Volume 15 Number 3 1987 Nucleic Acids Research

Structure and organization of the chicken H2B histone gene family

David K.Grandy+ and Jerry B.Dodgson

Departments of Microbiology and Biochemistry, Michigan State University, East Lansing, MI 48824, USA

Received September 9, 1986; Revised and Accepted January 9, 1987

ABSTRACT

The results of Southern blotting experiments confirm that the chicken H2B histone gene family contains eight highly homologous members. One or two more sequences which are considerably divergent from the others appear to exist in the chicken genome. Seven of the eight H2B genes have been cloned and sequenced. All seven genes fall in two histone gene clusters, but no common arrangement exists for the clusters themselves. Three different H2B protein variants are encoded by these seven genes. The nucleotide sequence homology among the genes within their coding sequences appears to exceed that required for the corresponding protein sequences, suggesting that histone H2B mRNA sequence and structure are both selected during evolution. An analysis of the 5' flanking sequence data reveals that these genes possess CCAAT and TATA boxes, elements commonly associated with genes transcribed by RNA polymerase II. In addition, these genes all share an H2B-specific element of the form: ATTTGCATA. The 3' sequences of these genes contain the hyphenated symmetrical dyad homology and downstream purine-rich sequence shared by histone genes in general.

INTRODUCTION

The histone genes constitute a complex, multigene family, whose structure and organization have been studied in a variety of eucaryotes (1,2). Since there are five major types of histone, the overall histone gene family can be considered to be composed of five subfamilies of H1, H2A, H2B, H3 and H4 histone genes. In addition, the nucleated erythrocytes of birds and other vertebrates contain a unique histone, H5 (3,4), which in the most general terms may be thought of as a member of the H1 subfamily. Furthermore, in most multicellular eucaryotes several variants exist for each of the various types of histone which differ in their primary structure and expression pattern. Zweidler (5) has divided the variants into replication-type histones whose expression is coupled to the S-phase of the cell cycle and replacement-type histones which seem to be expressed at a low constitutive level throughout the cell cycle (6,7,8). Indeed, we have shown that the chicken H3 histone gene subfamily is actually composed of two very

different sets of H3 genes, the H3.2 replication variants and the H3.3 replacement variants (9).

In most multicelled organisms, there exist from 10 to several hundred histone genes of each major subfamily per haploid genome (1) whereas Neurospora (10) and yeast (11) contain only one or two copies of each type of histone gene, respectively. Among the multicelled organisms, chicken appears to have one of the lowest histone gene repetition frequencies ranging from six for the H1 histone genes to 8-10 for the core histone genes (4,12,13).

Given the number of different genes in each subfamily, the prospect of individual regulation of the expression of each histone gene seemed worthy of consideration. This was given further credence when it was shown that in avians (14,15), mammals (16,17), and, to a lesser extent, amphibians (18), there was little or no uniformity in the position and orientation of a given type of gene within a histone gene cluster nor any long range homology between the sequences that flank the coding regions of two histone genes of the same type. Examples of differential regulation of different histone gene clusters and/or different members of the same histone gene subfamily have accumulated over the past few years. Childs et al. (19) have shown that sea urchin late stage histone genes form a separate and very different cluster from the previously studied blastula-stage genes. We demonstrated that the structure and organization of the replacement H3.3 variant histone genes of chicken are very different from the replication H3.2 genes (9,20). Similar conclusions were drawn for an apparent replacement H2A chicken histone gene by Harvey et al. (21). The chicken H5 histone gene is also very different in structure and organization as well as in expression from the replication histone genes (3,4). Furthermore, Perry et al. (22) have shown that Xenopus laevis contain at least two very different types of histone gene clusters.

Beyond these isolated examples, however, it has been difficult to estimate the actual diversity in structure and expression among a complete histone gene subfamily. This is because of the relatively high copy numbers of the subfamilies and the high level of sequence homology between the coding regions of all members of the subfamily. In preparation for a comparative analysis of the expression of different members of the chicken H2B subfamily, we have isolated and determined the DNA sequence of most of the members of this gene family. The results of these experiments show that there is a high level of homology among all H2B genes in the coding regions,

but considerable diversity in the flanking regions except for small blocks of known consensus sequence.

MATERIALS AND METHODS

Materials

Restriction enzymes and other nucleic acid modifying enzymes were purchased from International Biotechnologies, Inc.; New England Biolabs, Inc. and Bethesda Research Labs, Inc. and used according to the manufacturer's specifications. Other materials and bacterial strains used were as described previously (23,24).

