Volume 15 Number 7 1987 Nucleic Acids Research

Mouse histone H2A and H2B genes: four functional genes and a pseudogene undergoing gene conversion with a closely linked functional gene

Ta-Jen Liu, Landie Liu and William F.Marzluff

Department of Chemistry, Florida State University, Tallahassee, FL 32306, USA

Received November 19, 1986; Revised and Accepted March 11, 1987 Accession no. Y00117

ABSTRACT

The sequence of five mouse histone genes, two H2a and three H2b genes on chromosome 13 has been determined. The three H2b genes all code for different proteins, each differing in two amino acids from the others. The H2b specific elements present 5' to H2b genes from other species are present in all three mouse H2b genes. All three H2b genes are expressed in the same relative amounts in three different mouse cell lines and fetal mice. The H2b gene with the H2b specific sequence closest to the TATAA sequence is expressed in the highest amount. One of the H2a genes lacks the first 9 amino acids, the promoter region, the last 3 amino acids and contains an altered 3' end sequence. Despite these multiple defects, there is only one nucleotide change between the two H2a genes from codon 9 to 126. This indicates that a recent gene conversion has occurred between these two genes. The similarity of the nucleotide sequences in the coding regions of mouse histone genes is probably due to gene conversion events targeted precisely at the coding region.

INTRODUCTION

Histone proteins have been among the most highly conserved proteins during evolution. The histones are a group of five classes of closely related proteins whose expression is coordinately regulated. In all higher eucaryotes the genes coding for the five histones are tightly clustered (1). The organization of these genes is unique in higher eucaryotes, in that genes coding for independent proteins have remained clustered over long evolutionary times. The precise gene organization varies from organism to organism. In some cases they are present in tandemly repeated units, each unit containing a single copy of a gene for each of the five histones. In birds and mammals, the histone genes are present in randomly organized clusters, containing several distinct, independent, genes for each histone (2-5). In lower eucaryotes the histone genes have been found in pairs, with H3 and H4 genes linked (6,7) and H2a and H2b genes linked (8).

There are at least two clusters of histone genes found on chromosomes 3 and 13 in the mouse (9) and on chromosomes 1 and 6 in humans (10). Both

these clusters are regulated coordinately in the mouse in some conditions (2) but may also be regulated independently in other conditions (11,12). The clusters contain multiple copies of genes for a given class of histone proteins. The genes for these proteins are very similar in the coding region but diverge significantly in the flanking region, allowing the products of individual genes to be detected by an S1 nuclease assay (2, 13). Not all of the genes for a particular class of histone protein code for proteins with the same amino acid sequence. There are a number of known protein variants which differ in a small number of amino acids (14, 15). In addition as reported previously (13) and extended here, there are a number of protein sequence variants deduced from the gene sequences which have not been previously reported.

We report here the sequence of three H2b genes present on mouse chromosome 13. Two of these genes are very tightly linked (<300 bases away) to H2a genes. All of these genes are expressed in all mouse cell lines tested in the same relative proportion. One of the H2a genes is a pseudogene with multiple defects at both the 5' and 3' end of the gene, although 85% of the protein coding region has remained intact and in frame. This sequence includes an amino acid variant present in both H2a sequences which is not present in other H2a genes. This suggests that gene conversion is occurring which primarily affects only the coding region and that the pseudogene has recently undergone a conversion event. The closely linked H2b genes have also undergone a gene conversion event which may have also included the 3' untranslated region.

MATERIALS AND METHODS

DNA Sequencing

The phage containing the histone genes, MM221 and MM291, have been described previously (2). The genes were sequenced either by the method of Maxam and Gilbert (16) or by the method of Sanger (17). For sequencing with dideoxynucleotides the DNA fragments were cloned into mp8 and mp9 or into pUC118 or pUC119. The sequence of both strands of most (>90%) of the sequences presented here were obtained. In the cases where only one strand was sequenced, each sequence was determined at least twice. The DNA sequences were analyzed using the computer programs of Dr. Jim Pustell (18).

S1 nuclease mapping

Mouse myeloma cells, L cells and 3T6 fibroblasts were cultured and total RNA was prepared by extraction with phenol as previously described (2). S1 nuclease mapping was performed exactly as previously described (2). To deter-

mine the 5' ends of the mRNAs the DNA fragments protected from S1 digestion were analyzed in parallel with the sequence of the same end-labeled fragment sequenced by the method of Maxam and Gilbert (16).

RESULTS AND DISCUSSION

Histone genes are present in multiple copies in the mouse and are dispersed onto at least two chromosomes. The majority of the genes are present on mouse chromosome 13. We previously reported the organization and expression of the H3, H2a and H2b genes on two phage derived from mouse chromosome 13 (2). The structure of these phage is shown in Fig. 1A. MM221 contains two H3 genes, an H2b gene closely linked to the H3.2 gene and part of an H2a gene (13). MM291 contains two H2a-H2b pairs separated by an H3 gene (2). These are arbitrarily designated 291A and 291B (2, Fig. 1A). The H2a and H2b genes are less than 300 bases apart with their 5' ends juxtaposed and hence they are transcribed from opposite strands of the DNA. One of the H2a genes, H2a.291B, was classified as a pseudogene with a defect at the 5' end on the basis of S1 nuclease mapping (2).

Sequences of the H2a and H2b Genes

The restriction map of the genes from MM291 is shown in Fig. 1B. The intergenic distance between the H2a and H2b genes is nearly identical. The restriction maps of the H2b and H2a genes are identical in the coding region

