Volume 12 Number 3 1984 Nucleic Acids Research

Pseudogenes for human U2 small nuclear RNA do not have a fixed site of 3′ truncation

Scott W.Van Arsdell* and Alan M.Weiner

Department of Molecular Biophysics and Biochemistry, Yale Medical School, 333 Cedar Street, P.O. Box 3333, New Haven, CT 06510, USA

Received 27 October 1983; Revised and Accepted 27 December 1983

ABSTRACT

We present the sequences of five additional human U2 pseudogenes which are very similar to the U2.13 pseudogene reported previously [Van Arsdell et al. (1981) Cell 26, 11-17]. All six U2 pseudogenes preserve the 5′ end of the mature U2 snRNA sequence, and all six are flanked by nearly perfect direct repeats that differ in sequence and range in length from 16 to 21 base pairs. The 3′ ends of the six U2 pseudogenes are truncated at five different sites between position 33 and 82, and in two cases the 3′ end of the pseudogene overlaps the downstream direct repeat by 5 or 6 base pairs. The structure of these six U2 pseudogenes contrasts with that of four human U3 pseudogenes [Bernstein et al. (1983) Cell 32, 461-472] all of which are identically truncated at position 69 or 70, and appear to be derived from a self-primed 74 base reverse transcript of U3 snRNA. Comparison of the U2 and U3 pseudogenes suggests a model for their generation in which the 3′ end of the pseudogene is always truncated relative to the initial cDNA template.

INTRODUCTION

Mammalian genomes are rich in sequences created by the reverse flow of genetic information from cellular RNA back into chromosomal DNA. Hollis et al. (1) introduced the term "processed gene" to describe pseudogenes that correspond to an integrated genomic copy of a spliced mRNA, and we shall interpret this useful term broadly to denote all nonviral genomic sequences that were generated by an RNA-mediated insertion event. In mammals, processed genes include many different sequences: pseudogenes for the small nuclear RNAs (snRNAs) U1, U2, U3, U4, and U6 (2-8), divergent genomic copies of the transposable middle repetitive Alu family sequence (8-10), and pseudogenes for proteins as diverse as the human immunoglobulin constant regions C_λ (1) and C_ε (12,13), human metallothionein II (14), rat α-tubulin (15), and human β-tubulin (16). We (4,5,8) and others (1,6, 9,10) have proposed that processed genes arise by reverse transcription of a cellular RNA species, followed by integration of the cDNA into new chromosomal sites in germline DNA. We do not know whether the reverse transcriptase activity is provided by a normal cellular DNA polymerase, an endogenous provirus, or transient retroviral infection of germline cells (for a discussion, see ref. 8);

however, the available DNA sequence data do suggest a general mechanism for integration of the reverse transcript once it is made (4,8).

MATERIALS AND METHODS

Isolation and initial characterization of recombinant lambda bacteriophage containing the U2 pseudogenes were described previously (17). Short restriction fragments spanning the pseudogenes were subcloned into the M13 vectors mp8 and mp9 (18) and sequenced by the technique of Sanger *et al.* (19) at least twice on one strand without ambiguity except in the case of U2.1, where the 5′ end of the pseudogene was more than 300 bp from the primer oligonucleotide.

