J Mol Evol (1985) 21:209–219

Journal of Molecular Evolution

An H1 Histone Gene from Rainbow Trout (*Salmo gairdnerii*)

J. Mezquita, W. Connor, R.J. Winkfein, and G.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. A 1.7-kbp DNA region from the 10.2-kb cluster containing the five rainbow trout histone genes has been subcloned in pBR322 and completely sequenced. It contains a trout histone H1 gene together with its 5′ and 3′ flanking sequences. This H1 gene codes for a H1 variant different from the major trout testis H1 previously sequenced by Macleod et al. (1977). Northern blots of total RNA from trout testis, kidney, and liver indicate that this H1 gene is expressed in all three tissues but that the level of H1 mRNA is much higher in testis than in other tissues. The lack of heterogeneity in the sizes and 5′ initiation sites of trout H1 mRNAs is surprising in view of the substantial heterogeneity of H1 variant proteins observed previously. The coding sequence of the H1 gene shows strong evidence of repeated partial duplications of a hexapeptide motif of the form Ala.Ala.Ala.Lys.Lys.Pro and of a pentapeptide phosphorylation-site sequence, Lys.Ser.Pro.Lys.Lys, during its evolution.

Comparisons are drawn between this gene and the coding sequences of other vertebrate H1 genes from chicken and *Xenopus*, and a strong homology is seen in the region of amino acids 22–101, which form the hydrophobic "head" of the H1 molecule. The 5′ and 3′ regulatory signals in the trout H1 are also compared with those of H1 genes from other sequences.

Key words: Histone genes — Histone H1 — DNA sequence — Rainbow trout — *Salmo gairdnerii*

Offprint requests to: G.H. Dixon

Introduction

When compared with the four core histones, histone H1 shows unique properties. It is larger, and is composed of three structural domains: a flexible N-terminal "nose"; a central, hydrophobic, globular "head"; and a flexible, very basic C-terminal "tail" (Hartman et al. 1977; McGhee and Felsenfeld 1980). The amino acid sequence of the globular domain has been conserved during evolution significantly more than have those of the termini (Von Holt et al. 1979; McGhee and Felsenfeld 1980). Unlike the core histones, for which a limited number of variants have been described (Cohen et al. 1975; Franklin and Zweidler 1977), histone H1 shows both species and tissue specificity (Kinkade and Cole 1966; Bustin and Cole 1968; Kinkade 1969; Cole 1984). Moreover, histones of the H1 group undergo faster metabolic turnover than do the core histones (Djondjurov et al. 1983). Histone H1 plays a fundamental role at several levels of chromatin organization: within the nucleosome, between neighboring nucleosomes, and in the packaging of chromatin into a higher-order organization (Klingholz and Stratling 1982). Thus, the heterogeneity of histone H1 might be related to functional differences between subfractions (Kinkade and Cole 1966; Bustin and Cole 1968; Kinkade 1969). In *Xenopus laevis*, five H1 subtypes have been described (Risley and Eckhardt 1981), while in mammals, six different H1 subtypes have been resolved by two-dimensional gel electrophoresis (Lennox and Cohen 1983; Lennox 1984). In the rat, a testes-specific H1t has been described (Seyedin and Kistler 1980; Seyedin et al. 1981).

In the developing trout testis there appears to be a single major H1 species (Macleod et al. 1977), but other trout tissues possess much more heterogeneous sets of H1 proteins (Seyedin and Cole 1981). A total of eight chromatographic fractions are found in trout liver, and five in kidney (Brown and Goodwin 1983). Some of these fractions are unusual in containing methionine as well as tyrosine residues. The extra tyrosine seems to be located in the N-terminal region of the molecule, and could have important effects on the interaction of H1 in chromatin. Kidney fractions (d) and (e) show high levels of arginine, and might correspond to subtypes such as erythrocyte type H5 (Miki and Neelin 1975) or $H1_0$ (Panyim and Chalkley 1969; Srebreva and Zlatanova 1983).

Connor et al. (1984a,b) have isolated and characterized a series of clones containing a common 10.2-kb DNA fragment from a lambda Charon 4A rainbow trout genomic library (States et al. 1982). The five-membered histone cluster is located in this fragment. The gene order, H4-H2B-H1-H2A-H3, is unique to the trout, and the overwhelming majority of trout histone genes show this organization. There is no sign in the trout of the divergence in histone-gene organization seen in many higher vertebrates (Maxson et al. 1983). The gene sequences for rainbow trout H2A and H3 have recently been determined (Connor et al. 1984b); in the present article, we describe the coding and flanking nucleotide sequences of the H1 gene. Knowledge of the 5′ and 3′ regions of the H1 gene, in which regulatory signals are likely to be located, is essential if we are to obtain an understanding of the mechanism by which the expression of H1 genes is regulated during development and differentiation. In addition, the availability of comparative sequence information for the H1 gene from several different species will provide a basis for the study of the evolution of the H1 gene.

Materials and Methods

Materials. Restriction enzymes were purchased from Bethesda Research Laboratories or New England Biolabs and used as recommended by the manufacturers. DNA polymerase 1, calf intestine alkaline phosphatase, and unlabeled deoxynucleotides were from Boehringer Mannheim; T4 polynucleotide kinase and S1 nuclease from Pl Biochemicals; terminal transferase from Amersham; radiochemicals from Amersham, New England Nuclear, or ICN; agarose from BioRad; nitrocellulose sheets and disks from Schleicher & Schuell; Gene Screen from New England Nuclear; Pall Biodyne from Pall Canada; and formamide from Fluka. Most other chemicals were purchased from Fisher or Sigma.

Southern Blotting. An M13 subclone containing the newt (*Notophthalmus viridescens*) H1 gene was kindly supplied by Dr. J. Gall, Department of Biology, Yale University, New Haven, Connecticut. A 900-bp Ava I–Eco RI fragment containing the H1 gene was isolated from a 1% agarose gel using low-melting-temperature agarose, and after nick translation (Maniatis et al. 1975) was used as a probe. Recombinant DNAs were digested with restriction enzymes and electrophoresed on 1% agarose gels. Fragments were denatured in 0.4 N NaOH–1 M NaCl for 20 min at room temperature; neutralized in 0.5 M Tris–Cl^-, pH 8.0, for 15 min; and electroblotted to Gene Screen in agarose gel buffer. Hybridization was carried out at 45°C in a solution containing 50% formamide; 0.6 M NaCl; 0.12 M Tris–Cl^-, pH 8.0; 4 mM ethylenediaminetetraacetate (EDTA); 0.1% sodium dodecyl sulfate (SDS) and sodium pyrophosphate; 0.2% Ficoll, PVP, and bovine serum albumin (BSA); 100 mg/ml yeast RNA; and 10% dextran sulfate. Washing of the filters was done at 50°C at a salt concentration of 0.3 M.