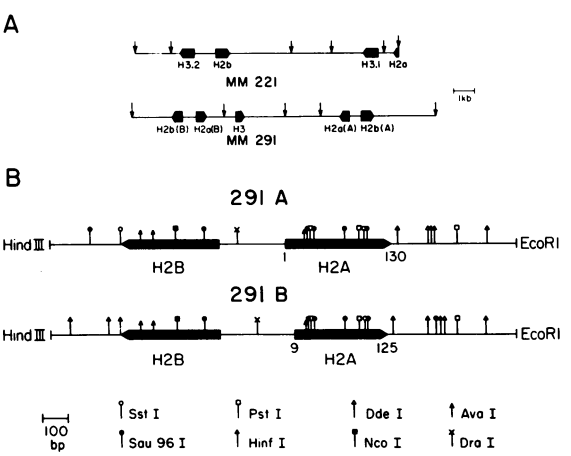
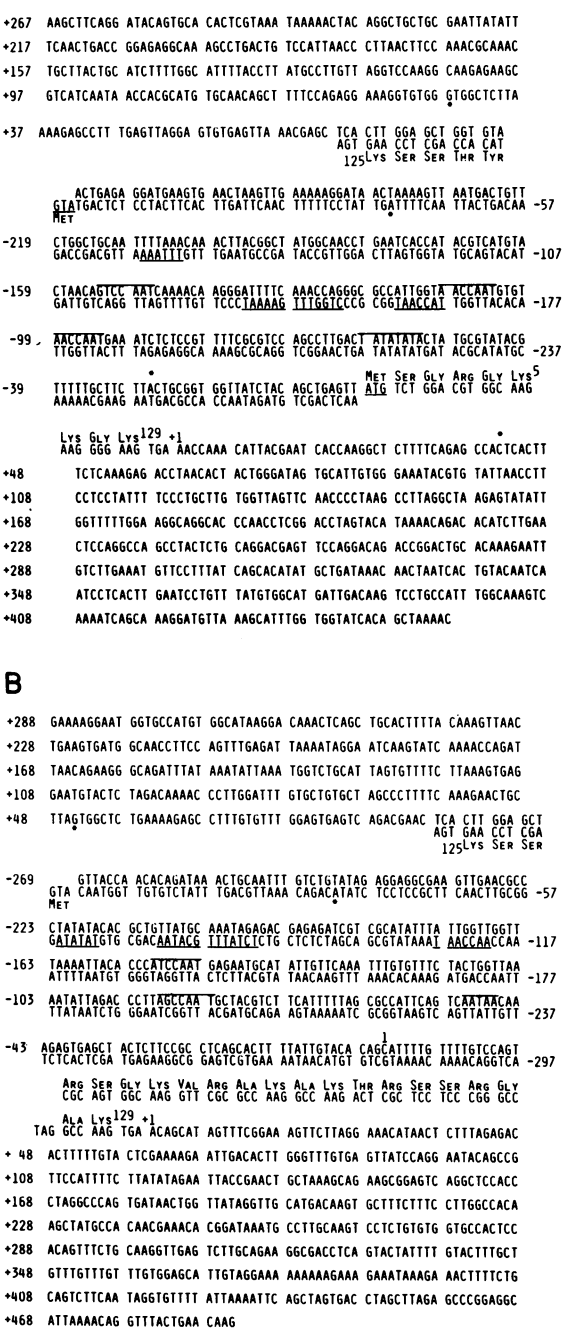


FIG. 1. Mouse histone clusters MM221 and MM291. A. The structure of the two recombinant phage with the position and direction of transcription of the histone genes is shown. B. The restriction map of the two H2a-H2b gene pairs from MM291 is shown.


A
+267 AAGCTTCAGG ATACAGTGCA CACTCGTAAA TAAAAACTAC AGGCTGCTGC GAATTATATT
+217 TCAACTGACC GGAGAGGCAA AGCCTGACTG TCCATTAACC CTTAACTTCC AAACGCAAAC
+157 TGCTTACTGC ATCTTTTGGC ATTTTACCTT ATGCCTTGTT AGGTCCAAGG CAAGAGAAGC
+97 GTCATCAATA ACCACGCATG TGCAACAGCT TTTCCAGAGG AAAGGTGTGG GTGGCTCTTA
+37 AAAGAGCCTT TGAGTTAGGA GTGTGAGTTA AACGAGC TCA CTT GGA GCT GGT GTA
AGT GAA CCT CGA CCA CAT
125Lys Ser Ser Thr Tyr
ACTGAGA GGATGAAGTG AACTAAGTTG AAAAAGGATA ACTAAAAGTT AATGACTGTT
GTATGACTCT CCTACTTCAC TTGATTCAAC TTTTTCCTAT TGATTTTCAA TTACTGACAA -57
Met
-219 CTGGCTGCAA TTTTAAACAA ACTTACGGCT ATGGCAACCT GAATCACCAT ACGTCATGTA
GACCGACGTT AAAATTTGTT TGAATGCCGA TACCGTTGGA CTTAGTGGTA TGCAGTACAT -107
-159 CTAACAGTCC AATCAAAACA AGGGATTTTC AAACCAGGGC GCCATTGGTA ACCAATGTGT
GATTGTCAGG TTAGTTTTGT TCCCTAAAAG TTTGGTCCCG CGGTAACCAT TGGTTACACA -177
-99 AACCAATGAA ATCTCTCCGT TTTCGCGTCC AGCCTTGACT ATATATACTA TGCGTATACG
TTGGTTACTT TAGAGAGGCA AAAGCGCAGG TCGGAACTGA TATATATGAT ACGCATATGC -237
Met Ser Gly Arg Gly Lys5
-39 TTTTTGCTTC TTACTGCGGT GGTTATCTAC AGCTGAGTT ATG TCT GGA CGT GGC AAG
AAAAACGAAG AATGACGCCA CCAATAGATG TCGACTCAA
Lys Gly Lys129 +1
AAG GGG AAG TGA AACCAAA CATTACGAAT CACCAAGGCT CTTTTCAGAG CCACTCACTT
+48 TCTCAAAGAG ACCTAACACT ACTGGGATAG TGCATTGTGG GAAATACGTG TATTAACCTT
+108 CCTCCTATTT TCCCTGCTTG TGGTTAGTTC AACCCCTAAG CCTTAGGCTA AGAGTATATT
+168 GGTTTTTGGA AGGCAGGCAC CCAACCTCGG ACCTAGTACA TAAAACAGAC ACATCTTGAA
+228 CTCCAGGCCA GCCTACTCTG CAGGACGAGT TCCAGGACAG ACCGGACTGC ACAAAGAATT
+288 GTCTTGAAAT GTTCCTTTAT CAGCACATAT GCTGATAAAC AACTAATCAC TGTACAATCA
+348 ATCCTCACTT GAATCCTGTT TATGTGGCAT GATTGACAAG TCCTGCCATT TGGCAAAGTC
+408 AAAATCAGCA AAGGATGTTA AAGCATTTGG TGGTATCACA GCTAAAAC
B
+288 GAAAAGGAAT GGTGCCATGT GGCATAAGGA CAAACTCAGC TGCACTTTTA CAAAGTTAAC
+228 TGAAGTGATG GCAACCTTCC AGTTTGAGAT TAAAATAGGA ATCAAGTATC AAAACCAGAT
+168 TAACAGAAGG GCAGATTTAT AAATATTAAA TGGTCTGCAT TAGTGTTTTC TTAAAGTGAG
+108 GAATGTACTC TAGACAAAAC CCTTGGATTT GTGCTGTGCT AGCCCTTTTC AAAGAACTGC
+48 TTAGTGGCTC TGAAAAGAGC CTTTGTGTTT GGAGTGAGTC AGACGAAC TCA CTT GGA GCT
AGT GAA CCT CGA
125Lys Ser Ser
-269 GTTACCA ACACAGATAA ACTGCAATTT GTCTGTATAG AGGAGGCGAA GTTGAACGCC
GTA CAATGGT TGTGTCTATT TGACGTTAAA CAGACATATC TCCTCCGCTT CAACTTGCGG -57
Met
-223 CTATATACAC GCTGTTATGC AAATAGAGAC GAGAGATCGT CGCATATTTA TTGGTTGGTT
GATATATGTG CGACAATACG TTTATCTCTG CTCTCTAGCA GCGTATAAAT AACCAACCAA -117
-163 TAAAATTACA CCCATCCAAT GAGAATGCAT ATTGTTCAAA TTTGTGTTTC TACTGGTTAA
ATTTTAATGT GGGTAGGTTA CTCTTACGTA TAACAAGTTT AAACACAAAG ATGACCAATT -177
-103 AATATTAGAC CCTTAGCCAA TGCTACGTCT TCATTTTTAG CGCCATTCAG TCAATAACAA
TTATAATCTG GGAATCGGTT ACGATGCAGA AGTAAAAATC GCGGTAAGTC AGTTATTGTT -237
1
-43 AGAGTGAGCT ACTCTTCCGC CTCAGCACTT TTATTGTACA CAGCATTTTG TTTTGTCCAGT
TCTCACTCGA TGAGAAGGCG GAGTCGTGAA AATAACATGT GTCGTAAAAC AAAACAGGTCA -297
Arg Ser Gly Lys Val Arg Ala Lys Ala Lys Thr Arg Ser Ser Arg Gly
CGC AGT GGC AAG GTT CGC GCC AAG GCC AAG ACT CGC TCC TCC CGG GCC
Ala Lys129 +1
TAG GCC AAG TGA ACAGCAT AGTTTCGGAA AGTTCTTAGG AAACATAACT CTTTAGAGAC
+ 48 ACTTTTTGTA CTCGAAAAGA ATTGACACTT GGGTTTGTGA GTTATCCAGG AATACAGCCG
+108 TTCCATTTTC TTATATAGAA TTACCGAACT GCTAAAGCAG AAGCGGAGTC AGGCTCCACC
+168 CTAGGCCCAG TGATAACTGG TTATAGGTTG CATGACAAGT GCTTTCTTTC CTTGGCCACA
+228 AGCTATGCCA CAACGAAACA CGGATAAATG CCTTGCAAGT CCTCTGTGTG GTGCCACTCC
+288 ACAGTTTCTG CAAGGTTGAG TCTTGCAGAA GGCGACCTCA GTACTATTTT GTACTTTGCT
+348 GTTTGTTTGT TTGTGGAGCA TTGTAGGAAA AAAAAAGAAA GAAATAAAGA AACTTTTCTG
+408 CAGTCTTCAA TAGGTGTTTT ATTAAAATTC AGCTAGTGAC CTAGCTTAGA GCCCGGAGGC
+468 ATTAAAACAG GTTTACTGAA CAAG