Isolation and Purification of Recombinant DNA Molecules

The λCharon4A chicken DNA library used in these studies was constructed by Dodgson et al. (23) and contains genomic fragments generated by AluI/HaeIII partial digestion of chicken DNA ligated to synthetic EcoRI linkers. Phage were plated at a density of 50,000 plaques per plate and approximately 10 genomic equivalents were screened. The library was screened using a previously described internal H2B gene-specific 0.3 kilobase pairs (kb) BstEII fragment from pKR1a-1.3 as probe (24). Hybridization screening, nick translation labeling of DNA, phage and plasmid DNA purification, and restriction mapping were as described previously (24-26). All manipulations of recombinant DNA molecules were done in accordance with current National Institutes of Health guidelines.

Southern Hybridization

High molecular weight genomic chicken red blood cell DNA was digested with restriction enzymes, precipitated, loaded onto 0.7% agarose gels at a concentration of 15 micrograms per lane, electrophoresed in a vertical gel support, stained with ethidium bromide and photographed. The DNA was denatured in alkali, the gel was then neutralized and DNA was transferred to nitrocellulose filters by capillary action as described by Southern (27). The filters were prehybridized and probed in a hybridization solution previously described (14) at 37° or 42° overnight. The Southern blots were washed and exposed as described in the legend to Fig. 1.

DNA Sequencing

DNA sequencing was performed by a modified Maxam and Gilbert (28) chemical degradation technique as described by Smith and Calvo (29). DNA fragments with 5' protruding ends were phosphatase treated and end-labeled with polynucleotide kinase as described (28). Blunt-ended fragments were labeled either by exonuclease III digestion followed by filling in with the

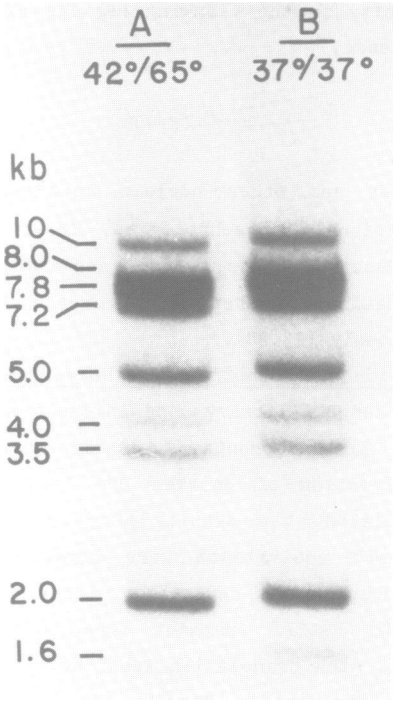

Figure 1. Genomic Southern blot of EcoRI/HindIII digested chicken red blood cell DNA. The blots were probed with a BstEII 0.3 kb H2B-specific sequence, nick-translated to a specific activity of 1 x 10^8 cpm. 100 ng of the 0.3 kb probe was used in 7 ml of hybridization solution. The washed blots were exposed to film with an intensifying screen for 10 days at -70°. A. The hybridizations were carried out at 42° and then the filter was washed at 65° in 0.1 x SSC. B. The hybridization was performed at 37° and the filter was washed at 37°.

appropriate labeled nucleotide (29) or by directly kinasing the phosphatased ends as described (28). In most cases, the additional T>G reaction (30) was used to corroborate the results of the four Maxam and Gilbert reactions.

RESULTS AND DISCUSSION

Estimate of the Number of Chicken H2B Histone Genes

In order to assess the number of chicken H2B histone genes present in the haploid chicken genome, Southern blotting experiments were performed. Chicken red blood cell DNA was prepared and digested with the restriction enzymes EcoRI and HindIII. These enzymes were chosen because they do not cut within either of the two H2B genes previously characterized (12). The

chromosomal DNA was electrophoresed, blotted to nitrocellulose filters and then probed with a nick-translated 0.3 kb BstEII restriction fragment prepared from pKR1a-1.3 (24). As previously described, this clone contains a single H2B gene, and the 0.3 kb BstEII fragment consists solely of H2B coding sequence. Hybridization was carried out at 37° or 42°; the filters were then washed at 37° or 65°, respectively, in .015 M NaCl, 1 mM Na citrate, pH 7.0.