Subclones. Restriction fragments from the recombinant λTH2 were subcloned in pBR322. pTH243 contains the Bam HI–Cla I fragment inserted into the Bam HI and Cla I sites of pBR322. pTH224 was prepared from pTH243 by cutting with Nde I and religating the 4-kb fragment that contains the origin of replication. pTH210 and pTH211 contain each of the Bal I fragments inserted in the Cla I site of pBR322 after the addition of Cla I linkers. The subclones were transformed into *Escherichia coli* RR1 and the transformed bacteria were selected by their resistance to ampicillin. The plasmids were screened by restriction analysis of minipreparations of DNA from each colony (Birnboim and Doly 1979). The DNA was amplified with chloramphenicol and purified through CsCl gradients (Maniatis et al. 1975).

DNA Sequencing. The DNA fragments were sequenced by the method of Maxam and Gilbert (1980). Five sequencing reactions were used: G, G+ A, A > C, C + T, and C. The DNA was 5′-end-labeled with polynucleotide kinase after pretreatment with alkaline phosphatase, or with terminal transferase at the 3′ end; labeled fragments were then cleaved with a second restriction enzyme. The single-end-labeled fragments were separated in 1–1.5% agarose gels and isolated using low-melting-temperature agarose or electroelution. The reaction mixtures were electrophoresed in 80-cm 8% or 6% polyacrylamide–urea gels and 40-cm 20% polyacrylamide–urea gels (Maxam and Gilbert 1980). Most sequences were determined from both strands at least twice. The sequences were analyzed using the computer programs of Dr. D. McKay, Department of Medical Biochemistry, University of Calgary, Calgary, Alberta, Canada.

Northern Blotting. Five micrograms of testes mRNA and 10 μg of mRNA from kidney and liver were electrophoresed in 2% agarose methylmercury gels (Bailey and Davidson 1976) and electroblotted to Pall Biodyne membranes in phosphate buffer. The blot was hybridized to a nick-translated Nae I-Hinc II fragment (2×10^7 cpm/μg). Hybridization was performed as recommended by the manufacturers of Pall Biodyne membranes, and washing was at 50°C in 0.075 M salt.

Results and Discussion

Restriction Maps and Sequencing Strategies for the Trout H1 Gene

In a previous paper in this series (Connor et al. 1984a), detailed restriction analysis of seven histone-gene-containing clones selected from a partial Eco RI genomic library prepared in lambda Charon 4A showed that six of the seven contained a 10.2-kb common Eco RI fragment including all five histone

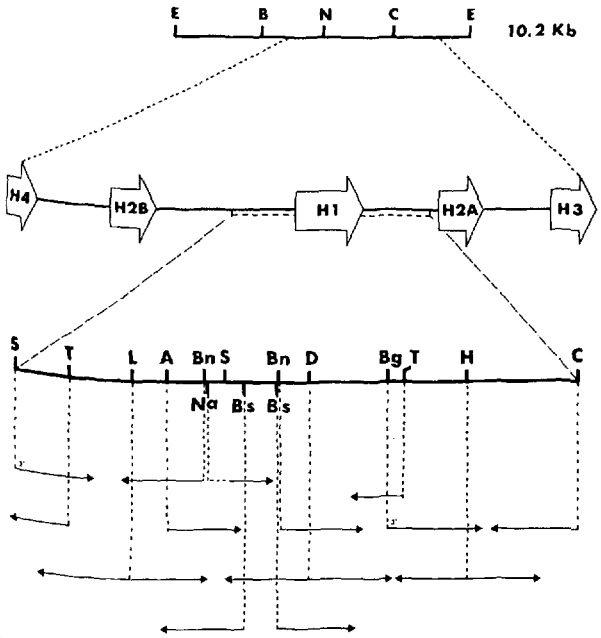

Fig. 1. Top: Organization of the λTH2 clone, showing the position and orientation of the H1 histone gene. A simplified restriction map is also shown, with restriction sites labeled as follows: E, Eco RI; B, Bam HI; N, Nde I; C, Cla I. Bottom: Strategy for sequencing of the H1 histone gene. The fragments were labeled at the 5′ or 3′ ends where indicated and sequenced by the method of Maxam and Gilbert (1980). In the case of the Bal I site, the corresponding fragments were sequenced by labeling the Eco RI (23 bp apart) and Hind III (6 bp apart) sites of the vector from subclones pTH210 and pTH211. Restriction sites are labeled as follows: A, Acc I; Bs, Bst NI; Bn, Ban I; Na, Nae I; D, Dde I; Bg, Bgl I; T, Taq I; H, Hinc II; L, Bal I; S, Sst I; C, Cla I

genes in the order H4-H2B-H1-H2A-H3. Initially considerable difficulty was encountered in obtaining hybridization with nonhomologous probes for H1. No hybridization was observed with a *Drosophila* H1 probe. However, an H1 probe from the newt, *Notophthalmus viridescens* (Stephenson et al. 1981), hybridized to a 1.1-kb fragment bordered by Sst I and Cla I sites. In Fig. 1 (top) a simplified restriction map of λTH2 is presented that shows the position and orientation of the H1 gene in the five-membered histone cluster. To facilitate the sequencing of the H1 gene, several subclones were constructed in pBR322 as follows: The Bam HI–Cla I fragment was inserted into the Bam HI–Cla I site of pBR322 to form pTH243. The Nde I–Cla I fragment was inserted between the Nde I site near the origin of replication and the Cla I site of pBR322; a large portion of pBR322 between the two Nde I sites was eliminated by cutting pTH243 with Nde I and religating the 4-kb fragment to give rise to pTH224. In addition, two different Bal I fragments were ligated to Cla I linkers and inserted into the Cla I site of pBR322. One (pTH210) comprises a 5′ region flanking the H1 gene and the second (pTH211) is immediately 3′ to the first and contains the H1 coding region as well as its 3′ flanker. There was a general paucity of suitable restriction sites within the H1 coding region, but using the four subclones listed above and employing the sequencing strategy depicted in Fig. 1 (bottom), it was possible to deduce an unequivocal DNA sequence by the Maxam–Gilbert method. Most of the sequence determinations were repeated several times on both DNA strands.