RESULTS

The human U2 pseudogenes

In Fig. 1 we compare the DNA sequences of six U2 pseudogenes with the DNA sequence of the human U2.24A gene (20); the sequence of the U2.13 pseudogene was reported previously (4). Each U2 pseudogene contains a truncated 5′ fragment of the mature U2 snRNA sequence. In all six pseudogenes the homology with U2 RNA begins precisely at the 5′ end of the RNA, but homology with U2 RNA is lost at different downstream positions (nucleotides 33, 35, 39, 57, or 82) in the six pseudogenes. The variety of 3′ truncations between nucleotides 33 and 82 in the human U2 pseudogenes should be contrasted with the consistent 3′ truncation of four human U3 pseudogenes at nucleotide 69 or 70 in the U3 RNA sequence (8). Perfect (or nearly perfect) direct repeats flank the U2 homology in all six U2 pseudogenes. These repeats are generally 16 base pairs long, but they do vary from 16 base pairs (U2.1, U2.4, U2.5, and U2.8) to 18 (U2.13) or even 21 base pairs (U2.6). Such variability contrasts sharply with the defined length of the direct repeats made by such eucaryotic transposable elements such as Ty1 in yeast, copia and the P element in Drosophila, or the vertebrate retroviruses such as MMTV or RSV (for discussion, see refs. 21 and 22). As expected for insertion of snRNA information into random chromosomal sites, the direct repeats do not share any obvious consensus sequence. On the strand synonymous with the snRNA, the direct repeats flanking the six U2 and four U3 pseudogenes are very rich in adenine (50% to 60%). Thus both strand asymmetry and richness in A + T may influence the choice of chromosomal target sequences for insertion; local melting of the DNA cannot be the sole determinant.

For consistency, we have drawn the box in Fig. 1 to maximize the length

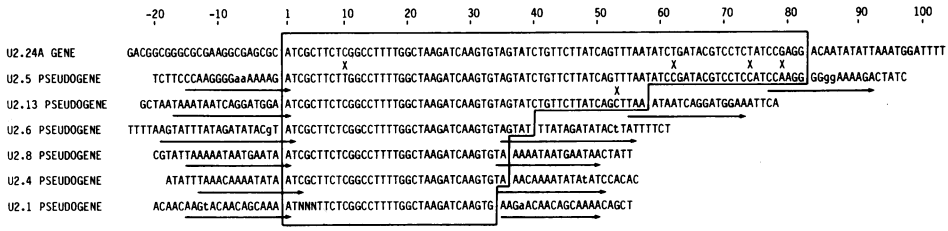

Fig. 1. The sequences of six human U2 pseudogenes are compared to a human U2 gene candidate (17). Homologies between the pseudogene and the gene are boxed; mismatches between the pseudogenes and the gene are indicated by the symbol "x"; flanking direct repeats are denoted by an arrow; imperfections within the direct repeats are represented by lower case letters; ambiguous nucleotides (see Materials and Methods) are denoted by an "N".

of the U2 pseudogenes, and the arrows have been drawn to maximize the length of the direct repeats. When this is done, the upstream direct repeat in each pseudogene overlaps the first nucleotide in the U2 sequence (an adenine); the downstream direct repeat does not overlap the U2 sequence at all in one pseudogene (U2.1) but can overlap by as much as 5 or 6 base pairs (U2.5 and U2.6). While an overlap of only 2 (U2.4 and U2.8) or 3 base pairs (U2.13) could be discounted as fortuitous, a 5 or 6 base pair overlap suggests that homology between the snRNA and the chromosomal target site can, but need not, play a role in determining the final configuration of the pseudogene.

DISCUSSION

Human U2 and U3 pseudogenes have similar structures.

We previously characterized the structure of four different U3 pseudogenes (8), each of which contains an identical 5′ fragment of the U3 snRNA sequence (nucleotide 1 to 69 or 70). Two of the U3 pseudogenes (U3.5 and U3.7) are flanked by perfect direct repeats of 18 or 19 base pairs; the other two pseudogenes (U3.2 and U3.6) have no direct repeats whatsoever. We also demonstrated that human U3 snRNA can serve *in vitro* as a self-priming template for avian myeloblastosis virus reverse transcriptase; the product of this reaction is a 74 base cDNA corresponding to the first 74 nucleotides of the U3 sequence. The ability of U3 snRNA to function as a self-priming template for reverse transcription might be unrelated to the formation of U3 pseudogenes, but we prefer the more optimistic interpretation that the covalent U3 snRNA-cDNA hybrid (or a U3 cDNA derived from it) is in fact the natural intermediate for insertion *in vivo*. We therefore proposed that the four characterized human U3 pseudogenes were created by direct insertion of the 74 base cDNA into

a new chromosomal site, with concurrent and consistent loss of 4 to 5 bases from the 5′ end of the cDNA.