C

```
-1135          TTCTACAAAGAAATAGGAAATGCAGAGGTTAGGTTCTGGGTTCCTTTATA
-1085  GAAGGAACATGCTCCTGATTGGTTGAAATCTTCAAATGCCTATGATGATGTATGCGGAAG
-1025  TACTTGGTAATTGGATAACAGGTCTCAACTTCAAATTTGTGACCAATAGAAATCTGGTTT
-965   GTTGCATAGCTTAGCTCCTCTTTGTTAACTATTGCTTCCTCAGCCACATTGTTACCATTA
-905   AGTACCGTAATTGTTACATTTTCGTCAAAACATTTTAATAAGAGATACTATGTAGATTTT
-845   GGCGATACCACAGCGGCCCCAAGGATTACTTGTTGAACTTGGTTACCCGAACTGAGTTTT
-785   TTCTTTTTCCTTACTTTATATTTAAGGAGATGTAAGGGAGCAGTTTACTATCCGGACCAA
-725   AATAAAACCGAACAGGCACACCCAACTTTGTACATCTTTTAATGACTTTTACAAATGTCT
-665   TAGCATGAGTAACCCTAATTATCTTCAGAAAAACACTAGCATGTTTTCTAGTAACTATGT
-605   ACAGGATTCATTGAAATTGCACAACTGTTGTAGCGTGGCAGTACCTCTAAGGGTTCTTGG
-545   GCCAAATGAAAAGCATGCCTACTTTAGCCGGACTGTACAATAATTATTTCTGGAAGCTCT
-485   AGGTGAATAGTATGTGATGTCTCCCATTCTTAATCTTTCTTCATAGACAATTTCAAAACA
-425   GGAAACCAACTTTACATGTAGCTAGAATTTATATATGAATTTTGTATCTAGGGAGAGATT
-365   TTTATTGATACTACTAACAAAATAACTTTTAAAAGAAAAATGGCTTTGTACTTTAGTACA
-305   GTATCAAACCAACTTTTTCCTAAAAACTATTTCGTCTAGTTTTCATAGAGTCTATTGGCT
-245   AAACTTGGCCAATCAGAAGAAAGAAACGAGTCTTCATTTGAATAGAGCCCTGTAAAGTTC
-185   GGTCCGCTTCCTCCGTGATACTTACGCAACTAACCACTGAGCGAATATGCTTCCTTGATG
-125   GACAGTTAGTGCTTGACGTTTGCAGACTCTCTGACAAGGACAGCCACCGCTTTATTTAAA
-65    GAGCAGGAAAGGAACGGAACAGTTCAATATCTCTTTTCTTGGCCTACCTTCATTCTCTGTTCACT
```

Fig. 2. Sequence of the mouse histone H2a and H2b genes. A. Sequence of cluster MM291A. B. Sequence of cluster MM291B. The TATAA box, CCAAT box and the H2b 5' consensus sequence are marked. The underlined sequences are 5' to the H2a gene and the overlined sequences are 5' to the H2b genes. The * marks the first and last nucleotides of the mRNAs. The sequence of the complete coding region is not shown here since these are compared in Fig. 3. The numbers refer to the distance from the start of translation, except for the H2a.291B pseudogene where they indicate the distance from where the ATG codon would have been. C. The sequence of the intergenic region between the H3.2 and H2b genes from MM221. The strand shown is the strand coding for the H2b gene. The underlined sequences are the consensus sequences 5' to the H2b gene and the overlined sequences the consensus sequences 5' to the H3.2 gene. The * indicates the first nucleotide of the mRNAs. The numbers refer to the distance from the start of translation of the H2b genes. The first nucleotide of the sequence is the nucleotide adjacent to the ATG codon of the H3.2 gene.

but diverge significantly in the flanking regions. In particular the intergenic regions have different restriction maps. We have previously reported the sequence of the complete H3.2 gene (13, 19) and the H2b gene on MM221 (13). Here we report in addition the complete sequence between these two genes. Since all these genes are on the same chromosome and are coordinately controlled, comparing these sequences may give some insight into how the genes have evolved and what common sequences important in control of expression may have been retained.