DNA Sequence of the Trout H1 Gene

Figure 2 shows the sequence of 1707 bp of DNA containing the H1 gene plus its flanking regions and spanning positions −540 to +1167 (with the T of the initiating Met codon ATG taken as position +1).

Coding Region

In the coding region of H1, an open reading frame predicts a polypeptide of 206 amino acids (excluding the initiating methionine) that is a characteristic H1 histone in being very rich in Lys (30%), Ala (28%), and Pro (9%) and in showing three structural domains: a hydrophobic "head" from amino acid 45 to amino acid 117 that corresponds to the strongly conserved globular region previously described in several H1 histones (Von Holt et al. 1979); an Ala-, Pro-, and Lys-rich "nose" (residues 1–44), which has been shown to be very variable in different histones; and a very basic C-terminal half (residues 118–206) characterized by a hexapeptide motif, Ala.Ala.Ala.Lys.Lys.Pro, that repeats, with minor variations, five times.

In 1977, Macleod et al., from this laboratory, determined the complete amino acid sequence of the major H1 component of trout testis; in Fig. 3, the two sequences are compared. Clear differences are seen in both the N- and the C-terminal regions. Between positions 25 and 40 in the "nose" region, there are five single-amino-acid substitutions: at residue 27, Ala for Ser [with the amino acid predicted from the DNA sequence indicated first and that from the Macleod et al. (1977) protein sequence second]; at residue 30, Ser for Ala; at residue 35, Ile for Ala; at residue 36, Val for Gly; and at residue 40, Ser for Ala. It is interesting that in the amino acid sequence determination of trout testis H1 (Macleod et al. 1977), a microheterogeneity was observed at position 35, with Ala the major residue and Ile the minor. This would suggest that the H1 gene sequenced here could be that for the minor trout H1 variant. The Ser → Ala changes at residues 30 and 40, if present as microheterogeneities, would not have been recognized in the amino acid sequence determination since back-hydrolysis of the phenylthiazolinones of the amino acids was used as

H1 sequence

5'-noncoding region

```
      -532       -522       -512       -502       -492       -482       -472
GAGCTCTCCT TTAGAACACC AAGGAGCACA TCAGGGGACG CCAAACCATT TGATTCCCTT GCCAGACATC

      -462       -452       -442       -432       -422       -412       -402
TCGGCTCACT CCACAATCAA TGGTGCTCGC AATCGGATGC ACTTTTAGGC CATTTAGGAG ACAAATAGCA

      -392       -382       -372       -362       -352       -342       -332
CAGGAAAAGC AGTGCGCACC GCTCCACCCG ACAGTCGAAC GGTGAGGTTT TGAGCGAGTC GCAGCAAAAT

      -322       -312       -302       -292       -282       -272       -262
GCACGGGCTT CTGCAGCCCA CATGACTTTA TTCTGAACGG ACACAAGTCT GCTCGCTGGG CCGTTCGCTT

      -252       -242       -232       -222       -212       -202       -192
TTGGGCCAAA AACACGGCTC CGTCGGTGAC TTTTGGCCCG ATATTGGCGA CCAGAAAACA CAAGTGAAAG

      -182       -172       -162       -152       -142       -132       -122
AGCATTTGGC CAGCCCGGAG AAGCCGAGCT GGGTGGCTTG AGTCTACATG GTTCTCATGT CGCGTTTAAG

      -112       -102        (+1)       -82        -72        -62        -52
GCCAGCCCCC TGCACGGTGT GGAGCTTCAA TAGCGCAGAG CAGCGTCTAC AGCAAAGTAC TCCTCCTCAC

       -42        -32        -22        -12         -2
AGACTACCGT AGTGTTGTAA GTGTGTGTGT TTACCGGACC GAACGACAGA C
```

Coding region

```
                    14                  29                  44                  59
ATG GCA GAA GTC GCA CCA GCA CCC GCC GCC GCG CCG GCC AAG GCA CCC AAG AAG AAG
Met Ala Glu Val Ala Pro Ala Pro Ala Ala Ala Ala Pro Ala Lys Ala Pro Lys Lys Lys

                    74                  89                 104                 119
GCA GCA GCC AAG CCC AAG AAA GCG GGA CCC AGC GTA GGC GAG CTC ATC GTC AAG GCG GTG
Ala Ala Ala Lys Pro Lys Lys Ala Gly Pro Ser Val Gly Glu Leu Ile Val Lys Ala Val

                   134                 149                 164                 179
TCC GCC TCC AAG GAG AGG AGC GGC GTG TCC CTG GCC GCG CTC AAG AAG TCT CTG GCG GCA
Ser Ala Ser Lys Glu Arg Ser Gly Val Ser Leu Ala Ala Leu Lys Lys Ser Leu Ala Ala

                   194                 209                 224                 239
GGC GGC TAC GAC GTG GAG AAG AAC AAC TCC CGT GTC AAG ATC GCC GTC AAG AGC CTC GTC
Gly Gly Tyr Asp Val Glu Lys Asn Asn Ser Arg Val Lys Ile Ala Val Lys Ser Leu Val

                   254                 269                 284                 299
ACC AAG GGC ACC CTG GTC CAG ACC AAG GGC ACC GGT GCT TCC GGC TCC TTC AAG CTC AAC
Thr Lys Gly Thr Leu Val Gln Thr Lys Gly Thr Gly Ala Ser Gly Ser Phe Lys Leu Asn

                   314                 329                 344                 359
AAG AAG GCC GTC GAG GCA AAG AAG CCC GCC AAG AAA GCC GCA GCC CCC AAA GCT AAG AAG
Lys Lys Ala Val Glu Ala Lys Lys Pro Ala Lys Lys Ala Ala Ala Pro Lys Ala Lys Lys

                   374                 389                 404                 419
GTG GCC GCC AAG AAG CCC GCC GCC GCC AAG AAG CCC AAG AAG GTA GCA GCC AAG AAG GCC
Val Ala Ala Lys Lys Pro Ala Ala Ala Lys Lys Pro Lys Lys Val Ala Ala Lys Lys Ala

                   434                 449                 464                 479
GTG GCC GCA AAG AAG TCC CCC AAG AAG GCC AAG AAG CCC GCT ACA CCC AAA AAG GCC GCC
Val Ala Ala Lys Lys Ser Pro Lys Lys Ala Lys Lys Pro Ala Thr Pro Lys Lys Ala Ala

                   494                 509                 524                 539
AAG AGC CCA AAG AAG GTG AAG AAG CCC GCC GCA GCG GCC AAG AAA GCG GCC AAG AGC CCC
Lys Ser Pro Lys Lys Val Lys Lys Pro Ala Ala Ala Ala Lys Lys Ala Ala Lys Ser Pro

                   554                 569                 584                 599
AAG AAG GCT ACC AAG GCA GCG AAG CCC AAA GCC GCC AAG CCC AAG GCG GCC AAG GCC AAG
Lys Lys Ala Thr Lys Ala Ala Lys Pro Lys Ala Ala Lys Pro Lys Ala Ala Lys Ala Lys

                   614
AAG GCA GCC CCC AAG AAG AAG TAA
Lys Ala Ala Pro Lys Lys Lys Ter
```