The six U2 pseudogenes described above bear a striking resemblance to the four U3 pseudogenes. All ten pseudogenes preserve the 5′ end of the mature snRNA sequence, and whenever the pseudogene is flanked by direct repeats, the upstream direct repeat overlaps the 5′ end of the snRNA sequence by at least one nucleotide. The two major differences between the human U2 and U3 pseudogenes are that (i) all four characterized U3 pseudogenes are truncated identically at nucleotide 69 or 70, whereas the U2 pseudogenes are a truncated at different sites between nucleotides 33 and 82; and (ii) in U2 but not U3 pseudogenes, the downstream direct repeat occasionally overlaps the 3′ end of the snRNA homology. Because the U3 pseudogenes appear to have been generated by insertion of a U3 cDNA, we feel it is reasonable to assume that U2 pseudogenes with a similar structure were created by insertion of a comparable U2 cDNA; however, we have been unable to synthesize the predicted U2 cDNA in vitro although we have used both naked U2 snRNA and purified U2 snRNPs (small nuclear ribonucleoprotein particles) as the template for reverse transcriptase (L.B. Bernstein and A.M. Weiner, unpublished results).

Why do U2 pseudogenes have different sites of 3′ truncation?

The DNA sequence analysis of five additional U2 pseudogenes presented in this paper reveals two new characteristics of snRNA pseudogenes which must be explained by any model for the reverse flow of genetic information from cellular RNA back into the genome. First, the 3′ truncation of five U2 pseudogenes with respect to the sixth strongly resembles the 3′ truncation of four U3 pseudogenes relative to the U3 cDNA from which they appear to be derived (8). Second, unlike the U3 pseudogenes, the 3′ ends of several U2 pseudogenes exhibit significant overlap with the downstream direct repeat. As discussed below and shown in Fig. 2, steps 3 and 4, both characteristics of the U2 pseudogenes can be accommodated naturally by simple refinements of our earlier model for cDNA insertion (4).

We previously suggested that the 3′ end of a cDNA is likely to attack the 5′ end of a double-stranded chromosomal break, because in this way the 5′ sequence of the snRNA can be preserved and the 3′ end of the chromosomal break can prime synthesis of the second cDNA strand (Fig. 2, steps 2 and 3). Were the 5′ end of a cDNA to have been joined to a protruding 3′ end at the target site, there would have been no available 3′ end to prime second strand synthesis without loss of 3′ sequence from the cDNA. It was quite remarkable to us to realize that preservation of the 5′ end of the snRNA sequence in each