The results of such an experiment are shown in Figure 1. The H2B-specific probe hybridizes to eight distinct bands under the more stringent conditions (Fig. 1A). These bands correspond to sizes: 10 kb, 8.0 kb, 7.8 kb, 7.2 kb, 5.0 kb, 4.0 kb, 3.6 kb and 2.0 kb. The pattern observed is in agreement with the results of Ruiz-Carrillo et al. (4), except for the fact that they used the full pKR1a-1.3 insert as probe rather than an internal coding probe. Also in agreement with Ruiz-Carrillo et al. (4), comparison of such a chromosomal blot to serially diluted standard amounts of the pKR1a-1.3 H2B gene shows that, for example, the 2.0 kb band intensity corresponds well with one copy per haploid genome (S.-Y. Son and J. Dodgson, unpublished results). As will be discussed below, the comparison of histone gene-containing clone restriction maps with the chromosomal Southern blot patterns (Fig. 1A; 12,13,31) indicates that the 2.0 kb, 7.2 kb and 8.0 kb bands result from one H2B gene each and the 7.8 kb and 10 kb bands result from two genes each. The 10 kb band appears less intense than expected by this analysis, perhaps because of decreased transfer of larger DNA fragments to the filter. We have not yet cloned the gene corresponding to the 5.0 kb band, but it appears to result from a single H2B gene (13). The 3.5 and 4.0 kb bands consistently appear much less intense than the others listed above. Comparing the stringent hybridization used in Fig. 1A to the more relaxed criteria used in Fig. 1B shows that a 1.6 kb band as well as the 4.0 kb band (and perhaps the 3.5 kb band) clearly result from only partial homology to the H2B probe used. These bands could result from one or more highly variant H2B genes, H2B pseudogenes and/or H2B-unrelated fortuitously hybridizing sequences elsewhere in the genome. Although no replacement variant H2B histones have yet been observed by protein analysis, spermatocyte-specific variant H2B histones exist in mammals (5). If a spermatocyte-specific H2B histone exists in chickens or if there are replacement variant H2B histones that have not been identified by gel electrophoresis, the corresponding gene(s) could account for these weakly-hybridizing bands. These sequences are not present on any of the histone gene clones we have tested to date.