3'-noncoding region

```
       633        643        653        663        673        683        693
ACCTATTACA AACAGTGTTC TTTCTACTCG ACACATGTTG TTACCACAAA AGGCTCTTTT AAGAGCCACC

       703        713        723        733        743        753        763
CACCTCTTTC CATAAAAGCG CATGTCATTC CATTCCACCT ACCTACCCGT GGTGCAAAAG AAATGAAATG

       773        783        793        803        813        823        833
AATGACTTTT ACGCACCACA TTTTCGAGTG GCTAAATGGC TTTACATTTG TCACTCAAGA GTGCAGCACC

       843        853        863        873        883        893        903
CTCACCAGTC AACATTGTTG TGATGTGTTA GAATTGGCAT GTCACATTGA TCCTGATAAT AAGAAGCATA

       913        923        933        943        953        963        973
CATACTATAG TGGGTTGGAC AGCAGCCATT GGTATTATGA GCCACTCTCG AATTGATTGA TACACGTTTC

       983        993       1003       1013       1023       1033       1043
AGTGATTTGA GTAGTCACTT TGGCGCTCCC GCCAACGTGA AAAAGTCACA AAAGTCGGAG GGAAACAGAG

      1053       1063       1073       1083       1093       1103       1113
TAGCCTAGCC CTCGTCACTC TGCTGCCTGC CTGCCTGCGG GTGGGTGGGG GCGGGCTTAG GGCTGACTGC

      1123       1133       1143       1153       1163       1173
CTGTCTGTCT GTCCCCTCCC TCACTCCAAT GGATAGGCCA CACCGGTCCG GTGGCCAATC GAT
```

Fig. 2. Sequence of the H1 histone gene. Some of the putative control sequence elements are underlined in both the 5' and 3' flanking regions

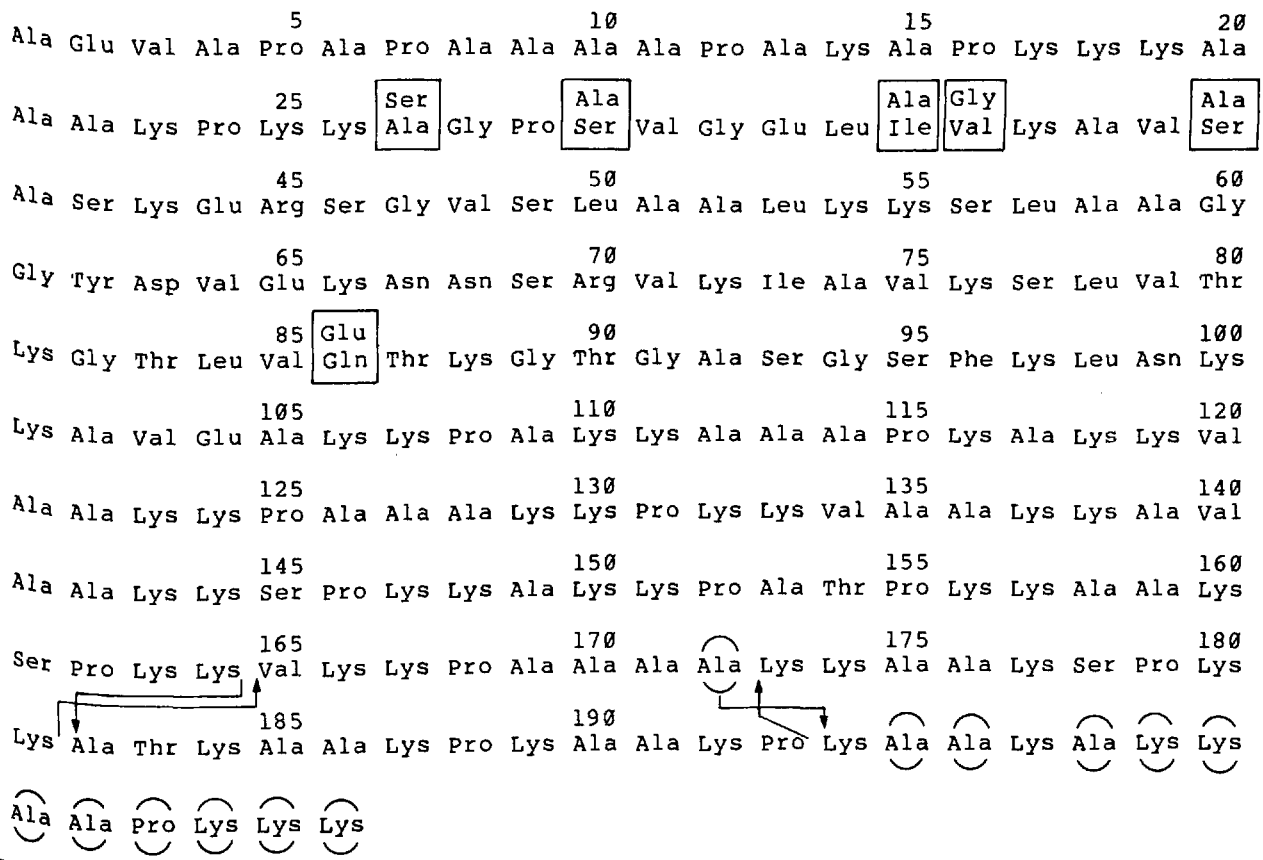

Fig 3. Amino acid sequence comparison between the H1 protein predicted from the gene sequence and the sequence determined by Macleod et al. (1977) for the testis-specific H1 protein. The amino acid differences are shown in boxes, with the predicted residue at bottom. The amino acids that are not present in the testis H1 protein are indicated in circles. The arrows indicate rearrangements that would increase the homology between the sequences

the major identification procedure, and both Ser and Ala derivatives are converted to Ala in this procedure (although in low and high yield, respectively). Similarly, the Gln → Glu change at position 86 would not be distinguished in the protein study. However, the Val → Gly change at position 36 must be a real change, since these residues are easily distinguished during amino acid sequence determination.