Figure 2A and 2B shows the sequence of the H2A-H2B gene clusters on MM291. The double stranded sequence is shown for the intergenic region. The TATAA sequences and the CCAAT sequences are underlined and the first and last

A

```
                            Ala                     10                              Val
H2B-291A    ATG CCT GAG CCC GCC AAG TCC GCT CCC GCC CCG AAG AAG GGC TCC AAG AAG GCC GTC ACC
H2B-291B                  A A G
H2B-221                                       T                                     C G
                            Thr                                                     Leu

            20                                      30
H2B-291A    AAG GCC CAG AAG AAG GAC GGC AAG AAG CGC AAG CGC AGC CGC AAG GAG AGC TAC TCG GTG
H2B-291B                          T
H2B-221

            40                                      50
H2B-291A    TAC GTG TAC AAG GTG CTG AAG CAA GTG CAC CCC GAC ACC GGC ATC TCC TCC AAG GCC ATG
H2B-291B                          A
H2B-221

            60                                      70                Ser               •
H2B-291A    GGC ATC ATG AAC TCG TTC GTG AAC GAC ATC TTC GAG CGC ATC GCG AGC GAG GCT TCC CGC
H2B-291B                                                          A G
H2B-221                                                             G A     G
                                                                    Gly

            80                                      90
H2B-291A    CTG GCG CAT TAC AAC AAG CGC TCG ACC ATC ACG TCC CGG GAG ATC CAG ACG GCC GTG CGC
H2B-291B
H2B-221       A

            100                                     110
H2B-291A    CTG CTG CTG CCC GGG GAG CTG GCC AAG CAC GCG GTG TCG GAG GGC ACC AAG GCA GTC ACC
H2B-291B                                                                      G
H2B-221                                           C                           T

            120
H2B-291A    AAG TAC ACC AGC TCC AAG TGA
H2B-291B
H2B-221
```

B

```
                                                    10                      Thr
H2A-291A    ATG TCT GGA CGT GGC AAG CAA GGA GGC AAG GCC CGC GCC AAG GCC AAG ACG CGC TCC TCC
H2A-291B    CAT  T   TT TTG TC   GT  GC A T          TT                       T

            20                                      30
H2A-291A    CGG GCC GGC CTG CAG TTC CCC GTG GGC CGC GTG CAC CGG CTG CTC CGC AAG GGC AAC TAC
H2A-291B

            40
            Ser                                     50  Leu
H2A-291A    TCG GAG CGC GTG GGC GCC GGC GCC CCG GTG TAC CTG GCG GCC GTG CTG GAG TAC CTG ACG
H2A-291B

            60                                      70
H2A-291A    GCC GAG ATC CTG GAG CTG GCG GGC AAC GCG GCC CGC GAC AAC AAG AAG ACG CGC ATC ATC
H2A-291B

            80                                      90                                  Arg
H2A-291A    CCG CGC CAC CTG CAG CTG GCC ATC CGC AAC GAC GAG GAG CTC AAC AAG CTG CTG GGC CGC
H2A-291B
H2A-221     ----------------------------------------------------- G             T

            100                                     110
H2A-291A    GTG ACC ATC GCG CAG GGC GGC GTC CTG CCC AAC ATC CAG GCC GTG CTG CTG CCC AAG AAG
H2A-291B
H2A-221

            120                     Ala
H2A-291A    ACC GAG AGC CAC CAC AAG GCC AAG GGG AAG TGA
H2A-291B                        T               --- ---
H2A-221                       T     C         A      A
                                    Pro
```

Fig. 3. Comparison of the Coding Region Sequences. A. The coding region of three H2b genes are compared. The sequence of the H2b.221 gene is taken from ref. 13. Only those nucleotides that differ from the sequence of the H2b.291A gene are indicated. Where there are amino acid differences among the genes

these are indicated. B. The coding region sequence of the H2a.291a gene, the H2a.291B pseudogene and the portion of the H2a.221 gene previously sequenced (13) are compared. The amino acids characteristic of the H2a.1 protein variant as well as the serine at position 40 not previously reported in H2a proteins are indicated. The 5' end of the H2a.291B sequence has been aligned with the H2a.291A without any deletions or insertions. The - in the H2a.221 indicates that this sequence has not been determined (13). The - in the H2a.291B sequence indicates a deletion.

base of the mRNA is marked with an *. Also underlined are the regions 5' to the H2b gene which are similar to the sequence found 5' of most H2b genes (20-22). The complete coding region sequence is not shown here since these sequences are compared in detail in Fig. 3.

The complete intergenic region between the H3.2-221 and H2b.221 gene is presented in Figure 2C. Again the putative regions involved in the expression of these two genes are indicated, including CCAAT sequences far 5' to both the H2b and H3.2 genes (see below).

Each of the genes codes for a different protein variant

Each of the 3 H2b genes differ from one another in two amino acids (Table 1, Fig. 3A). As previously reported the H2b.221 gene has a leucine substituted for valine normally found in H2b genes at amino acid 18 (13). The H2b.291A gene has a serine at position 75, giving it the identical sequence of the H2b.2 protein reported by Franklin and Zweidler (14), while the H2b.291B and H2b.221 have a serine at this position typical of the H2b.1 variant (14). The H2b.291B gene has a threonine substituted for alanine normally found in H2b at amino acid 4. Since these genes are each expressed in small amounts and the proteins would not necessarily be resolved from the major H2b proteins by gel electrophoresis, it is not surprising that these minor protein variants have not been previously reported.

The H2a genes also have amino acid changes from the major H2a protein sequence (Fig. 3B, Table 1). The H2a.291A gene has the three amino acids, threonine 18, leucine 51 and arginine 99, characteristic of the most abundant H2a histone, the H2a.1 variant (15). This is probably the same variant encoded by the H2a.221 gene (13). The H2a.291A gene and H2a.291B pseudogene both contain an amino acid change, a serine for alanine at position 40, which has not previously been reported in mammalian H2a proteins (Table 1).

The sequence of the H2a.291B gene confirms our previous conclusion that it is a pseudogene. It contains several defects. It lacks an ATG codon and the first 8 amino acids (Fig. 3B). It has a termination codon at position 126, as well as the original termination codon at codon 129. It also lacks

Table 1. Amino acid changes in the H2a and H2b genes.

