J Mol Evol (1988) 27:1–7

Journal of Molecular Evolution

A Repetitive DNA Sequence in the Salmonid Fishes Similar to a Retroviral Long Terminal Repeat

Robert D. Moir and Gordon H. Dixon

Department of Medical Biochemistry, Faculty of Medicine, Health Sciences Centre, University of Calgary, 3330 Hospital Drive N.W., Calgary, Alberta T2N 4N1, Canada

Summary. We describe the characteristics of a repetitive DNA sequence from the rainbow trout and related salmonid fishes that is similar to a retroviral long terminal repeat (LTR). The repeat is 160 bp long and contains a region of homology to the LTR of the avian sarcoma virus. Two clones with this repeat from the chum salmon also have a polypurine tract and tRNA binding site, respectively, and these clones may represent the two LTRs of a retrovirus or retroviral-like repetitive element. Copies of the repeat are also adjacent to rainbow trout and chum salmon protamine genes. These repeats may be "solo" LTRs. There appears to be some polymorphism in restriction sites between individual rainbow trout and considerable differences between salmonid fish species when the repeat is used as a probe.

Key words: Long terminal repeat — Protamine gene — Salmonid fishes

Introduction

A number of repeated DNA elements that are present in eukaryotes have structural similarities to retroviruses (Weiner et al. 1986). The similarities include the presence of direct long terminal repeats (LTRs) flanking an internal sequence that may contain one or more open reading frames. Immediately following the 5′ LTR there is a sequence complementary to a tRNA. In retroviruses the tRNA acts as a primer for viral DNA synthesis (Varmus 1982). A polypurine tract may also be present preceding the 3′ LTR and it may also be involved in retroviral DNA synthesis (Varmus 1982).

The transposable element, TY1, of yeast transposes through an RNA intermediate (Boeke et al. 1985) and encodes a protein with reverse transcriptase activity. Moreover, virus-like particles appear in the cytoplasm of cells in which transcription has been induced, although the particles are not infectious (Garfinkel et al. 1985). These observations suggest that TY1, and by inference the repeated DNA sequences that have structural similarities, may share a common evolutionary origin with retroviruses (Weiner et al. 1986).

In this report, we describe the structure of a repeated DNA sequence from the salmonid fishes that also has some similarities with a long terminal repeat. In two clones, sequences are present that are analogous to a complete retroviral element with two LTRs, while other clones contain "solo" LTRs (Rotman et al. 1984).

Materials and Methods

Construction and Screening of Libraries. The construction of the rainbow trout (*Salmo gairdneri*) Charon 4A library and the isolation of the protamine clones have been described previously (States et al. 1982; Aiken et al. 1983). The chum salmon (*Oncorhynchus keta*) genomic library was also made in Charon 4A using the 12–20-kb size fraction from a complete EcoRI digest. The library was screened using Gene Screen membrane circles according to the manufacturer's instructions (New England Nuclear). Purified bacteriophage were grown in liquid culture and purified as described using density gradient centrifugation (Maniatis et al. 1982). DNA was isolated by proteinase K sodium dodecyl sulfate (SDS) treatment (Maniatis et al. 1982).

Subcloning and DNA Sequencing. DNA fragments from purified lambda clones that hybridized to a particular probe were

Offprint requests to: G.H. Dixon

subcloned into pBR322. The plasmid DNA was isolated from liquid culture using the alkaline lysis method (Garger et al. 1983) followed by density gradient centrifugation (Maniatis et al. 1982).

A number of restriction enzyme sites were mapped in the subcloned inserts and used for end-labeling of DNA fragments with polynucleotide kinase (Maniatis et al. 1982). The Maxam–Gilbert method (Maxam and Gilbert 1980) or the rapid chemical methods described by Bencini et al. (1984) were used to determine the nucleotide sequence.

Southern Blotting and Hybridizations. Southern blotting was done essentially as described (Maniatis et al. 1982) except nylon membranes were used (Hybond N, Amersham). The DNA was fixed to the filter after transfer using long-wave ultraviolet light according to the manufacturer's instructions.