At the C-terminus, a much more substantial difference between the two sequences is apparent. The gene sequence predicts a coding region 12 amino acid residues longer than the determined protein sequence, for a total length of 206 rather than 194 residues. There is precise identity (with the exception of the Glu → Gln change at 86) from residue 41 to residue 164, but beyond that point, right to the C-terminus, direct homology is nonexistent. However, it was noticed that substantial stretches of homology could be reestablished if sections of the determined amino acid sequence were rearranged as indicated in Fig. 3.

In any comparison between a coding region predicted from the DNA sequence of a gene and the cognate sequence determined directly from an isolated protein, the possibility of error in the determination of either the DNA or the protein sequence and the possibility that a multi-gene family exists and that the gene sequenced codes for a protein variant different from the one previously sequenced must be considered. In the present case, the differences are sufficiently substantial that it is probable that the sequenced H1 gene codes for a different H1 variant.

However, this conclusion cannot be a definite one, in view of the unusual features of the C-terminal region of trout H1, which posed very difficult problems in the protein sequence determination (Macleod et al. 1977). The predominance of three amino acids, Lys, Ala, and Pro, disposed in an extended series of tandemly repeating hexapeptide units made it very difficult to be certain either that these almost identical units were correctly ordered or that one or more completely identical units might not be repeated twice. For example, a misordering of the peptide R2 sequence Ser-Pro-Lys-Lys-Val-Lys-Lys-Pro-Ala-Ala-Ala-Lys-Lys (Macleod et al. 1977), obtained after the N→O acyl shift followed by cleavage with hydroxylamine in the amino acid sequence determination, could partially account for the C-terminal rearrangement. However, this pep-

Table 1. Codon usage in H1 histone gene

TTT-Phe	0	TCT-Ser	1	TAT-Tyr	0	TGT-Cys	0
TTC-Phe	1	TCC-Ser	7	TAC-Tyr	1	TGC-Csy	0
TTA-Leu	0	TCA-Ser	0	TAA-Ter	1	TGA-Ter	0
TTG-Leu	0	TCG-Ser	0	TAG-Ter	0	TGG-Trp	0
CTT-Leu	0	CCT-Pro	0	CAT-His	0	CGT-Arg	1
CTC-Leu	4	CCC-Pro	16	CAC-His	0	CGC-Arg	0
CTA-Leu	0	CCA-Pro	2	CAA-Gln	0	CGA-Arg	0
CTG-Leu	3	CCG-Pro	1	CAG-Gln	1	CGG-Arg	0
ATT-Ile	0	ACT-Thr	0	AAT-Asn	0	AGT-Ser	0
ATC-Ile	2	ACC-Thr	5	AAC-Asn	3	AGC-Ser	5
ATA-Ile	0	ACA-Thr	1	AAA-Lys	6	AGA-Arg	0
ATG-Met	1	ACG-Thr	0	AAG-Lys	57	AGG-Arg	1
GTT-Val	0	GCT-Ala	4	GAT-Asp	0	GGT-Gly	1
GTC-Val	7	GCC-Ala	31	GAC-Asp	1	GGC-Gly	7
GTA-Val	2	GCA-Ala	14	GAA-Glu	1	GGA-Gly	1
GTG-Val	6	GCG-Ala	9	GAG-Glu	4	GGG-Gly	0

tide was placed at the C-terminus in the protein determination by virtue of the presence of two residues of Lys at its C-terminus that could be removed by carboxypeptidase B. Internal peptides resulting from N→O acyl shift followed by hydroxylamine cleavage are blocked to carboxypeptidases at their C-termini by hydroxamic acid functions, and would not be susceptible to carboxypeptidase B. A N→O acyl rearrangement of the sequence predicted from the DNA would give rise to a much longer 29-reside peptide (178–206) of composition Ser_1, Pro_4, Ala_{10}, Lys_{13}, Thr_1. No peptide of this size or composition was isolated from the acyl-shift reaction mixture, and the ones that were (R1–R3, Macleod et al. 1977) are not consistent with such a peptide.

However, in view of the widespread occurrence of tissue-specific variants of H1 (Cole 1984; Lennox 1984), and particularly the observation by Seyedin and Cole (1981) that the H1 subtype profile in trout testis is clearly less complex than that in trout kidney or liver, it is not at all surprising that the first trout H1 gene sequence to be cloned should code for a H1 variant other than the major one expressed in testis, a tissue for which testis-specific H1t's have been described in other species (Seyedin and Kistler 1980; Seyedin et al. 1981). Furthermore, Ajiro et al. (1981) have shown that in HeLa S-3 cells, the H1A and H1B variants differ in size by ca. 1000 daltons and that this difference occurs near the C-terminus. It is precisely in this C-terminal region that the major difference between the trout testis H1 protein and the protein predicted from the trout H1 gene coding region—the 12 additional amino acids predicted from the gene coding region—is located. This would lead to a size difference of 1337 daltons—close to the estimated 1000-dalton difference between HeLa H1A and H1B. Thus, the overall conclusion is that the H1 gene characterized here codes for a H1 variant different from the major one expressed in trout testis and previously sequenced.

Codon Usage in the Trout H1 Gene

As in the total trout histone cluster (Connor et al. 1984a,b), there is a strong preference in the H1 coding region for usage of codons ending in C or G (Table 1), with a significant bias in favor of C over G. For example, seven of eight serines are encoded by TCC, 31 alanines by GCC, and 9 by GCG out of 58, 13 of 15 valines by GTC (7) or GTG (6), and seven of nine glycines by GGC.

The AAG codon for Lys is overwhelmingly used (57 of 63 codons) in preference to AAA. It is interesting in this connection that the Arg content of trout H1 as well as of other H1 histones should be so extremely low in view of the fact that in the 63 Lys positions, mutation of either the first codon position from A to C or the second position from A to G would give rise to an Arg codon. This would imply that there is strong selection in favor of Lys at the protein level in H1 and suggests that although Lys and Arg are chemically similar in charge and length of side chain, the two residues must be quite distinct in function in their interactions with DNA in chromatin.