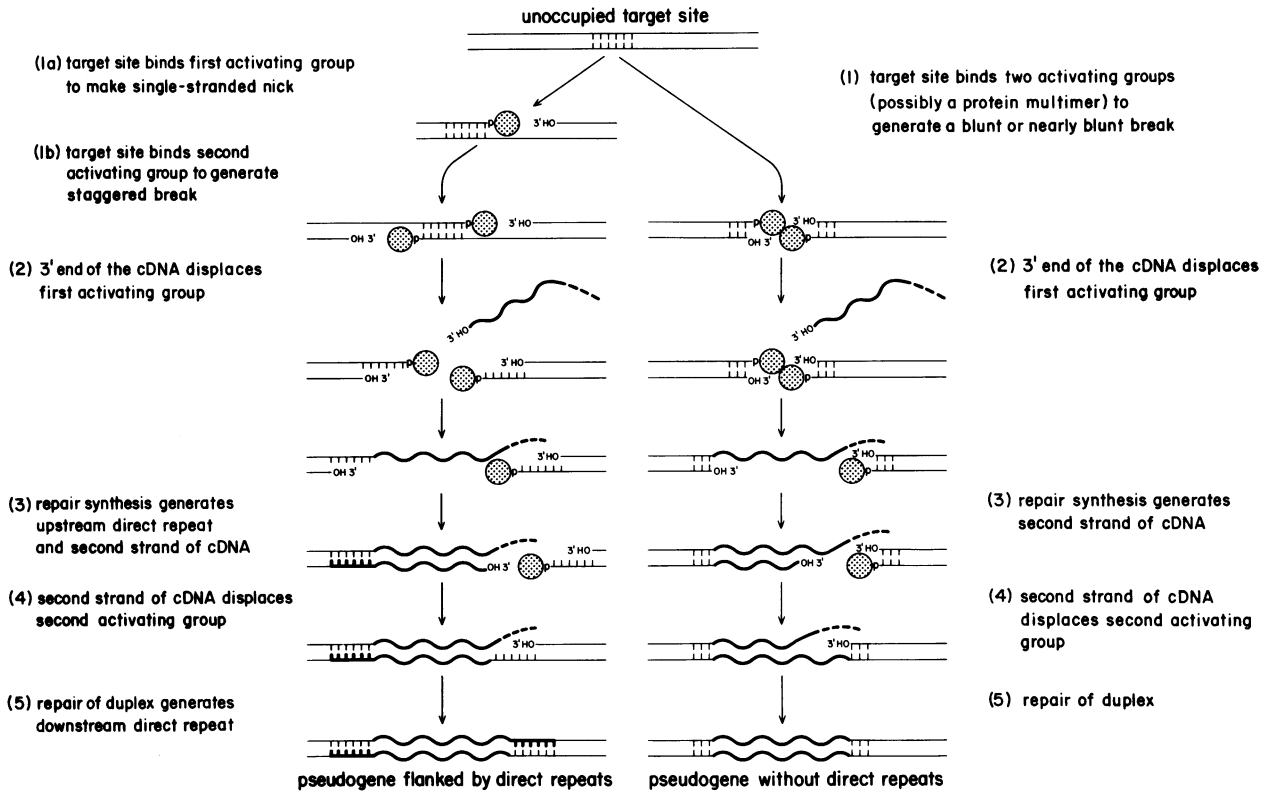

Fig. 2. A model for the generation of small nuclear RNA pseudogenes.

of the pseudogenes made such a strong prediction regarding the mechanism of insertion.

All polynucleotide joining reactions require a source of energy to make the new phosphodiester bond. In principle, either the 5′ end of the chromosomal break or the 3′ end of the cDNA could be activated for the strand transfer reaction shown in Fig. 2, step 2. We have deliberately drawn Fig. 2 with an activated chromosomal break (step 1) because there are examples of chromosomal activation by DNA topoisomerases (albeit with the opposite polarity to that invoked in Fig. 2; refs. 23-26) and ligases (for a review, see ref. 27) but no known precedent for activating the end of a nonviral linear extrachromosomal DNA single strand. For simplicity, we have drawn the upstream and downstream chromosomal activating groups (Fig. 2, stippled circles) in a symmetrical fashion, but we do not mean to imply that the same enzyme necessarily mediates the two separate strand transfer reactions (steps 2 and 4). Because two of the four U3 pseudogenes previously characterized lack direct repeats (8) while other snRNA pseudogenes are flanked by direct repeats of at least 14 base pairs (2-8), we have drawn Fig. 2 with two separate pathways for cDNA insertion, starting from a blunt or a staggered chromosomal break; however, the absence of direct repeats could equally well be attributed to recombination between two U3 pseudogenes or between a U3 pseudogene and a _bona fide_ U3 gene.

After covalent attachment of the 3′ end of the cDNA to a 5′ end at the chromosomal break, a cellular DNA polymerase begins synthesis of the second cDNA strand using the 3′ end at the break as a primer. In the case of a staggered break, this synthesis completes the upstream direct repeat. The ability of DNA polymerase to copy the complete cDNA will determine the length of the pseudogene; incomplete copying will result in 3′ truncation of the pseudogene relative to the original cDNA. Self-primed _in vitro_ reverse transcription of U3 snRNA produces a covalent U3 snRNA-cDNA hybrid (in Figs. 2 and 3 the cDNA is shown as a wavy solid line and the attached snRNA as a dotted line). Although the U3 cDNA is 74 bases long, the U3 pseudogenes are all truncated at position 69 or 70 in the RNA sequence (8). We believe that variable truncation at the 3′ end of U2 pseudogenes and the consistent loss of 4 to 5 bases from the 5′ end of the U3 cDNA during insertion both reflect factors that can influence the length of the second cDNA strand: (a) The DNA polymerase responsible for second strand synthesis may be partially or completely blocked by secondary structure within cDNA, secondary structure within the covalent cDNA-snRNA hybrid, the presence of an RNA:RNA duplex, or