H2A GENES					
Amino Acid #	16	40	51	99	126
H2a.291A	Thr	Ser	Leu	Arg	Ala
H2a.291B(12-125)	Thr	Ser	Leu	Arg	---
H2a.221(92-129)	---	---	---	Arg	Pro
H2B GENES					
Amino Acid #	4	18	75		
H2b.291A	Ala	Val	Ser		
H2b.291B	Thr	Val	Gly		
H2b.221	Ala	Leu	Gly		

The amino acids characteristic of the H2a.1 protein variant (15) as well as amino acids that vary among the H2a and H2b genes are shown.

the hairpin loop structure characteristic of histone mRNA 3' ends. It remains almost identical (one base change) from codon 11-124 with the H2a.291A gene. This degree of similarity is not expected since the H2a.291B gene is clearly a pseudogene with multiple defects. There are potential CCAAT and TATAA box sequences present in the 5' flanking region of this pseudogene (Fig. 2B), consistent with the interpretation that this was once a functional gene.

Comparison of the Coding Region Sequences

As previously reported for the mouse histone H3 genes (19) the coding regions of the histone genes are very similar. This similarity is partly due to a highly constrained codon usage. The codon usage in the H2a and H2b genes is essentially identical to that of the mouse histone H3 genes (13, 19, not shown).

The H2b genes differ among each other by 8-14 nucleotides (Table 2), with two nucleotide substitutions between each pair of genes resulting in amino acid changes (Fig. 3A). This is within the range of nucleotide changes observed among the 3 H3 genes on chromosome 13 (19). Like the H3 genes, wherever there is an amino acid change there are multiple (2-3) substitutions in a 4 base region around the replacement substitution. The remaining substitutions

Table 2. Nucleotide changes among H2a and H2b genes.

	Silent Changes	Replacement Changes
291A-291B H2b	8	2
291A-221 H2b	9	2
291B-221 H2b	14	2
291A-291B H2a	1	-
291A-221 H2a	6(out of 112)	1(out of 37)
Ave. H3 on chromosome 13	9	

The number of silent and replacement changes in the H2a and H2b genes is shown. This is compared with the average number of silent changes among the three H3 genes on chromosome 13 (19).

are distributed throughout the gene, although there are so few it is impossible to tell if there is an underlying pattern, as has been observed for the H3 genes (19) and histone genes of other species (23).

The divergence of the H2a.291A and H2a.291B genes at codon 8 could represent the boundary of a recent gene conversion event, since these two genes are more similar in the coding region than any other pair of mouse histone genes we have sequenced (13, 19). The alternative, that there has been a sizable (30 nt) deletion at the 5' end, we regard as less likely since the intergenic distance between the H2a and H2b genes is similar in both gene pairs. At the 3' end of the gene the new TAG codon at codon 125 of the coding region is due to a point mutation. In addition there has been a deletion of 6 nucleotides removing the last two codons leaving the original TGA codon still present.

Comparing the H2a.291 genes with the portion of the functional H2a.221 gene (also probably coding for an H2a.1 protein) previously sequenced (13), there are six nucleotide changes, one of which results in an amino acid change, out of 112 nucleotides (Fig. 3B). This is a number similar to that observed between the H2b and H3 genes on chromosome 13.

5' Flanking regions of the 291A genes

The potential consensus sequences required for expression of the H2a.291A and H2b.291A genes, the TATAA and CCAAT sequences, overlap to some extent due

to the extremely short intergenic distance. The H2b CCAAT sequence is closer to the H2a gene than to the H2a gene and vice versa. Examining the sequences 5' of the H2a.291B pseudogene there are two CCAAT sequences (underlined in Fig. 2B) which are located in appropriate positions to have been part of the promoter for this gene. There is also a possible residual TATAA box (also underlined in Fig. 2B). We have not observed any stable transcripts from this gene by S1 nuclease mapping (2), although we cannot rule out the possibility that it is still transcribed. Since the gene lacks a functional histone 3' end it is possible that any transcripts would not be transported to the cytoplasm or would be very unstable.

Comparison of the 5' flanking region sequences of the H2B genes

Comparing the 5' flanking region sequences of the three H2b genes (for which all the 5' sequence up to the next gene has been determined), one can identify several elements which are found in similar positions and are conserved among all three H2b genes in the 200 bases 5' to the gene. All three genes have two copies of the H2b consensus sequence located 5' to the gene (see below, Fig 4). In addition to these sequences there are two other regions of great similarity among the 3 H2b genes located upstream of the first H2b specific sequence and between the two H2b specific sequences. Each of these genes also contains the typical RNA polymerase II consensus sequences, TATAA and CCAAT boxes, but these differ in the position and the exact sequence This is particularly surprising for the H2b.291A and H2b.291B genes since they presumably arose by a gene duplication and provides further support for the idea that there has been gene conversion limited to the coding region sequences of the histone genes.

Harvey et al. (20) first pointed out the existence of a sequence common to most H2b genes upstream of the TATAA box and this sequence has been further defined by Dixon and coworkers (21) and Wells (22). There is a 13 nucleotide consensus sequence located 3' to a CCAAT sequence. A sequence similar to this sequence is present twice in the H2b.291A and H2b.291B genes (Fig. 4). The most similar sequence to the consensus (22) sequence in each gene is just 3' to a CCAAT sequence. A similar location of this sequence 3' to the CCAAT sequence has been reported for other H2b genes (20,21). In the H2b.291A gene the furthest upstream element is in best agreement with the consensus sequence and is 3' to the only consensus CCAAT sequence in this gene. In the H2b.291B gene the element closest to the TATAA box matches the consensus sequence best and is also downstream of a CCAAT sequence. Thus it is likely that, if the H2b consensus sequence has a function, the functional copy is located at