Hybridizations were done at 42°C in 50% formamide, 0.5% SDS, 6× standard saline citrate (SSC) (1× SSC = 0.015 M sodium citrate, 0.15 M sodium chloride), 5× Denhardt's solution (1× Denhardt's = 0.02% Ficoll 400, 0.02% polyvinylpyrrolidone, 0.02% bovine serum albumin) and 100 μg/ml sonicated calf thymus DNA. The filters were washed in 2× SSC, 0.1% SDS at 65°C. A high stringency wash of 0.1× SSC, 0.1% SDS at 65°C was included for the hybridization shown in Fig. 5. All probes were prepared using the random primer method of Feinberg and Vogelstein (1983).

Results

Identification of Clones Carrying the Repetitive Element

While screening a chum salmon (*Oncorhynchus keta*) genomic library for protamine genes, we discovered a repetitive element located adjacent to the rainbow trout (*Salmo gairdneri*) p101 protamine gene (States et al. 1982), which had been used as a probe to screen the library. The element was identified on the basis of clones that hybridized to a protamine gene probe with a substantial amount of 5′ and 3′ flanking DNA (the BglII to BamHI fragment in Fig. 1) but not to a probe confined mostly to the coding region (the AvaII to HpaII fragment in Fig. 1). Eleven of 15 clones that hybridized to the BglII–BamHI probe did not subsequently hybridize to the smaller probe. Four clones hybridized to both probes and contained chum salmon protamine genes. The detailed characterization of these protamine clones is presented in Moir and Dixon (1988). Two of the 11 clones, termed pC2 and pC3, which apparently contained homology to only the flanking regions of the p101 gene but not to the protamine coding region, were selected for sequencing. The regions in pC2 and pC3 that hybridized to the BglII–BamHI probe could be localized to 1.6- and 1.5-kb fragments, respectively. The entire nucleotide sequences of these fragments were determined (not shown). For most of the inserts only one strand was sequenced.

Homology between the Clones

The clones pC2, pC3, and the rainbow trout protamine clone, p101 (States et al. 1982), contain a common 160-bp sequence. The location of this sequence within p101 is indicated in Fig. 1 with thick underlining. The homology ends approximately 200 bp before the start of transcription of the protamine gene. The same sequence is present in exactly the same location in the chum salmon protamine clone, pC241. Fig. 2 is an alignment of the sequence common to all four clones, pC2, pC3, and two protamine gene clones p101 (rainbow trout), and pC241 (chum salmon). Clones pC2 and pC3 do not contain further homology to each other or to p101 or pC241. Clones p101 and pC241 are very similar for long distances 5′ and 3′ of the protamine gene (Moir and Dixon 1988). The homology between the clones begins with the sequence TGT, except in pC2 which has TAT (see Fig. 2, first boxed sequence). The end of homology varies slightly but the clones can be aligned so that the common end is ACA (Fig. 2, middle and last box), the complement of TGT. The pC2 repeat is apparently slightly shorter than the other sequences. The same pentanucleotide, CAGAT, is found at each end of the p101 and pC241 sequences (arrowed in Fig. 2). However, pC2 and pC3 do not have direct repeats at their termini.

The homology between any two of the salmonid sequences in Fig. 2 is at least 90%, even though p101 is from the rainbow trout and the other clones are from the chum salmon. Also included in Fig. 2 is the homology to the avian sarcoma virus long terminal repeat (bp 151–193), which had been previously described for p101 (Jankowski et al. 1986).

Characterization of the Element

We did some Southern blots of genomic DNA using a fragment from p101 that contained the repeat plus short flanking sequences as a probe (the BglII to DraI fragment in Fig. 1) and also a probe that starts only 80 bp downstream from the end of the repeat and includes the protamine gene, a small amount of 5′ sequence, and the entire 3′ flanking sequence. This protamine probe is actually from the rainbow trout protamine clone, TP21, which is virtually identical to p101 over the length of the probe (Aiken et al. 1983). In Fig. 3, the genomic DNA was from eight individual rainbow trout from several locations in Canada and Europe. There is obvious polymorphism between individuals with both probes but this is particularly marked with the repeat probe. The overall pattern of banding is similar for all individuals but only D2 and D3 appear to be identical. The number of bands seen with the repeat probe is much greater than that seen with the protamine gene probe, suggesting a much greater copy number. This difference also suggests that the end of the repeat as we have defined it by homology is also the end in vivo, since there is only a short gap of 80 bp between these two probes.