Repeating Peptide Motifs in Trout H1

There is, particularly in the C-terminal region of the trout H1 gene, a repeating hexapeptide motif, Ala.Ala.Ala.Lys.Lys.Pro, that appears five times with strong conservation (residues 120–125, 126–131, 134–139, 140–145, and 170–175) and once more (residues 201–206) with somewhat more divergence (see Table 2). In addition, examples of the more

Table 2. Repeating hexapeptide and phosphorylation-site motifs in the trout H1 coding region

Repeating Hexapeptide Motif in H1

Highly conserved motifs

C-terminal half

120	125
Val . Ala . Ala . Lys . Lys . Pro	
GTG.GCC.GCC.AAG.AAG.CCC	
126	131
Ala . Ala . Ala . Lys . Lys . Pro	
GCC.GCC.GCC.AAG.AAG.CCC	
134	139
Val . Ala . Ala . Lys . Lys . Ala	
GTA.GCA.GCC.AAG.AAG.GCC	
140	145
Val . Ala . Ala . Lys . Lys . Ser	
GTG.GCC.GCA.AAG.AAG.TCC	
170	175
Ala . Ala . Ala . Lys . Lys . Ala	
GCA.GCG.GCC.AAG.AAA.GCG	

More diverged motifs

C-terminal half

201	206
Ala . Ala . Pro . Lys . Lys . Lys	
GCA.GCC.CCC.AAG.AAG.AAG	

N-terminal half

20	25
Ala . Ala . Ala . Lys . Pro . Lys	
GCA.GCA.GCC.AAG.CCC.AAG	
51	56
Ala . Ala . Leu . Lys . Lys . Ser	
GCC.GCG.CTC.AAG.AAG.TCT	
103	108
Val . Glu . Ala . Lys . Lys . Pro	
GTC.GAG.GCA.AAG.AAG.CCC	
107	112
Lys . Pro . Ala . Lys . Lys . Ala	
AAG.CCC.GCC.AAG.AAA.GCC	

Repeating Pentapeptide Phosphorylation Site Motifs

144	148
Lys . Ser . Pro . Lys . Lys	
AAG.TCC.CCC.AAG.AAG	
160	164
Lys . Ser . Pro . Lys . Lys	
AAG.AGC.CCA.AAG.AAG	
177	181
Lys . Ser . Pro . Lys . Lys	
AAG.AGC.CCC.AAG.AAG	

Both the amino acids and the corresponding segments of DNA sequence are given; the numbers above the amino acids indicate their positions in the coding sequence

diverged motif appear four times in the N-terminal half of the molecule (residues 20–25, 51–56, 103–108, and 107–112). The highly conserved peptide motif is coded for by an 18-bp DNA repeating sequence, Gtg GCc GCc AAG AAg cCc, in which 11 of the 18 positions are totally conserved (upper-case letters) and five of the remainder are in third codon positions and do not affect the coding sense. The pair of Lys residues is coded for by AAG.AAG except in positions 170–175, where the second Lys is coded for by AAA.

The most probable explanation for the generation of an array of tandemly repeated motifs in H1 is a process of repeated partial duplication (Macleod et al. 1977; Von Holt et al. 1979). In the set of more diverged hexapeptide motifs in the N-terminal half of the molecule (Table 2), new residues appear, e.g., Leu in motif 51–56 and Glu in motif 103–108; in motif 20–25, the two C-terminal amino acids appear in reverse order.

There is some evidence of a gradient of base substitutions in the molecule: The two central and adjacent motifs 120–125 and 126–131 differ in only two bases, whereas the motifs at the extreme ends of the molecule, 201–206 and 20–25, differ by six and nine bases, respectively. This could argue that the two central, adjacent motifs were the ones most recently duplicated during the evolution of histone H1.

Repeating Pentapeptide Phosphorylation Motif

A repeating pentapeptide motif of amino acid sequence Lys.Ser.Pro.Lys.Lys occurs three times in the C-terminal region (144–148, 160–164, and 177–181), without any amino acid variation. It has been shown that the Ser residues at these three sites are phosphorylated during the rapid mitotic divisions characteristic of spermatogonial stem cells (Louie and Dixon 1973; Dixon et al. 1975; Macleod et al. 1977). As shown in Table 2 the DNA sequences for two of these sequences (160–164 and 177–181) differ by only a single base substitution in the third position of the Pro codon (CCA → CCC), whereas the DNA for the third (144–148) differs from both of these at two positions in the Ser codon and, in fact, switches from one Ser codon series to the other (AGC → TCC). This is consistent with the idea that sequences 160–164 and 177–181 were partially duplicated within the H1 gene more recently than was sequence 144–148.

Homology Comparisons with Other Vertebrate H1 Genes

In Fig. 4, the trout H1 gene sequence is compared with those published for *Xenopus* (XLHW8) by Turner et al. (1983) and for chicken embryonic H1 by Sugarman et al. (1983). Both the DNA and predicted amino acid sequences are compared, and identical sequences are boxed in both cases. There

Fig. 4. Sequence comparison of the H1 coding regions of trout, chicken embryo (Sugarman et al. 1983), and *Xenopus* (XLHW8) (Turner et al. 1983). The three sequences have been aligned for maximum homology. Regions of homology are boxed

is very strong homology in the region of amino acids 22–101, which includes the hydrophobic head previously observed to be strongly conserved (Von Holt et al. 1979). There is much more variability in the N-terminal nose and the highly basic C-terminal tail domains, although the repeating hexapeptide motif as well as the pentapeptide phosphorylation-site motif can be seen in all three species.

Sequences Flanking the Trout H1 Codon Region and Containing Possible Regulatory Signals

5′ Flanking Sequences. H1 genes that have previously been sequenced in the 5′ region include those from two species of sea urchin, *Psammechinus miliaris* (Schaffner et al. 1978) and *Strongylocentrotus purpuratus* (Levy et al. 1982); a chicken embryonic H1 (Sugarman et al. 1983); and two *Xenopus laevis* H1 genes plus one pseudogene (Turner et al. 1983). The mRNA initiation site for the trout H1 gene has been mapped by the S1 nuclease technique (Sharp et al. 1980; States et al. 1982), which clearly indicates that the first A at +1 (−93 in Fig. 2) in the sequence CTTCAATA is the initiation site (data not shown). There is a potential Goldberg–Hogness box of the form TTTAAGG at position −35 (underlined at −127 in Fig. 2), where the second nucleotide of the box, A, has been changed to a T. Table 3 compares the Goldberg–Hogness boxes and mRNA initiation (CAP) sites of the known H1 genes. There is conservation of the sequence TATA in all the genes except that of trout, which shows TTTA; however, this change in the second position from T to A is the only difference between *Xenopus* and *Salmo* over a longer sequence, TATAAGG.