the formation of base pairs between the first cDNA strand and the downstream direct repeat (see discussion below). (b) The cDNA or cDNA-snRNA hybrid may be trimmed by nucleases prior to or during the insertion process. For U3 pseudogenes the trimmed cDNA might be a consistent 69 or 70 nucleotides long, whereas for U2 pseudogenes with different degrees of truncation the trimmed cDNA could be quite heterogeneous. (c) The initial U2 cDNA may be primed at multiple sites corresponding to different extents of 3′ truncation in the six U2 pseudogenes; however, multiple priming sites for U2 reverse transcription would be at variance with the unique priming site found for U3 snRNA *in vitro* (8).

Does homology between the cDNA and the downstream direct repeat influence healing of the chromosomal break?

In our model, the double-stranded break is healed after insertion of the cDNA by a reaction analogous to the initial attack of the 3′ hydroxyl group of the *first* cDNA strand on an activated 5′ phosphate bond at the target site: the 3′ hydroxyl group of the *second* cDNA strand attacks the other activated 5′ phosphate (Fig. 2, step 4). In pseudogenes where the downstream direct repeat does not overlap the snRNA sequence (e.g., U2.1), the 3′ hydroxyl group of the second cDNA strand would attack the activated 5′ phosphate directly to heal the chromosomal break. To explain overlaps of as much as 5 or 6 base pairs between the 3′ end of the truncated snRNA sequence and the downstream direct repeat (U2.5 and U2.6), we propose that the first cDNA strand can search for homology in the exposed single-stranded DNA of the downstream direct repeat. The formation of base pairs between the first cDNA strand and the downstream direct repeat could block the progress of the DNA polymerase responsible for copying the template cDNA; synthesis of the second cDNA strand would then

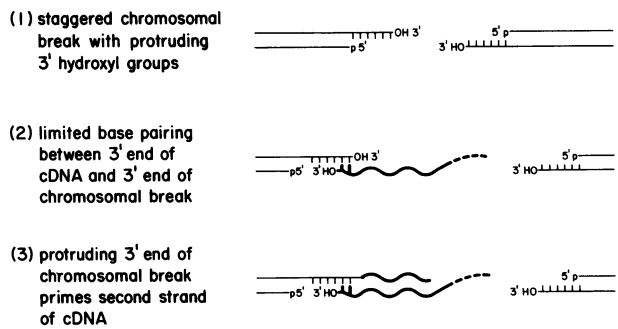

Fig. 3. An alternative model for the generation of snRNA pseudogenes.

proceed only as far as the sequence overlap, before the chromosomal break was healed by attack of the 3′ hydroxyl group of the second cDNA strand on the remaining activated 5′ phosphate group. In the final step of our model for cDNA insertion, DNA repair restores the uninterrupted DNA duplex; for a staggered break, this synthesis would complete the downstream direct repeat.

An alternative model for cDNA integration

In the complete absence of any data bearing directly on the nature of the initial chromosomal break, we must also consider models in which the staggered break has protruding 3′ ends (Fig. 3, step 1). In this case, the 5′ end of the pseudogenes could be preserved if the protruding 3′ end of the chromosomal break primed synthesis of the second cDNA strand by forming a limited number of base pairs with the 3′ end of the cDNA (Fig. 3, steps 2 and 3). We think this model is less plausible than that in Fig. 2 for two related reasons: (1) The model shown in Fig. 3 is difficult to reconcile with the observation that there is only a single base pair of overlap between the upstream direct repeat and the 5′ end of all the known snRNA pseudogenes except for U2.4 and U2.6; and (2) it is also difficult to imagine how a single base pair of overlap between the 3′ end of the cDNA and the protruding 3′ end of the staggered break would suffice for priming DNA synthesis.