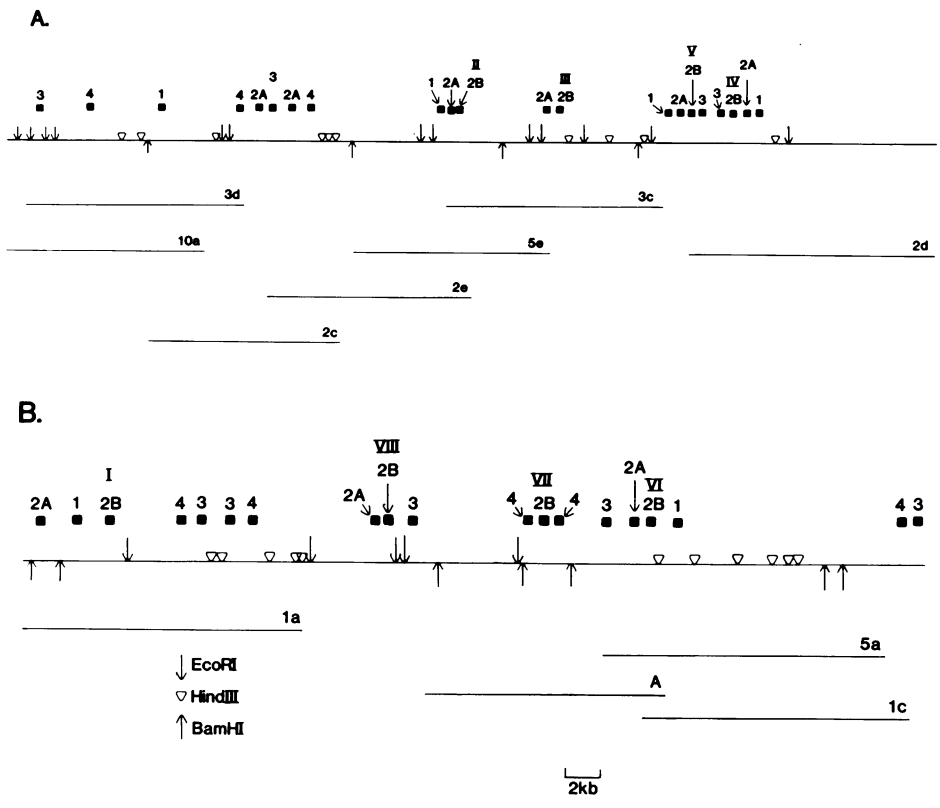

Figure 2. Organization of cloned chicken H2B genes. Location of known histone genes in the clusters are indicated by filled boxes with the appropriate class type designated by numbers over the box. The restriction map of the chromosomal DNA containing the histone genes is shown by a line with restriction sites indicated. Lines below the restriction map indicate the portion of chicken chromosomal DNA contained in a λ clone with the given designation (e.g., 3d for λCH3d). In several cases, preliminary restriction maps of the clones were given in Sugarman et al., (32). Areas not covered by our λ clones were inferred from the results of D'Andrea et al., (13). A. Histone gene cluster containing 21 histone genes including four H2B genes. B. Histone gene cluster containing 19 histone genes including four H2B genes. Roman numerals are placed above the 8 H2B genes in order to help identify the corresponding genes in Fig. 3. H2B-I is the first chicken H2B gene sequenced (24) designated by its plasmid, pKR1a-1.3. H2B-II, the gene on pRR2e-3.5, has also been previously sequenced (12, 31). H2B-III to H2B-VII are, in order, on plasmids: pRR3c-3.5, pPP2d-2.3, pPP2d-4.0, pBRA-5.4, and pBBA-3.0 (see Fig. 3). The genes are listed in this numerical order in Figs. 4-7. H2B-VIII was not present among our clones and is not described further.

Isolation of Chicken H2B Histone Genes

Upon screening a chicken genomic library in λCharon4A with sea urchin H3 and H2A histone gene probes, about 50 histone gene-containing recombinants were isolated (14,32). About 20 of these have been mapped and characterized in detail. Several of these clones were shown to contain H2B-hybridizing regions. We have previously described the complete sequence of two of these H2B genes (12,24). These two genes were found to differ in both their coding and flanking sequences. The same two genes were independently analyzed by Harvey et al. (31).

Since an H2B probe was not used in the initial screening, the chicken genomic library was rescreened with an internal 0.3 kb H2B-specific fragment in order to expand our collection of chicken H2B genes. Several of the resultant H2B-containing phage clones were overlapping with or identical to previously characterized phage (32), whereas others were unique. Restriction mapping and hybridization analysis with various histone gene probes identified seven unique H2B genes among our collection which fall into the three clusters identified by D'Andrea et al. (13). Furthermore, one of our phage clones (λCHA) overlaps two of the clusters identified by that group thereby demonstrating that all eight H2B genes and almost all replication variant chicken histone genes fall into one of two major histone gene clusters. It remains to be seen if these two clusters are also linked in the chicken genome. Figure 2 shows the restriction maps of these two histone gene clusters along with the positions of the various chicken histone genes identified to date. As has been noted previously for the replication variant histone genes (13,14,32), there is no consistent order or arrangement of the clusters. However, six of the eight H2B genes exist in closely-linked divergently-transcribed H2A-H2B gene pairs (13).

Fig. 2A shows that four H2B genes fall within one histone gene cluster which contains 21 genes in total. This includes the H2B gene on λCH2e that was sequenced previously (12,31), an additional H2B gene which is present (along with the λCH2e gene) on λCH3c, and the two genes on λCH2d (the gene at the left of λCH2d as shown in Fig. 2A is only partially within the cloned DNA). Fig. 2B shows the other four H2B genes located in a cluster of 19 histone genes. These include the previously sequenced gene on λCH1a (24,31), the two genes on λCHA (one of which is also on λCH5a), and the gene on a 5 kb EcoRI fragment that is not among our clones.