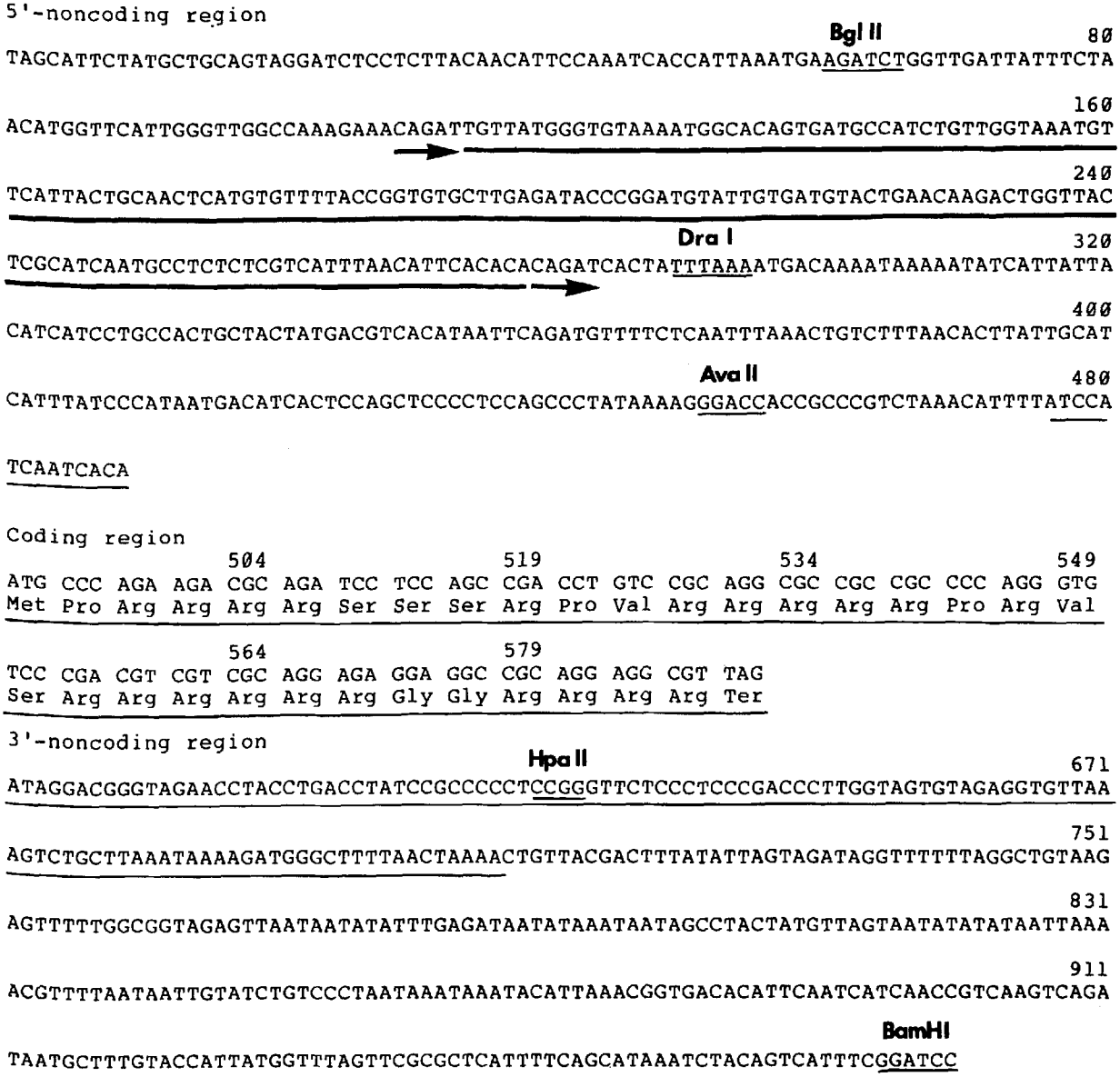

Fig. 1. The nucleotide and predicted amino acid sequence for the rainbow trout protamine gene p101 and flanking regions (States et al. 1982). The transcription unit is indicated with thin underlining and the LTR-like repeat with thick underlining. The arrows indicate short direct repeats at termini of the repetitive element. Restriction enzyme sites referred to in the text are also shown.