Work on procaryotic transposable elements in many laboratories has demonstrated that DNA sequence analysis of random transpositions can never resolve enzymological or mechanistic questions; the mechanism which generates snRNA pseudogenes will likewise remain elusive until the frequency of these rare insertion events can be increased in vivo or until the entire process can be recreated in vitro.

ACKNOWLEDGEMENTS

We are grateful to Leroy Liu for a catalytic discussion, and to Nancy Maizels, Cathy Joyce and Nigel Grindley for a great deal of help on the manuscript. We thank the referee for encouraging us to take the model presented in Fig. 3 more seriously. This work was supported by Grant GM31073 from the National Institutes of Health and Grant PCM 7821799 from the National Science Foundation.

*Present address: Department of Microbiology and Immunology, University of California, Berkeley, CA 94720, USA

REFERENCES

1. Hollis, G.F., Hieter, P.A., McBride, O.W., Swan, D. and Leder, P. (1982) Nature 296, 321-325.
2. Hayashi, I. (1981) Nucl. Acids Res. 8, 3379-3388.
3. Oshima, Y., Okada, N., Tani, T., Itoh, Y. and Itoh, M. (1981) Nucl. Acids

Res. 9, 5145-5158.
4. Van Arsdell, S.W., Denison, R.A., Bernstein, L.B., Weiner, A.M., Manser, T. and Gesteland, R.F. (1981) Cell 26, 11-17.
5. Denison, R.A. and Weiner, A.M. (1982) Mol. Cell. Biol. 2, 815-828.
6. Hammarstrom, K., Westin, G. and Pettersson, U. (1982) EMBO Journal 1, 737-739.
7. Piechaczyk, M., Lelay-Taha, M., Sri-Wadada, J., Brunel, C., Liautard, J-P. and Jeanteur, P. (1982) Nucl. Acids Res. 10, 4627-4640.
8. Bernstein, L.B., Mount, S.M. and Weiner, A.M. (1983) Cell 32, 461-472.
9. Jagadeeswaran, P., Forget, B.G. and Weissman, S.M. (1981) Cell 26, 141-142.
10. Schmid, W. and Jelinek, W.R. (1982) Science 216, 1065-1070.
11. Grimaldi, G. and Singer, M.F. (1982) Proc. Natl. Acad. Sci. USA 79, 1497-1500.
12. Battey, J., Max, E.E., McBride, O.W., Swan, D. and Leder, P. (1982) Proc. Natl. Acad. Sci. USA 79, 5956-5960.
13. Ueda, S., Nakai, S., Nishida, Y., Hisajima, H. and Honjo, T. (1982) EMBO Journal 1, 1539-1544.
14. Karin, M. and Richards, R.I. (1982) Nature 299, 797-802.
15. Lemischka, I. and Sharp, P.A. (1982) Nature 300, 330-335.
16. Lee, M.G-S., Lewis, S.A., Wilde, C.D. and Cowan, N.J. (1983) Cell 33, 477-487.
17. Denison, R.A., Van Arsdell, S.W., Bernstein, L.B. and Weiner, A.M. (1981) Proc. Natl. Acad. Sci. USA 78, 810-814.
18. Messing, J. and Vierira, J. (1982) Gene 19, 269-276.
19. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
20. Van Arsdell, S.W. and Weiner, A.M. (1984) Mol. Cell. Biol. 4, in press.
21. Temin, H. (1980) Cell 21, 599-600.
22. O'Hare, K. and Rubin, G.M. (1983) Cell 34,
23. Been, M.D. and Champoux, J.J. (1981) Proc. Natl. Acad. Sci. USA 78, 2883-2887.
24. Gellert, M. (1981) Ann. Rev. Biochem. 50, 879-910.
25. Reed, R.R. and Grindley, N.D.F. (1981) Cell 25, 721-728.
26. Halligan, B.D., Davis, J.L., Edwards, K.A. and Liu, L.F. (1982) J. Biol. Chem. 257, 3995-4000.
27. Kornberg., A. (1980) DNA Replication (Freeman, San Francisco).