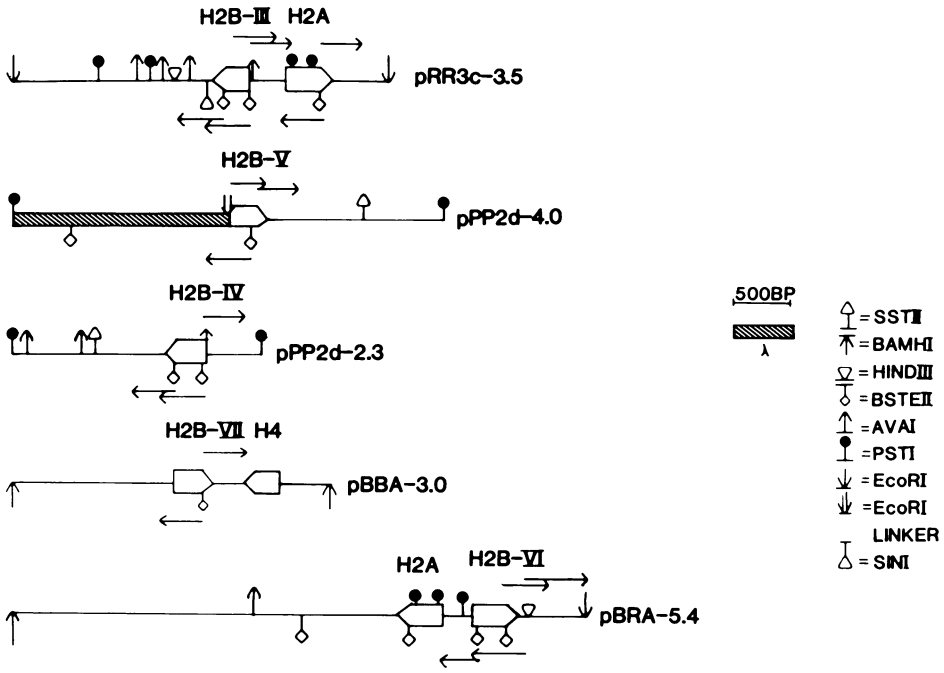

Figure 3. Restriction map and sequencing strategy of H2B gene subclones. Partial restriction maps are given for the subclones containing the five H2B genes whose sequences have not previously been reported. The open horizontal arrows mark the various genes and point in the direction of transcription. The horizontal arrows below the map indicate the extent of the DNA sequence determined for the template strand. H2B genes are designated by Roman numerals as described in the legend to Fig. 2.

Chicken H2B Histone Gene Coding Sequences

The seven H2B genes on our λ clones were subcloned into plasmid vectors, and the restriction maps of the subclones derived by standard procedures. The maps and the sequencing strategy used for the five H2B genes whose sequence was not reported previously are given in Figure 3. Some sequence data was also obtained for two H2A genes and an H4 gene which are closely linked to particular H2B genes (results not shown). The location and transcriptional orientation of these genes are also indicated on the maps.

The H2B coding regions were identified based on sequence comparisons with the two previously characterized genes in pKR1a-1.3 and pRR2e-3.5 (12,24,31). Figure 4 presents a compilation of seven (six complete plus

```
                1   2   3   4   5   6   7   8   9  10  11  12  13  14
              5'  Pro Glu Pro Ala Lys Ser Ala Pro Ala Pro Lys Lys Gly Ser
pKR1a-1.3     ATG CCT GAG CCG GCC AAG TCC GCA CCC GCC CCC AAG AAG GGC TCC
pRR2e-3.5       C           T       G       G               T
pRR3c-3.5       C           T       G       G               T
pPP2d-2.3       C                   G       G
pBBA- 3.0                   A   A   T   A   A           A       G

               15  16  17  18  19  20  21  22  23  24  25  26  27  28  29
               Lys Lys Ala Val Thr Lys Thr Gln Lys Lys Gly Asp Lys Lys Arg
pKR1a-1.3      AAG AAG GCG GTC ACC AAG ACC CAG AAG AAG GGC GAC AAG AAG CGC
pBRA- 5.4                  G
pBBA- 3.0                   T               A       T           A A A

               30  31  32  33  34  35  36  37  38  39  40  41  42  43  44
               Lys Lys Ser Arg Lys Glu Ser Tyr Ser Ile Tyr Val Tyr Lys Val
pKR1a-1.3      AAG AAG AGC CGC AAG GAG AGC TAC TCG ATC TAC GTG TAC AAG GTG
pBBA- 3.0        GA GCG A G   A   A
                Arg Ala

               45  46  47  48  49  50  51  52  53  54  55  56  57  58  59
               Leu Lys Gln Val His Pro Asp Thr Gly Ile Ser Ser Lys Ala Met
pKR1a-1.3      CTG AAG CAG GTG CAC CCC GAC ACG GGC ATC TCG TCC AAG GCC ATG
pPP2d-2.3                                           A
pBBA -3.0                                           A   T   A

               60  61  62  63  64  65  66  67  68  69  70  71  72  73  74
               Gly Ile Met Asn Ser Phe Val Asn Asp Ile Phe Glu Arg Ile Ala
pKR1a-1.3      GGC ATC ATG AAC TCG TTC GTC AAC GAC ATC TTC GAG CGC ATC GCC
pBBA- 3.0      A
               Ser

               75  76  77  78  79  80  81  82  83  84  85  86  87  88  89
               Gly Glu Ala Ser Arg Leu Ala His Tyr Asn Lys Arg Ser Thr Ile
pKR1a-1.3      GGC GAG GCG TCG CGC CTG GCG CAC TAC AAC AAG CGC TCG ACC ATC

               90  91  92  93  94  95  96  97  98  99 100 101 102 103 104
               Thr Ser Arg Glu Ile Gln Thr Ala Val Arg Leu Leu Leu Pro Gly
pKR1a-1.3      ACG TCG CGG GAG ATC CAG ACA GCC GTG CGG CTG CTG CTG CCC GGC
pRR2e-3.5                           G
pRR3c-3.5                           G
pPP2d-2.3                           G
pPP2d-4.0                           G

              105 106 107 108 109 110 111 112 113 114 115 116 117 118 119
              Glu Leu Ala Lys His Ala Val Ser Glu Gly Thr Lys Ala Val Thr
pKR1a-1.3     GAG CTG GCC AAG CAC GCG GTC TCC GAG GGT ACC AAG GCG GTC ACC
pRR2e-3.5                               G       C
pRR3c-3.5                               G       C
pPP2d-2.3                               G
pPP2d-4.0                                       C
pBBA- 3.0                                       C

              120 121 122 123 124 125
              Lys Tyr Thr Ser Ser Lys    3'
pKR1a-1.3     AAG TAC ACC AGC TCC AAG TAG
pRR2e-3.5                           A
pBBA- 3.0                   A   A   A
pPP2d-4.0         T
                 Ile
```