Benoist et al. (1980) have described a model eukaryotic sequence GGC$^{C}_{T}$AATCT around positions −70 to −80 from the CAP site. A variant sequence GGCCAgcCc (−91) with six identities (upper-case letters) out of nine positions is also present 5′ to the

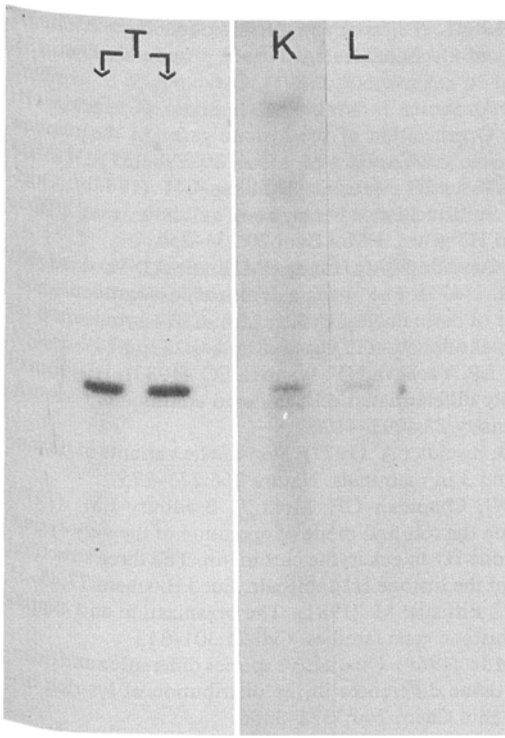

Fig. 5. Northern hybridization of the H1 histone gene to total mRNA from testes (5 μg) (lane T), kidney (10 μg) (lane K), and liver (10 μg) (lane L). The blot was hybridized to nick-translated Nae I–Hinc II fragment (~2 × 10^7 cpm/μg), washed, and exposed for 8 h (lane T), 35 h (lane K), and 150 h (lane L)

Table 3. Comparison of putative regulatory sequences in the 5′ flanking regions of H1 genes

P. miliaris	--------	-28 TATATGG	-17-CGTCACTT-34-ATG
S. purpuratus	-36 CAACAATG	-28 TATATTG	-17-GCTTATTC-34-ATG
Gallus domesticus	-56 CACCAATC	-34 TATAAAT	-23-GCCCAGTG-34-ATG
Xenopus laevis	-91 GGCCAATC	-37 TATAAGG	-26-TTTTAGTT-24-ATG
Salmo gairdnerii	-89 GGCCAGCC	-34 TTTAAGG	-23-CTTCAATA-86-ATG
Consensus	---CA---	T-TA---	----A-T-

The data for *P. miliaris* are from Hentschel and Birnstiel (1981), those for *S. purpuratus* from Maxson et al. (1983), those for *Xenopus laevis* from Turner et al. (1983), those for chicken embryo from Sugarman et al. (1983), and those for *Salmo gairdnerii* from the present paper. The number above the sequence indicates the position upstream from the CAP site (indicated by a dot above the capped A). The degree of consensus is indicated at bottom, with upper-case letters denoting complete conservation

trout H1 gene. An exact nine-residue repeat of this sequence is also found in the trout H1 gene immediately downstream of the TTTAA box.

There is little conservation of sequence at the mRNA initiation (CAP) site; as may be seen in Table 3, only the capped A and a T at position +3 are totally conserved. However, the two residues at −1 and −2 are always pyrimidines, for a consensus of PyPyANT.

Other motifs that have been seen in 5′ flanking regions of H1 and other histone genes, such as the sequences GATCC (Hentschel and Birnstiel 1981) and GCGGGG (Sugarman et al. 1983), are not present in the trout H1 5′ flanking region. Their possible universal functional significance must therefore be questioned.

3′ Flanking Region. In the 3′ flanking region is found a 25-bp conserved sequence (at positions +47 to +72) that includes the hyphenated dyad found in all eukaryotic histone genes so far sequenced and has an important function in histone mRNA processing, as well as the sequence ACCCA (+68) known to be the termination site of histone genes. Twelve residues further downstream is another sequence, AAAAGCGC, which contains an A-rich motif, AAAAGG (C) common to all five trout histone genes (Connor et al. 1984) and was previously described in the sea urchin histone genes. The sequence (C)TT TCCA(C) is repeated four times, at positions +41, +76, +96, and +101. This sequence is very close to the complement on the opposite DNA strand of the sequence GTGGTTTG AAA G, which has been described as a transcription-enhancer element in several systems (Weiher et al. 1983).

Is This H1 Gene Transcribed?

Total RNA prepared from trout testes, kidney and liver was electrophoresed in denaturing, methylmercury gels and blotted to Pall Biodyne membranes, and the blots were probed with a nick-translated Nae I–Hinc II fragment (Fig. 1) of λTH2 containing most of the coding region of the trout H1 gene. Results are shown in Fig. 5. For testis and liver, only a single, sharp H1 mRNA band hybridized to the probe. For kidney, which is also a major hemopoietic tissue in fishes (Catton 1951), there was, in addition to the same major band as in the other tissues, a faint, faster-running band, as well as some streaking that might indicate the presence of other H1 mRNA species in low amounts. The concentration of H1 mRNA is many-fold higher in trout testis than in kidney, and only prolonged exposure revealed the liver H1 band.

The finding of a single predominant H1 mRNA band in all three tissues was somewhat unexpected in view of previous observations of multiple H1

protein species in a variety of mammalian and fish tissues (Kinkade and Cole 1966; Bustin and Cole 1968; Kinkade 1969; Seyedin and Cole 1981; Brown and Goodwin 1983; Cole 1984). Moreover, the discrepancy between the coding region predicted from the present H1 DNA sequence and that of the trout testis H1 sequenced at the protein level by Macleod et al. (1977) would predict that more than one mRNA species would be found in trout testis. The fact that only one H1 mRNA band is seen, together with the fact that at the 5′ end, S1 protection mapping revealed only a single mRNA initiation site, requires further explanation and study. One possibility might be that the differences between the H1 proteins observed by ion-exchange chromatography might be so minor as to have no discernible effect on the overall size of the cognate mRNAs, which, in addition, must possess identical CAP sites. However, this would mean that the single H1 mRNA band would not be sequentially pure but would be composed of a family of mRNAs of similar but not identical sequence. A second possibility is that post-synthetic proteolytic processing or enzymatic modification of H1 proteins might give rise to the observed H1 heterogeneity. Further work is under way to distinguish between these possible explanations of the observed heterogeneity of H1 proteins.

Acknowledgments. This research was supported by a Medical Research Council of Canada operating grant to G. H. Dixon, by the Alberta Heritage Foundation for Medical Research, and by the Juan March Foundation of Spain.