We have used the entire pC2 and pC3 subclones (1.6 kb and 1.5 kb, respectively) and a third repeat clone, C251, as probes in Northern blots of RNA from various rainbow trout tissues. We have not been able to detect any transcripts, suggesting that the repeat itself is not transcribed and in the case of pC2, pC3, and C251 is not close enough to other genes to be included in their mRNA sequences. However, a recent report (Suzuki et al. 1986) has identified an LTR-like middle repetitive sequence in the rat that is expressed in tumor cell lines but not in normal tissues. The same may be true of the sequence element from the fish species.

In Fig. 4 the repeat fragment from rainbow trout p101 (BglII to DraI; Fig. 1) was used as a probe to a Southern blot of DNA from various related salmonid fish species. There are large differences in the number of bands seen, suggesting there may be variation in the copy number of repeats between species. Chum salmon DNA, the species from which pC2, pC3, and pC241 were derived, hybridizes very well to the rainbow trout probe.

The Absence of the Repeat from Other Protamine Clones

We used the same probes as in Fig. 3 to determine if the repeat sequence was adjacent to the other rainbow trout protamine genes that have been sequenced in our laboratory (Aiken et al. 1983). The

```
      103                                                          150
p101  AAGAAACAGATTGTTATGGGTGTAAAATGGC...ACAGTGATGCCATCT..GTTG
pC241 --T---------------------------G-----AGC---------T-----GA----
pC2   CCAGCC-CTCC-A-----------G-----...---------------..----
pC3   ----GGG---A-----------------...---------------..----

      151                                                          198
ASV   --------AGTC---.-----TA---T---A--C--GG-AACAT---AAC
p101  GTAAATGTTCA.TTACTGCAA..CTCATGT.GTTTT.AC....CGGTGTGGAGAT
pC241 ----------.---------..---G---.-----.--.....-----T-T....-
pC2   ----------.---------..-------.-----.--....------------
pC3   -----C---A-.---------..---CA--.-----T--....------------

      199                                                          253
p101  GAGATACCCGGATGTA.TTGTGATGTACTGAACAAGACTGGTTACTCGCATCAAT
pC241 ....----------C-A--A-----------------------------------
pC2   --------------C-.-------------------------------T------
pC3   -------T--------.---CA----CA-----------------T--------

      254                                                    285
p101  .GCC.T...CTCTCGTCATTTAACATTC..ACACAGATCACTAT
pC241 G---C-GT.----------------..----------..-G--
pC2   .---.-...-..-------.C-.ACAAAAAATGGT-CTCTTG-G-
pC3   .---.-GTG-----------------AAACAAC-G-G--ATG
```

Fig. 2. An alignment of the sequence common to all four clones that contain the putative repeat. Clones pC2 and pC3 are from chum salmon and contain only the repeat sequence, while p101 and pC241 are rainbow trout and chum salmon clones that contain both the repeat and protamine gene. ASV refers to the sequence from avian sarcoma virus LTR that has homology with the other clones (Jankowski et al. 1986). The TGT/ACA trinucleotides are boxed and the direct repeats in p101 and pC241 are shown by arrows. A dash indicates a base conserved with p101 and a dot represents a gap introduced to maximize homology. The numbering system refers to the numbers for the p101 sequence shown in Fig. 1.