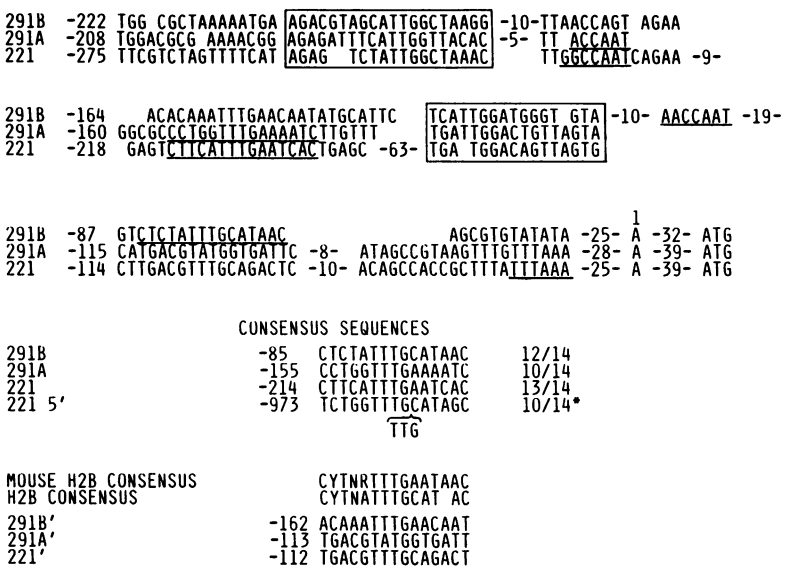


Fig. 4. Similarities in the 5' flanking regions of the 3 H2b genes. The 5' flanking regions of the 3 H2b genes are aligned to show regions of maximal similarity. The numbers refer to the distances from the ATG codon. The two blocks of sequence similar in all three genes that do include previously reported consensus sequences are boxed. The underlined sequences are the H2b consensus sequence (22) and the CCAAT and TATAA sequences. The final nucleotide shown is the first nucleotide of the respective mRNAs. The numbers within the sequences indicate the number of dissimilar nucleotides between the regions shown. The presumed "functional" consensus sequences are compared with the H2b consensus sequences of other species (22). The H2b.2215' sequence is the sequence located 3' to a CCAAT sequence near the H3.2 gene. This sequence contains a three base insertion in the consensus sequence. Beneath these sequences are shown the presumably "non-functional" remnants of the H2b consensus sequence which are not downstream of a CCAAT sequence but which are in regions which show similarity among all three genes.

different distances with respect to the promoter in the two genes.

There are three regions which are similar to the H2b consensus sequence present 5' to the H2b.221 gene. Two copies are present in a location similar to those in the other genes, suggesting that the three promoters arose from the same primordial sequence. As in the H2b.291A gene, the upstream element is most similar to the consensus sequence and present 3' to a CCAAT sequence. It is likely that this is the functional element. The third copy of the element is located about 970 bases 5' to the gene. The only other CCAAT sequence in the flanking region is located in an appropriate position upstream of this sequence. The coincident occurrence of these two elements suggests

that the far upstream element may have a function, perhaps in facilitating the entry of the RNA polymerase. In agreement with this interpretation there is also a CCAAT sequence located 855 bases 5' to the H3.2 gene in this long spacer sequence (Fig. 2C.). Again this is the only CCAAT sequence 5' to the H3.2 gene on this strand of the DNA other than those found close to the H3.2 gene and it nearly overlaps the probable functional CCAAT sequence 5' of the H2b gene. In contrast to the H2b genes, we have not found conserved sequences other than those around the CCAAT sequence upstream of the mouse H3 genes on chromosome 13 (19) and highly conserved sequences have not been reported 5' of other core histone genes (22). Thus both of the histone genes in this divergently transcribed cluster have possible functional upstream sequences located near the promoter of the other gene. It is very unlikely that this arrangement is coincidental but suggests that there are may be entry points for RNA polymerase at both ends of the spacer, which then direct the polymerase toward the appropriate gene.

There are two additional regions of similarity among the 5' flanking sequences of the H2b genes. A 22 base sequence 5' of the distal H2b consensus sequence has been highly conserved among the three genes. The most similar sequences in these two regions are between the H2b.221 and H2b.291A genes rather than between the closely linked H2b genes. The same is true of a 14 base sequence located between the two H2b-specific 5' sequences. Whether these regions are functionally important, or whether the similarity is due to a common precursor to these three genes is not known.

Comparison of the 3' flanking sequences

The 3' flanking sequences of the mouse histone genes generally show no similarity other than the sequence around the hairpin loop at the 3' end of the mRNA. The H2b genes provide the first example where this is not true. The first 120 nucleotides of the 3' flanking sequences of the 3 H2B genes are compared in Figure 5A. There is clear sequence similarity between the H2b.291A and H2b.291B for 50 nucleotides 3' to the gene, extending to the end of the hairpin loop. This could be due to homology remaining from the duplication event or from gene conversion followed by divergence in the less highly constrained flanking region. The 3' untranslated region of the H2b.221 gene which is located on the same chromosome, presumably in the same cluster, also shows similarity to the other H2b genes. We infer from the S1 nuclease mapping data that the H2b.291A gene has a 3' untranslated region which is similar enough to several other H2b genes that it is not distinguished by S1

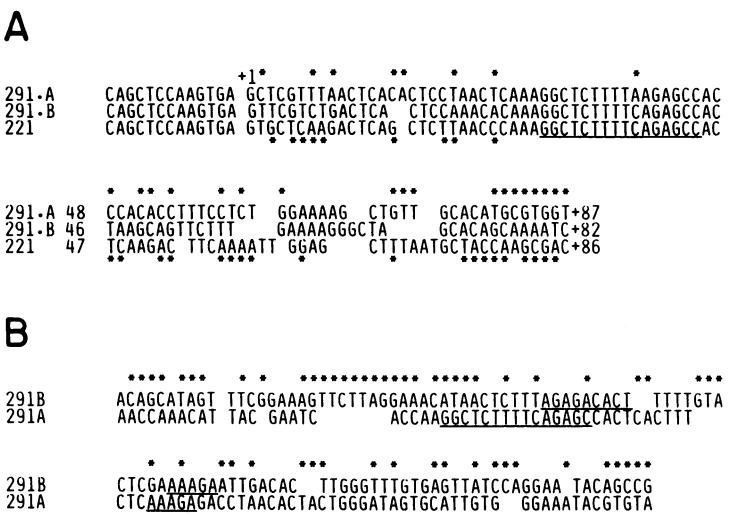


Fig. 5. Similarities in the 3' flanking regions of mouse histone A. The sequences 3' to the three histone H2b genes are shown aligned to give maximum similarity. The dyad region is underlined.
B. The sequences 3' to the H2a.291A gene and the H2a.291B pseudogene are shown aligned to show the similarity. The same sequences are underlined.

nuclease mapping (see below). The 3' untranslated regions of the mouse histone H3 genes are not similar to each other (13, 19).

The 3' ends of the H2A.291 genes are compared in Figure 6B. There is residual homology presumably remaining from the original hairpin loop present in the H2A.291B pseudogene. The purine rich sequence normally present 10-15 nucleotides downstream of the 3' end of the hairpin loop is also present in the pseudogene.