The authors would also like to thank Dr. Stephen Krawetz for his help in obtaining H1 gene sequences through Intelli-Genetics and to Marnie Cudmore for expertly typing the manuscript.

References

Ajiro K, Borun TW, Shulman SD, McFadden GM, Cohen LH (1981) Comparison of the structures of human histones 1A and 1B and their intramolecular phosphorylation sites during the HeLa S-3 cell cycle. Biochemistry 20:1454–1464.

Bailey JM, Davidson N (1976) Methylmercury as a reversible denaturing agent for agarose gel electrophoresis. Anal Biochem 70:75–85

Benoist C, O'Hare K, Breathnach R, Chambon P (1980) The ovalbumin gene-sequence of putative control regions. Nucleic Acids Res 8:127–142

Birnboim HC, Doly J (1979) A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Res 7:1513–1523

Brown E, Goodwin GH (1983) Isolation from trout liver of a methionine-containing H1 subfraction, THI^{met}. Eur J Biochem 134:255–259

Bustin M, Cole RD (1968) Species and organ specificity in very lysine-rich histones. J Biol Chem 243:4500–4505

Catton WT (1951) Blood cell formation in certain teleost fishes. Blood 6:39–60

Cohen LH, Newrock KM, Zweidler A (1975) Stage-specific switches in histone synthesis during embryogenesis of the sea urchin. Science 190:994–997

Cole RD (1984) A minireview of microheterogeneity in H1 histone and its possible significance. Anal Biochem 136: 24–30

Connor W, Mezquita J, Winkfein RJ, States JC, Dixon GH (1984a) Organization of the histone genes in the rainbow trout (*Salmo gairdnerii*). J Mol Evol 20:227–235

Connor W, States JC, Mezquita J, Dixon GH (1984b) Organization and nucleotide sequence of rainbow trout histone H2A and H3 genes. J Mol Evol 20:236–250

Dixon GH, Candido EPM, Honda BM, Louie AJ, Macleod AR, Sung MT (1975) The biological roles of post-synthetic modifications of basic nuclear proteins. In: CIBA symposium on structure and function of chromatin, vol 28, pp 229–258

Djondjurov LP, Yacheva NY, Ivanova EC (1983) Histones of terminally differentiated cells undergo continuous turnover. Biochemistry 22:4095–4102

Franklin SG, Zweidler A (1977) Non-allelic variants of histone 2a, 2b and 3 in mammals. Nature 266:273–275

Hartman PG, Chapman GE, Moss T, Bradbury EM (1977) Studies on the role and mode of operation of the very-lysine-rich histone H1 in eukaryote chromatin. The three structural regions of the histone H1 molecule. Eur J Biochem 77:45–51

Hentschel C, Birnstiel M (1981) The organization and expression of histone gene families. Cell 25:301–313

Kinkade JM Jr (1969) Qualitative species differences and quantitative tissue differences in the distribution of lys-rich histones. J Biol Chem 244:3375–3386

Kinkade JM Jr, Cole RD (1966) The resolution of four lysine-rich histones derived from calf thymus. J Biol Chem 241: 5790–5797

Klingholz R, Stratling WH (1982) Reassociation of histone H1 to H1-depleted polynucleosomes. J Biol Chem 257:13101–13107

Lennox RW (1984) Differences in evolutionary stability among mammalian H1 subtypes. J Biol Chem 259:669–672

Lennox RW, Cohen LH (1983) The histone H1 complements of dividing and non-dividing cells of the mouse. J Biol Chem 258:262–268

Louise AJ, Dixon GH (1973) Kinetics of phosphorylation and dephosphorylation of testis histones and their possible role in determining chromosomal structure. Nature (New Biol) 243:164–168

Macleod AR, Wong NCW, Dixon GH (1977) The amino-acid sequence of trout testis histone H1. Eur J Biochem 78:281–291

Maniatis T, Jeffrey A, Kleid D (1975) Nucleotide sequence of the rightward operator of phage λ. Proc Natl Acad Sci USA 72:1184–1189

Maxam A, Gilbert W (1980) Sequencing end-labelled DNA with base specific chemical cleavages. Methods Enzymol 65: 499–560

Maxson R, Cohn R, Kedes L, Mohun T (1983) Expression and organization of histone genes. Annu Rev Genet 17:239–277

McGhee JD, Felsenfeld G (1980) Nucleosome structure. Annu Rev Biochem 49:1115–1156

Miki BLA, Neelin JM (1975) The histones of rainbow trout erythrocytes include an erythrocyte-specific histone. Can J Biochem 53:1158–1169

Panyim S, Chalkley R (1969) A new histone found only in mammalian tissues with little cell division. Biochem Biophys Res Commun 37:1042–1049

Risley MS, Eckhardt RA (1981) H1 histone variants in *Xenopus laevis*. Dev Biol 84:79–87

Schaffner W, Kunz G, Dactwyler H, Telford J, Smith HO, Birnstiel ML (1978) Genes and spacers of cloned sea urchin histone DNA analyzed by sequencing. Cell 4:655–671

Seyedin SM, Cole RD (1981) H1 histones of trout. J Biol Chem 256:442–444

Seyedin SM, Kistler WS (1980) Isolation and characterization of rat testis H1t. J Biol Chem 255:5949–5954

Seyedin SM, Cole RD, Kistler WS (1981) H1 histones from mammalian testes. Exp Cell Res 136:399–405

Srebreva L, Zlatanova J (1983) Occurrence of histone $H1_0$-related protein fractions in trout liver. Biochim Biophys Acta 740:163–168

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

Stephenson EC, Erba HP, Gall JG (1981) Characterization of a cloned gene cluster of the newt, *Notophthalmus viridescens*. Nucleic Acids Res 9:2281–2295

Sugarman BJ, Dodgson JB, Engel JD (1983) Genomic organization, DNA sequence and expression of chicken embryonic histone genes. J Biol Chem 258:9005–9016

Turner PC, Aldridge TC, Woodland HR, Old RW (1983) Nucleotide sequences of H1 histone genes from *Xenopus laevis*—a recently diverged pair of H1 genes and an unusual H1 pseudogene. Nucleic Acids Res 11:4093–4107

Von Holt C, Strickland WN, Brandt WF, Strickland MS (1979) More histone structures. FEBS Lett 100:201–218

Weiher H, Konig M, Gruss P (1983) Multiple point mutations affecting the simian virus-40 enhancer. Science 219:626–631

Received August 28, 1984/Revised October 30, 1984