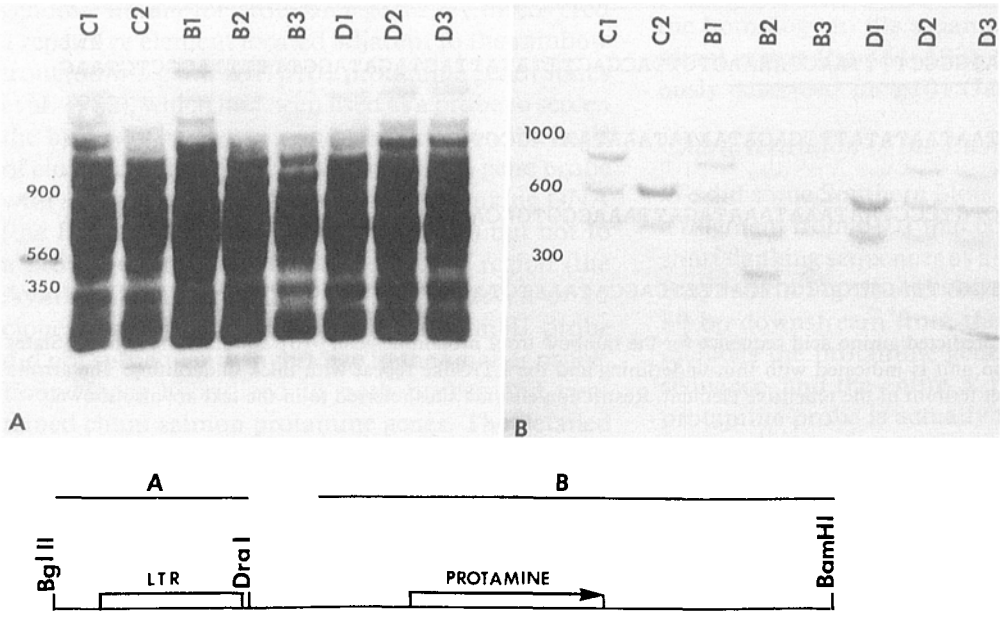

Fig. 3. Autoradiogram of a Southern blot of eight different rainbow trout genomic DNA samples digested with AluI and probed with a repeat-specific probe (**Panel A**) and a protamine gene-specific probe (**Panel B**). The tissue samples were obtained from Calgary in 1985 (C1, C2), Mission, British Columbia, 1983 (B1, B2, B3), and Denmark, 1974 (D1, D2, D3). The location of the CTR and protamine gene are shown by boxes. The direction of transcription is indicated by an arrow.

results of the Southern blots are shown in Fig. 5. Only one other rainbow trout protamine phage clone, TP14, hybridizes to the repeat probe. The 2-kb fragment that contains the hybridizing sequence from TP14 maps 14 kb downstream from the protamine gene (Aiken et al. 1983). Protamine clones TP16 and TP21 do not contain enough 5′ flanking information to determine whether or not the repeat is in

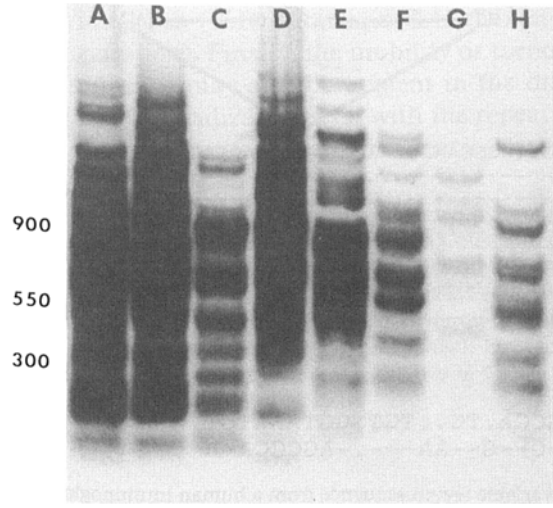

Fig. 4. Autoradiogram of a Southern blot of genomic DNA samples from various salmonid fish species. The samples were digested with AluI and the probe was the repeat-specific fragment (BglII–DraI fragment, Fig. 1). The species were: lane A, rainbow trout (*Salmo gairdneri*); lane B, cutthroat trout (*Salmo clarkii*); lane C, Atlantic salmon (*Salmo salar*); lane D, chum salmon (*Oncorhynchus keta*); lane E, coho salmon (*Oncorhynchus kisutch*); lane F, lake trout (*Salvelinus namaycush*); lane G, mountain whitefish (*Prosopium williamsonii*); lane H, Arctic grayling (*Thymallus arcticus*).