Only the H2b genes show extensive similarity in the 3' untranslated regions. Whether this is the result of gene conversion extending into the 3' untranslated region or to an unknown functional role of this sequence in H2b genes is not known. It is intriguing that only the H2b genes show similarity in both the 5' and 3' flanking regions.

Expression of the H2b genes

The amount of expression of the different H2b mRNAs was determined by S1 nuclease mapping. In addition to the start site of the mRNA, there is a major fragment protected from S1 nuclease which extends exactly to the ATG codon. This is the result of protection of the probe by a number of different mRNAs

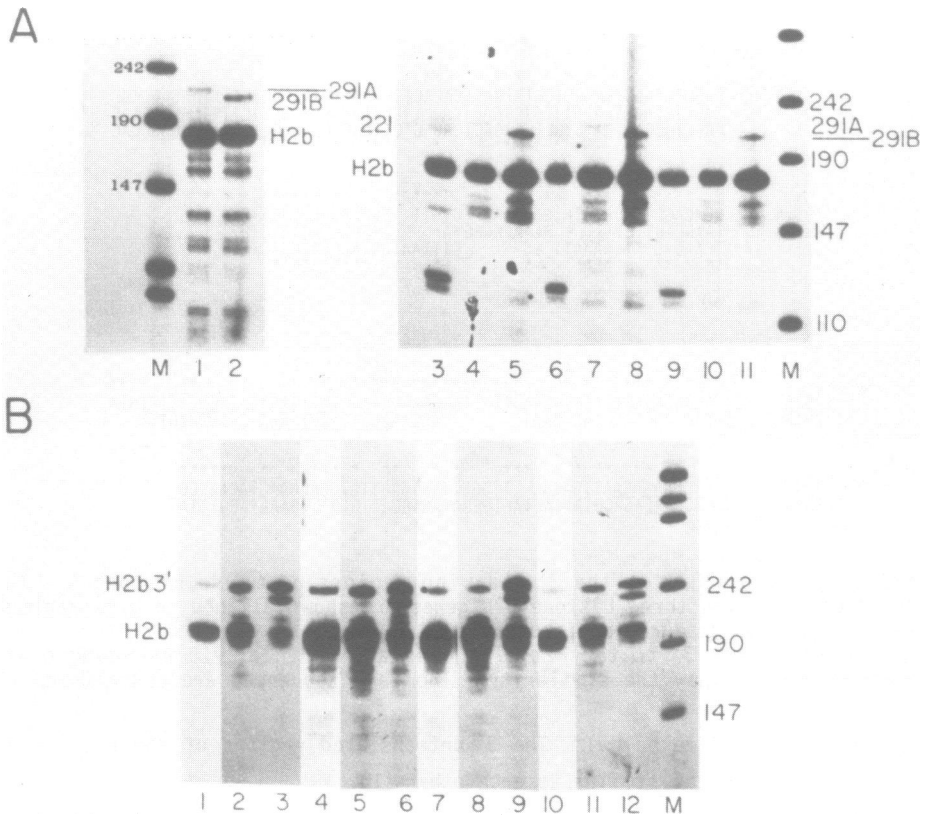


Fig. 6. Expression of the H2b genes. A. The 5' end of the NcoI site (codon 59 of the H2b gene) of the 3 H2b genes was labeled with $^{32}PO_4$ and hybridized to 3 µgms of RNA from exponentially growing mouse myeloma cells (lanes 1,2), mouse L cells (lanes 3-5), mouse 3T6 fibroblasts (lanes 6-8), 9 day mouse embryos (lanes 9-11). The S1 resistant hybrids were resolved by gel electrophoresis and detected by autoradiography as described previously (2). The H2b.219A mRNA was analyzed in lanes 1,4,7,10; the H2b.291B gene in lanes 2,5,8,11; and the H2b.221 gene in lanes 3, 6, 9. B. The 3' end of the same NcoI site was end-labelled and used as a probe in the S1 nuclease assay. The RNAs were from mouse L cells (lanes 1-3); mouse 3T6 fibroblasts (lanes 4-6); mouse myeloma cells (lanes 7-9); 9 day mouse embryos (lanes 10-12). The H2b.221 gene was analyzed in lanes 1,4,7,10; the H2b.291B gene in lanes 2,5,8,11; the H2b.291A gene in lanes 3,6,9,12. The marker (lane M) was pUC 18 digested with HpaII. The protected fragments were H2b- protection to the ATG codon (panel A) or the TAA codon (panel B). H2b.221- protection to the start of the H2b.221 mRNA (which appears heterogenous in this assay); H2b.291A- protection to the end of the H2b.291A mRNA; H2b.291B- protection to the 5' end of the H2b.291B mRNA. The band at 118 nts in the H2b.221 lanes (3,6,9, panel A) is due to the amino acid change at position 18 (13). H2b3'- protection to the end of the mRNA from the specific genes (in the case of H2b.291A protection of mRNAs from several different genes. All of these H2b genes have 3' untranslated regions which are of similar length.

with similar coding regions and divergent flanking regions (2). This allows one to measure the amount of expression of the different H2b genes. Figure 6 shows that the H2b.291A gene codes for a mRNA with a slightly longer 5' untranslated region than the H2b.291B mRNA. We measured the amount of expression of the three H2b genes in three different mouse cell lines (myeloma cells, C127 mouse fibroblasts, mouse L cells) and in 9 day fetal mice. Similar amounts of expression were found in all four samples; the H2b.291B gene was expressed more strongly than the H2b.221 or H2b.291A gene. The H2b.291B has the 5' H2b consensus element positioned closest to the TATAA sequence and it is attractive to think that this may be the reason for the high level of expression. The positioning of this sequence relative to the TATAA sequence may help determine the strength of the promoter. This element has been shown to be essential for maximal expression of human H2b genes (24).

Similar results were obtained when the same RNAs were analyzed using a probe which measures the 3' end of the mRNA (Fig. 6B). The H2b.291B mRNA was more abundant than the H2b.221 mRNA in all the cells tested. Using the 3' end of the H2b.291A gene as a probe produced an unexpected result. The major protected fragment extended well beyond the TAA codon, with over 50% of the protected DNA extending to the end of the H2b.291A mRNA. We interpret this as conservation of the 3' untranslated region of a large number of the H2b mRNAs. This is supported by the sequence comparisons above (Fig. 5).