Figure 4. Chicken H2B coding sequence comparison. The gene on pKR1a-1.3 was used as the prototype against which the other sequences were compared. Only those sequences which differ from the pKR1a-1.3 gene are shown for the other genes. The amino acid substitutions are indicated where they occur. Seven unique H2B genes are represented. The gene on pPP2d-4.0 contains the sequence 3' of nucleotide 106.

TABLE 1. Comparison of the H2B histone genes with respect to the H2B gene of pKR1a-1.3.

	Number of coding nucleotide differences compared to 1a	% sequence divergence	Protein coding differences		
			#	residue	substitution
Class I					
pKR1a-1.3	-	-	-	-	-
pRR2e-3.5	8	2.1	-	-	-
pRR3c-3.5	8	2.1	-	-	-
pPP2d-2.3	5	1.3	-	-	-
pBRA-5.4	1	0.3	-	-	-
Class II					
pBBA-3.0	29	7.7	3	31	Lys → Arg
				32	Ser → Ala
				60	Gly → Ser
Class III					
pPP2d-4.0*	3	1.1	1	122	Thr → Ile

*Sequence available only from codon 36 onward.

that available for the pPP2d-4.0 gene) H2B gene coding sequences. The H2B coding region in pKR1a-1.3 was chosen as the prototype against which the others were compared.

An analysis of these sequences reveals that all of the chicken H2B genes that have been completely sequenced are uninterrupted and code for proteins composed of 125 amino acids. The coding region of each gene is 375 nucleotides long. These genes tend to be rich in G and C residues which can account for 58% (pBBA-3.0) to 64% (pRR3c-3.5) of each coding region. The striking character of this GC bias is reflected in the number of CpG dinucleotides which can number as many as 40 in certain H2B genes, whereas CpG is generally relatively rare in eucaryotic DNA, although less so within coding sequences (33).

As shown in Table 1, each of the seven H2B coding regions is unique (pRR2e-3.5 and pRR3c-3.5 differ only in their stop codon), but they are all found to share extensive sequence homology to the pKR1a-1.3 H2B gene ranging from 92% to 99.7%. Most of the differences in sequence are silent, falling

in the third or, occasionally, the first codon nucleotide. The H2B gene on pBBA-3.0 is by far the most different gene from the 1a gene (and the other five as well). These two genes differ by 29 base pairs (bp) in the 375 bp of coding sequence, with three of the changes leading to altered amino acids (Table 1). The partial H2B gene on pPP2d-4.0 for which sequence is available only from codon 36 onward contains a single amino acid change relative to the pKR1a-1.3 gene, but contains only two other alterations in nucleotide sequence. The other four genes all code for the same amino acid sequence as does the 1a gene. They differ in nucleotide sequence from the 1a gene by from 1 to 8 bp (Table 1).