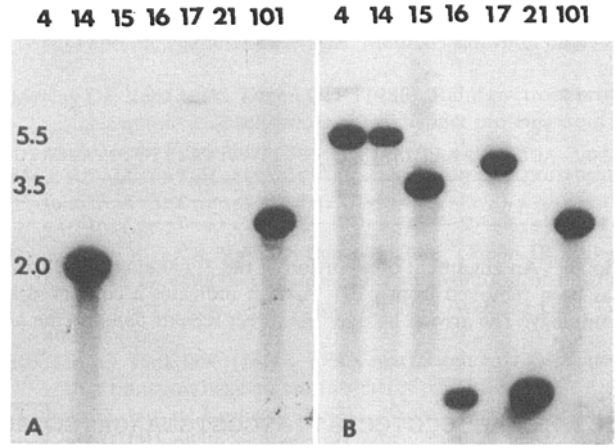

Fig. 5. Autoradiogram of a Southern blot of rainbow trout protamine gene clones probed with a repeat probe (**Panel A**) and a protamine gene probe (**Panel B**). The DNA samples were digested with BamHI and EcoRI. The numbers at the top of the photograph refer to the designation given to the different clones in reference 6. Size markers in kilobases are also shown.

the same position with respect to the gene as p101, and therefore we cannot exclude the possibility that these genes have the repeat.

If the repeat were inserted next only to p101 then, it should have interrupted the extensive homology between p101 and the other protamine genes in their 5′ flanking regions (Aiken et al. 1983). We "removed" the putative repeat sequence from p101 and then aligned the remaining sequence to three other rainbow trout clones (Fig. 6). Only p101 has a duplication of the pentanucleotide CAGAT at the point of insertion of the repeat. However, there are 20 well-conserved bases immediately upstream of this point and there is clearly homology farther upstream. Clones TP14, TP15, and TP17 are identical in this region while p101 is distinct. This alignment does indeed suggest that the insertion of the repeat sequence has interrupted 5′ homology between p101 and the other protamine genes. This insertion event must have occurred in p101 more recently in evolution than the divergence of the various protamine genes themselves.

Discussion

In this report we have described the characteristics of a short repetitive element from the salmonid fishes that has some similarities with long terminal repeats of retroviruses. With one exception the repeats begin and end, respectively, with the trinucleotide TGT and its complement ACA. These sequences are found at the termini of retroviral LTRs and are probably required for integration and excision of proviral DNA (Varmus 1982). The homology between the avian sarcoma virus LTR and the p101 5′ flanking DNA that had been previously identified (Jankowski et al. 1986) lies entirely within the limits of the repeat as we define them (Fig. 2). The representatives of the repeat next to the rainbow trout and chum salmon protamine genes have the same pentanucleotide, CAGAT, flanking the TGT/ACA termini. The same observation is also made with host sequences at the ends of intact retroviruses and represents a duplication of the target DNA during integration of the proviral DNA (Chen and Barker 1984). The clones pC2 and pC3 do not have direct repeats flanking their termini (Fig. 2), but we note that pC3 has a polypurine tract immediately preceding it. This sequence is found immediately 5′ of the 3′ LTR of most retroviruses and may serve a role in viral DNA replication (Varmus 1982). Clones pC2 and pC3 may, therefore, be the 5′ and 3′ LTRs, respectively, of two different representatives of a retrovirus or retroviral-like repetitive element. A sequence immediately 3′ of the repeat from pC2 (Fig. 2) shows similarities with the 18-bp sequence (TBS) of most avian retroviruses that binds tryptophan tRNA (Chen and Barker 1984). The tRNA acts as a primer for viral DNA synthesis (Varmus 1982). The first seven and last five base pairs are a good match between the two sequences, but the homology in the central region is weak. The CR1 elements of chicken, which also have some similarities to LTRs, also have good matches for TBS for only a few base pairs (Stumph et al. 1982). Although these

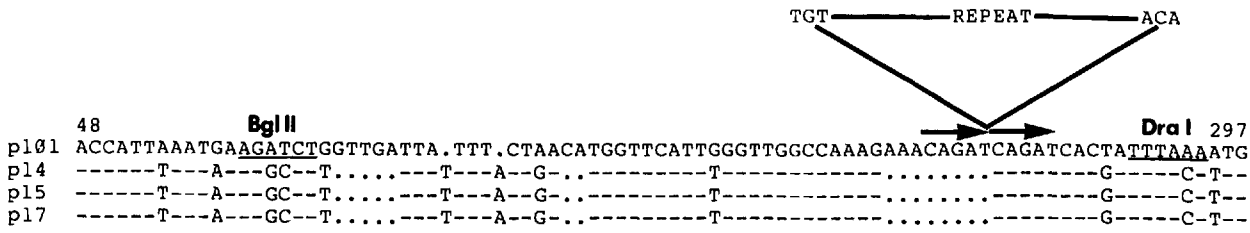

Fig. 6. An alignment of a portion of the 5′ flanking sequences of different rainbow trout protamine clones after the putative repeat has been removed from p101. A dash indicates a conserved base with p101 and a dot represents a gap introduced to maximize homology. The arrows lie over the direct repeats flanking the LTR-like sequence.