Organization and evolution of histone genes

The H2a and H2b gene pair on MM291 presumably arose by an original event resulting in the inverted duplication of the entire gene pair. This is indicated by the exact similarity in size between the two gene pairs. To our knowledge, this gene pair represents the closest juxtaposition (230 nucleotides) of two genes in vertebrates. As a result of the closeness of the two genes the potential regulatory sequences overlap somewhat. However the same overlap of potential regulatory sequences is also found in the H2b.221 and H3.2-221 genes which are >1 kilobase apart, suggesting that overlapping the signals may be functionally important. This type of organization with histone genes organized with their 5' ends juxtaposed is found commonly, but by no means exclusively. H2a-H2b gene pairs which are divergently transcribed are found in yeast (8), drosophila (25), chickens (19), newt (26), frogs (27), and humans (28). In the mouse there are clearly H2a and H2b genes present in different environments (13). There are also examples of histone genes closely linked (< 1 kb) and transcribed form the same strand (2).

The results reported here suggest that the degree of heterogeneity of

histone protein sequences is much greater than previously thought. It is not known whether there is a functional significance to the non-allelic histone variants. A number of the variants (e.g. H2a.1 and H2a.2, H3.2 and H3.3) are found in birds and mammals suggesting that there is a selective pressure maintaining the variants during evolution (29). One might expect that neutral mutations with respect to amino acid sequence could occur in a member of the multigene family and be maintained within a species. This could explain the variant H2a and H2b genes we have found. It is also possible that each one of these variants has a particular function. If this is the case then we would predict that similar variants will be found in other species. It seems most likely that these are neutral variants, since they have only been found in H2a and H2b genes and not in H3 genes, known to be under more severe selective constraints.

These results strongly suggest that the high conservation of histone coding sequences in the mouse is due to gene conversion which is targeted at the coding region. The gene conversion must generally involve small areas usually not larger than the coding region. Inspection of the sequence variation among the mouse histone genes sequenced thus far, particularly the H3 genes (18) but also the H2a.291A and H2a.221 genes reveals that a high proportion of the nucleotide changes occur near the ends of the gene, consistent with this interpretation. This may mean that most of the gene conversion events do not involve the whole coding region but a smaller target. There has been at least one gene conversion event which presumably extended into the 3' untranslated region of the H2b.291A and H2b.291B genes and this may have occurred in other H2b genes as well, since they have much more similar 3' flanking regions than other mouse histone genes.

The late sea urchin histone genes (30) and the chicken histone genes (20) which show a similar pattern of highly conserved coding regions and divergent flanking regions, may be evolving by a similar mechanism.

ACKNOWLEDGEMENTS

This work was supported by NIH grant GM 29832 to W.F.M. We thank Dr. Susan Wellman for providing the original MM291 clone and Dr. Reed Graves for helpful discussions.

REFERENCES

1. Hentschel, C. C. and Birnstiel, M. L. (1980) Cell 25 301-313.
2. Graves, R. A., Wellman, S.E. Chiu, I-M. and Marzluff, W. F. (1985) J. Mol. Biol. 183 179-94.

3. D'Andrea, R.J., Cloes, L.S., Lesnikowski, C., Tabe, L. and Wells, J.R.E. (1985) Mol. Cell. Biol. 5 3108-15.
4. Sierra, F., Lichtler, A., Marashi, F., Rickles, R., Van Dyke, T., Clark, S. Wells, J., Stein, G. and Stein, J. (1982) Proc. Nat. Acid Sci. 79 1795-99.
5. Heintz, N., Zernik, M. and Roeder, R. G. (1981) Cell 24 661-68.
6. Smith, M.M. and Murray, K. (1983) J. Mol. Biol. 169 641-661.
7. Woudt, L. P., Patink, A., Kempers-Veenstra, A. E., Jansen, A. E. M., Mager, W. H. and Planta, R. J. (1983) Nuc. Acids. Res. 11 5347-60.
8. Hereford, L., Fahner, K., Woolford, J., Rosbash, M. and Kaback, D.B. (1979) Cell 18 1261-71.
9. Marzluff, W. F. and Graves, R. A. (1984) in Histone Genes: Structure, Organization and Function (ed. Stein, G., Stein, J. and Marzluff, W. F.) J. Wiley and Sons, N.Y. 281-315.
10. Triputti, P., Emanuel, B.S., Croce, C.M., Green, L.G., Stein, G.S. and Stein, J.L. (1986) Proc. Nat. Acad. Sci. 83 3185-88.
11. Graves R.A., W.F. Marzluff, D.A. Giebelhaus, and G. Schultz. (1985) Proc. Nat. Acad. Sci. 82 5685-89.
12. Brown, D.T., Wellman, S.E., and Sittman, D.B. (1985) Mol. Cell. Biol. 5 2879-86.
13. Sittman, D. B., Graves, R. A. and Marzluff, W. F. (1983) Nuc. Acids Res. 11 6679-91.
14. Franklin, S. G. and Zweidler, A. (1977) Nature 266 273-74.
15. Zweidler, A. (1984) in Histone Genes: Structure, Organization and Regulation (eds. Stein, Stein and Marzluff) John Wiley and Sons, N.Y. p. 339-369.
16. Maxam, A.M. and Gilbert, W. (1980) Methods Enzymol. 65 499-560.
17. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Nat. Acad. Sci. 74 5463-67.
18. Pustell, J. M. and Kafatos, F. C. (1982) Nuc. Acids Res. 10 51-59.
19. Taylor, J.D., Wellman, S.E., and Marzluff, W.F. (1986) J. Mol. Evol., 23 242-49.
20. Harvey, R.P., Robins, A. J. and Wells, J.R.E. (1982) Nuc. Acids Res. 10 7851-63.
21. Winkfine, R.J., Connor, W., Mezquita, J. and Dixon, G.H. (1985) J. Mol. Evol. 22 1-19.
22. Wells, D. E. (1986) Nuc. Acids Res. 14 r119-r149.
23. Marzluff, W. F. (1986) in DNA Systematics (S.K. Dutta, ed.) CRC Press, Boca Raton, FL. p. 169-168.
24. Sive, H.L., Heintz, N., and Roeder, R.G. (1986) Mol. Cell Biol. 6 3329-40.
25. Lifton, R. P., Goldberg, M. L., Karp, R.W. and Hogness, D. S. (1978) Cold Spring Harbor Symp. Quant. Biol. 42 1047-51.
26. Stephenson, E.C., Erba, H.P. and Gall, J.G. (1981) Nuc. Acids Res. 9 2281-95.
27. Turner, P.C., Aldridge, T.C., Woodland, H.R. and Old, R.N. (1983) Nuc. Acids Res. 11 971-86.
28. Marashi, F., Prokopp, K., Stein, J. and Stein, G. (1984) Proc. Nat. Acad. Sci. 81 1936-40.
29. Urban, M. K. and Zweidler, A. (1983) Developmental Biology 95 421-28.
30. Roberts, S. B., Weisser, K. E. and Childs G. (1984) J. Mol. Biol. 174 647-62.