The amino acid sequence predicted for all seven genes contains isoleucine at position 61 whereas the only complete protein sequence of a chicken H2B histone (34) contains a serine at that position. However, the sequenced calf H2B histones contain isoleucine at position 61 (35), and isoleucine at 61 is also in agreement with the chicken H2B histone amino acid composition data of Urban et al. (36). Thus, it seems likely that in fact isoleucine is the most common, if not the only, amino acid at position 61 of the chicken H2B histone. Urban et al. (36) identified two forms of chicken H2B histone by gel electrophoresis, H2B.1 and H2B.2. Urban and Zweidler (37) subsequently showed that the major chicken H2B histone, H2B.1, is synthesized in the typical replication-dependent fashion, while the minor variant, H2B.2, appears to be expressed in a partially replication-independent manner. Since Urban et al. (36) showed that H2B.2 contains alanine at position 32 and serine at position 60, the H2B.2 variant clearly corresponds to the pBBA-3.0 gene (Class II in Table 1). It is interesting that this variant is closely linked to another H2B gene (the one on pBRA-5.4) which, however, codes for the H2B.1 variant. Our sequence data suggests a third class of H2B gene (Class III in Table 1) corresponding to the gene on pPP2d-4.0. A separate H2B histone corresponding to this gene has not been identified. The gene may not be expressed (we do not have the complete sequence of the gene, so it could potentially be a pseudogene), or may be expressed at relatively low levels. Alternatively, the protein product of the pPP2d-4.0 gene may not be separable from H2B.1. Except for the single amino acid change noted, the pPP2d-4.0 gene is very similar in nucleotide sequence to the Class I, H2B.1 genes.

Although most of the nucleotide differences seen between the various H2B genes are silent with respect to the histone amino acid sequence, their distribution throughout the gene is clearly not random. Most of the diver-