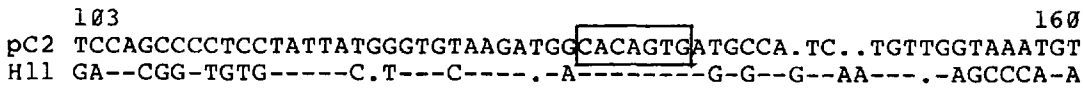

Fig. 7. Alignment between the 5′ region of pC2 and the 3′ region of H11, a variable region sequence from a human immunoglobulin heavy-chain gene (Rechavi et al. 1982)

structural similarities are suggestive, it would be necessary to isolate a clone that contains two of these repeats flanking open reading frames to demonstrate that the repeats are LTRs. A functional test for retrotransposition could be designed for such a clone as has been done for the TY1 element of yeast (Boeke et al. 1985). The p101 and pC241 repeat sequences may represent solo LTRs, unlinked to a retrovirus. Similar observations have been made for chicken, mouse, yeast, and rat retroviral-like LTRs (Roeder and Fink 1980; Rotman et al. 1984; Stumph et al. 1984; Suzuki et al. 1986). The solo LTRs may result from homologous recombination between the LTRs of a retrovirus (Rotman et al. 1984). Alternatively, it has been suggested that LTRs are capable of an independent existence, presumably as the objects of passive transposition (Temin 1982).

The repeat sequence we have described is only one-half the length of even the shortest retroviral LTR and lacks TATA and polyadenylation signals (Temin 1982). However, a TATA box and polyadenylation signal are found closely placed together 70 bp after the end of the repeat in pC2, and a good TATA sequence is found 20 bp upstream of the pC3 repeat (not shown). A TATA/polyadenylation sequence combination is also found immediately after the end of the p101 repeat (Fig. 1), and this TATA sequence is functional in vitro (Jankowski et al. 1986).

The LTR-like sequence also has a region of homology with the variable (V) region of an immunoglobulin heavy-chain gene. Fig. 7 shows an alignment between pC2 and the immediate 3′ region of the human V segment H11 (Rechavi et al. 1982). The homology includes the conserved heptamer (boxed in Fig. 8) which occurs at the 3′ limit of the V coding region and probably serves as a recombination signal for V–J joining (Sakano et al. 1980). The 5′ limit of homology between the two sequences occurs at the beginning of the pC2 repeat and includes several amino acids of the V coding region. The homology extends downstream of the heptamer for only a few bases. There is not a good match in pC2 for the conserved nonamer, which is a second recombination signal found in H11 and other immunoglobulin genes (Sakano et al. 1980). The insertion of LTR-like sequence elements into the genome may be mediated by the same system that rearranges immunoglobulin genes.

Although the LTR-like repeat has not yet been found adjacent to any of the other rainbow trout protamine genes, we cannot reach a definite conclusion about TP16 and TP21 because their 5′ sequences are too short. However, the LTR-like repeat is present in exactly the same position in a chum salmon protamine clone, which is a salmonid fish from a different genus (*Oncorhynchus*). It has been previously suggested that a retrovirus could have acted as a horizontal transmission vector for the introduction of protamine genes into salmonid fishes (Jankowski et al. 1986). If this repeat provides the evidence for such a process, then it has been specifically deleted from rainbow trout protamine clones other than p101. Alternatively, it may have been inserted next only to the p101 gene after the introduction of protamine genes into fish species. We cannot speculate on the effect of this repeat on p101 gene transcription. A protein corresponding to the one coded for by p101 and p15 is present in roughly equal parts in a maturing rainbow trout testis, although the relative amounts may vary during development (McKay et al. 1986).

The possible mobility in the genome of the LTR-like repeat or its parent retrovirus may be responsible for the polymorphism in restriction sites we see using both repeat and protamine gene-specific probes (Fig. 3). We see a far lower degree of polymorphism in the histone gene clusters of the same

individual fish (R.D. Moir and G.H. Dixon 1988, in preparation). Finally, the mobility or turnover of this sequence may also be evident in the different degrees of hybridization seen with the repeat probe (Fig. 4) to different salmonid fish species sequences.

