoming integrated at a new chromosomal site and requires, moreover, that integration occurs after staggered breaks have been introduced into the chromosomal DNA. After repair synthesis a pseudogene will be created, being flanked by direct repeats.

The results reported in the present communication support the model of van Arsdell *et al.* (1981) and a similar model has been proposed by Jagadeeswaran *et al.* (1981) for generation of interspersed repetitive DNA elements in eucaryotes. The model as originally proposed by van Arsdell *et al.* (1981) leaves, however, several questions yet to be answered. It would for instance require the existence of reverse transcriptase activity in germ line cells. Moreover, a priming event is necessary to generate a cDNA copy. Self-priming is a conceivable mechanism that could generate truncated cDNA copies of snRNA and the fact that the 5′ end is present in all pseudogenes so far studied, whereas 3′ sequences are often lacking, speaks in favour of a mechanism involving reverse transcription and self-priming. Based on the U4A RNA sequence, a structure can be drawn which, by self-priming, could give rise to a cDNA with approximately the same sequence as that present in the U4/5 clone (Figure 3). It should,

```
U4/5     AGAAAATGCCTGCAGTTTAG
           IIIIIII
U3.5      TAAAATGCTAATTATCCAA


U4/5     AGAAAATGCCTGCAGTTTAG
         IIIII
U1.101   AGAAACAGGCTTTTGC


U4/5     AGAAAATGCCTGCAGTTTAG
           IIII
U2.13      TAAATAATCAGGATGGAA
```

**Fig. 4.** A sequence comparison of the direct repeats that flank different snRNA pseudogenes. Data for the U1, U2, and U3 pseudogenes are from van Arsdell *et al.* (1981).

however, be pointed out that this structure is not the most energetically stable structure for U4 RNA. We are currently testing experimentally whether self-priming is a feasible mechanism for generation of truncated cDNA copies of U4 RNA.

Another interesting problem concerns the nature of the direct repeats. Are there specific endonucleases in mammalian cells which introduce staggered nicks in the DNA at preferred sites? A comparison between the direct repeats which have been found to flank different snRNA pseudogenes reveals similarities both with regard to length and primary sequence (Figure 4).

A difficult question, related to the model, is how the postulated events could take place in germ line cells. There is as yet no mechanism known which could explain how such events occur. There are, however, several examples reported in the literature which give strong support for reverse transcription being a mechanism by which genetic information can be reintroduced into the genome of mammalian cells. (Nishioka *et al.*, 1980; Vanin *et al.*, 1980; Hollis *et al.*, 1982; Wilde *et al.*, 1982).

## Materials and methods

### *Isolation of clones*

A library of human fetal liver fetal DNA was kindly supplied by T. Maniatis. Clones were identified by screening with $^{32}$P-labelled U4 RNA. DNA was prepared from the recombinant phages, grown and purified as described before (Westin *et al.*, 1981). Subcloning in the pBR322 plasmid was carried out as described by Stenlund *et al.* (1980).

### *DNA sequencing*

The protocol of Maxam and Gilbert (1980) was followed.

### *Hybridization*

Fragments were separated on 1% agarose gels before transfer to nitrocellulose according to the method of Southern (1975) and hybridization was carried out as described before (Westin *et al.*, 1981).

### *SnRNA*

SnRNA was extracted from $^{32}$P-labelled HeLa cells as described by Westin *et al.* (1981).

## Acknowledgements

We thank Ms. Marianne Gustafson for excellent secretarial aid. We also thank Hans-Jürg Monstein for $^{32}$P-labelled RNA preparations and Rolf Ohlson for carrying out some of the initial restriction enzyme cleavages. This investigation was supported by grants from the Swedish Medical Research Council, the Swedish Cancer Society, and the National Board for Technical Development.

## References

Benton,W.D., and Davis,R.W. (1977) *Science (Wash.)*, **196**, 180-182.

Calos,M.P., and Miller,J.H. (1980) *Cell*, **20**, 579-595.

Denison,R.A., van Arsdell,S.W., Bernstein,L.B., and Weiner,A.M. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 810-814.

Grunstein,M., and Hogness,D.S. (1975) *Proc. Natl. Acad. Sci. USA*, **72**, 3961-3965.

Hayashi,K. (1981) *Nucleic Acids Res.*, **9**, 3379-3388.

Hollis,G.F., Hieter,P.A., McBride,O.W., Swan,D., and Leder,P. (1982) *Nature*, **296**, 321-325.

Jagadeeswaran,P., Forget,B.G., and Weissman,S.M. (1981) *Cell*, **26**, 141-142.

Krol,A., Branlant,C., Lazar,E., Gallinaro,H., and Jacob,M. (1981) *Nucleic Acids Res.*, **9**, 2699-2716.

Lawn,R.M., Fritsch,E.F., Parker,R.C., Blake,G., and Maniatis,T. (1978) *Cell*, **15**, 1157-1174.

Manser,T., and Gesteland,R.F. (1981) *J. Mol. Appl. Genet.*, **1**, 117-125.

Maxam,A.M., and Gilbert,W. (1980) in Grossman,L., and Moldave,K. (eds.), *Methods in Enzymology*, **65**, Academic Press, NY, pp. 499-560.

Monstein,H.-J., Westin,G., Philipson,L., and Pettersson,U. (1982) *EMBO J.*, **1**, 133-137.

Nishioka,Y., Leder,A., and Leder,P. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 2806-2809.

Ohshima,Y., Okada,N., Tani,T., Itoh,Y., and Itoh,M. (1981) *Nucleic Acids Res.*, **9**, 5145-5158.

Southern,E.M. (1975) *J. Mol. Biol.*, **98**, 503-517.

Stenlund,A., Perricaudet,M., Tiollais,P., and Pettersson,U. (1980) *Gene*, **10**, 47-52.

Temin,H.M. (1980) *Cell*, **21**, 599-600.

van Arsdell,S.W., Denison,R.A., Bernstein,L.B., Weiner,A.M., Manser,T., and Gesteland,R.F. (1981) *Cell*, **26**, 11-17.

Vanin,E.F., Goldberg,G.I., Tucker,P.W., and Smithies,O. (1980) *Nature*, **286**, 222-226.

Westin,G., Monstein,H.-J., Zabielski,J., Philipson,L., and Pettersson,U. (1981) *Nucleic Acids Res.*, **9**, 6323-6338.

Wilde,C.D., Crowther,C.E., Cripe,T.P., Gwo-Shu Lee,M., and Cowan,N.J. (1982) *Nature*, **297**, 83-84.