Acknowledgments. This work was supported by an operating grant to G.H.D. from the Medical Research Council of Canada. R.D.M. was the recipient of a studentship from the Alberta Heritage Foundation for Medical Research. We thank Bob Winkfein for the loan of his Northern filters and Marnie Cudmore for preparation of this manuscript. We also thank Dr. Jacek Jankowski for pointing out the homology shown in Fig. 8.

References

Aiken JM, McKenzie D, Zhao H-Z, States JC, Dixon GH (1983) Sequence homologies in the protamine gene family of the rainbow trout. Nucleic Acids Res 10:7581–7592

Bencini DA, O'Donovan GA, Wild JR (1984) Rapid chemical degradation sequencing. Biotechniques 2:4–5

Boeke JD, Garfinkel DJ, Styles CA, Fink GR (1985) T_Y elements transpose through an RNA intermediate. Cell 40:491–500

Chen HR, Barker WC (1984) Nucleotide sequences of the retroviral long terminal repeats and their adjacent regions. Nucleic Acids Res 12:1767–1778

Feinberg A, Vogelstein B (1983) A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal Biochem 132:6–13

Garfinkel DJ, Boeke JD, Fink GR (1985) T_Y element transposition: reverse transcriptase and virus-like particles. Cell 42:507–517

Garger SJ, Griffith OM, Grill LK (1983) Rapid purification of plasmid DNA by a single centrifugation in a two-step cesium chloride–ethidium bromide gradient. Biochem Biophys Res Commun 117:835–842

Jankowski JM, States JC, Dixon GH (1986) Evidence of sequences resembling avian retrovirus long terminal repeats flanking the trout protamine gene. J Mol Evol 23:1–10

Maniatis T, Fritsch E, Sambrook J (1982) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor NY

Maxam A, Gilbert W (1980) Sequencing end-labeled DNA with base specific chemical cleavages. Methods Enzymol 65:499–559

McKay DJ, Renaux BS, Dixon GH (1986) Rainbow trout protamines: amino acid sequences of six distinct proteins from a single testis. Eur J Biochem 158:361–366

Moir RD, Dixon GH (1988) Characterization of a protamine gene from the chum salmon (*Oncorhynchus keta*). J Mol Evol 26:130–148

Rechavi G, Bienz B, Ram D, Ben-Neriah Y, Cohen JB, Zakut R, Givol D (1982) Organization and evolution of immunoglobulin U_H gene subgroups. Proc Natl Acad Sci USA 79: 4405–4409

Roeder SG, Fink GR (1980) DNA rearrangements associated with a transposable element in yeast. Cell 21:239

Rotman G, Hin A, Keshet E (1984) "Solo" large terminal repeats (LTR) of an endogenous retrovirus-like gene family (VL30) in the mouse genome. Nucleic Acids Res 12:2273–2282

Sakano H, Maki R, Kurosawa Y, Roeder W, Tonegawa S (1980) Two types of somatic recombination are necessary for the generation of complete immunoglobulin heavy-chain genes. Nature 286:676–683

States JC, Connor W, Wosnick MA, Aiken JM, Gedamu L, Dixon GH (1982) Nucleotide sequence of a protamine component CII gene of *Salmo gairdneri*. Nucleic Acids Res 10:4551–4563

Stumph WE, Hodgson CP, Tsai M-J, O'Malley BW (1984) Genomic structure and possible retroviral origin of the chicken CR1 repetitive DNA sequence family. Proc Natl Acad Sci USA 81:6667–6671

Suzuki N, Fujiyoshi T, Maehara Y, Takahasi K, Yamamoto M, Endo H (1986) A new family of LTR-like sequences abundantly expressed in rat tumors. Nucleic Acids Res 14:9271–9289

Temin HM (1982) Function of the retrovirus long terminal repeat. Cell 28:3–5

Varmus HE (1982) Form and function of retroviral proviruses. Science 216:812–820

Weiner AM, Deninger PL, Efstratiadis A (1986) Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu Rev Biochem 55:631–661

Received February 20, 1987/Revised July 10, 1987