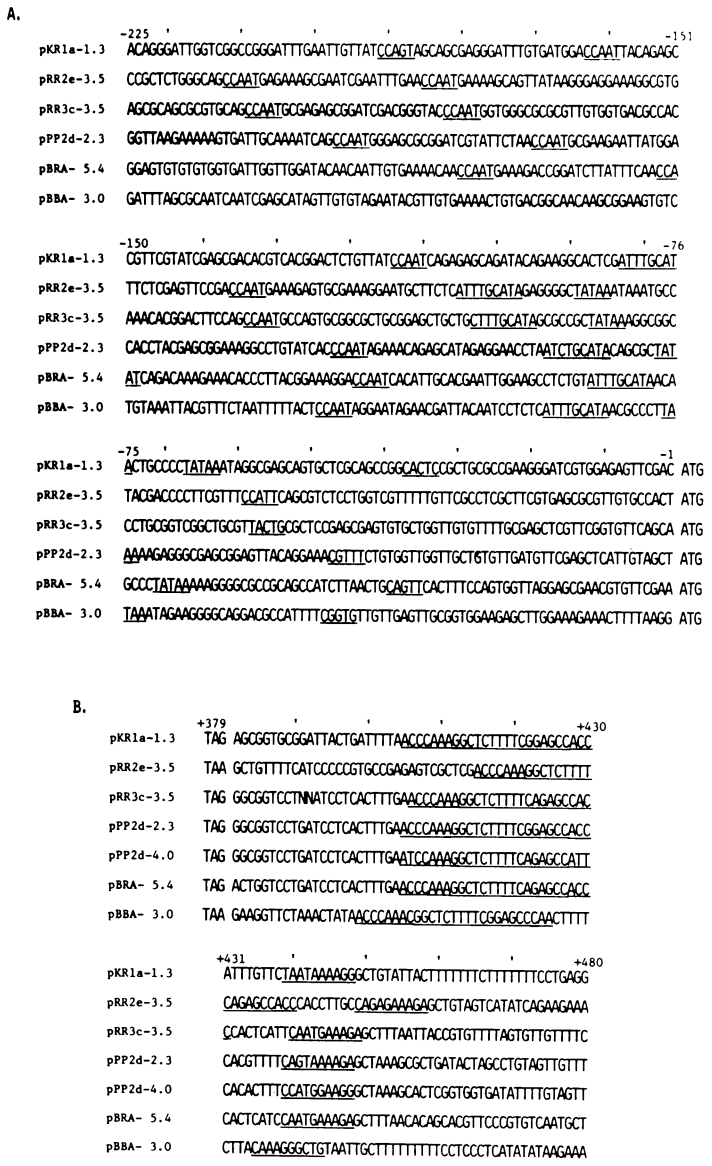

Figure 5. Flanking sequence regions of the chicken H2B genes. A. DNA sequences are compared for 225 nucleotides 5' to the initiation codons of the six chicken H2B genes for which they are available. Probable cap sites, TATA and CCAAT consensus sequences and the H2B-specific ATTTGCATA sequences are underlined. B. DNA sequences for 100 nucleotides 3' to the seven chicken H2B genes are compared. The hyphenated symmetrical dyad and the downstream purine rich sequences are underlined. Nucleotides are numbered such that the A in the ATG initiation codon is +1.

```
          -191                                                                +1
pKR1a-1.3    CCAGT-23bp-CCAAT-44bp-CCAAT-26bp-ATTTGCATACTGCCCCTATAAATA-22bp-CACTCCG-30bp-ATG

          -212                                                                +1
pRR2e-3.5    CCAAT-22bp-CCAAT-44bp-CCAAT-26bp-ATTTGCATAGAGGGGCTATAAATA-22bp-CCATTCA-52bp-ATG

          -209                                                                +1
pRR3c-3.5    CCAAT-22bp-CCAAT-43bp-CCAAT-26bp-CTTTGCATAGCGCCGCTATAAAGG-21bp-TACTGCG-51bp-ATG

          -197                                                                +1
pPP2d-2.3    CCAAT-22bp-CCAAT-43bp-CCAAT-24bp-ATCTGCATACAGCGCTATAAAAGA-22bp-CGTTTCT-40bp-ATG

          -180                                                                +1
pBRA-5.4     CCAAT-22bp-CCAAT-29bp-CCAAT-27bp-ATTTGCATAACAGCCCTATAAAAA-24bp-CAGTTCA-32bp-ATG

                              -124                                            +1
pBBA-3.0                       CCAAT-26bp-ATTTGCATAACGCCCTTATAAATA-21bp-CGGTGTT-41bp-ATG


Consensus Sequence:

CCAAT-22bp-CCAAT-29-44bp-CCAAT-24-27bp-ATTTGCATA-3bp-(G,C)4TATAAANR-21-24bp-YNNTNYN-30-52bp-ATG
      CCAAT BOXES                  H2B-SPECIFIC          TATA BOX       CAP SITE
                                     ELEMENT
```

Figure 6. Consensus sequences and their positions 5' to the chicken H2B genes. The sequence and arrangement are given for each of the six H2B genes sequenced in this region, and a consensus for all six genes is given below. N = any nucleotide, Y = pyrimidine, R = purine.

gence falls within the first 35 codons, and much of the remaining divergence in the last 15 codons. Between codons 61 and 111 there is only one site where a single difference in the coding sequence occurs. This, along with the high G-C bias in the third position of codons, suggests that there is in fact evolutionary selection for the sequence of the histone mRNA, irrespective of the resultant amino acid sequence. This may be due to requirements for a specific histone mRNA structure, possibly so that it can interact with regulatory proteins such as might function in the post-transcriptional control often seen for histone mRNA (38).

5' Sequences Flanking the H2B Histone Genes

The DNA sequence of the 5' and 3' flanking regions of each of the H2B genes is given in Figure 5 (except for the uncloned 5' portion of the pPP2d-4.0 gene). When the 5' flanking sequences of the completely sequenced chicken H2B genes are compared, each one is seen to be unique. However, within this variable background a number of blocks of sequence homology can be identified. These are listed in Figure 6. Many of these conserved sequences are shared by other genes transcribed by RNA polymerase II. One such canonical sequence is the CCAAT pentamer (39). At least one copy of this sequence is associated with every chicken H2B gene analyzed and in all but one case (pBBA-3.0) three copies are found. These sequences reside at a distance of 115-212 nucleotides prior to the ATG start codon. The distance

separating multiple CCAAT pentamers is 22-23 bp between the pair of CCAAT sequences furthest 5' and is usually 43-44 bp between the middle and 3' CCAAT pentamer (29 bp in one case). It may be of significance that 4 of the 5 genes with three CCAAT sequences are part of divergently-transcribed H2B-H2A pairs such that promoter elements for both genes of each pair must fall in intergenic regions of about 3-400 bp. It is not clear whether the fact that the pBBA-3.0 gene has only one CCAAT sequence has anything to do with the differences in expression of its corresponding H2B.2 histone variant (36) from the more common H2B.1 histone. About 40 bp downstream of the 3'-most CCAAT sequence in each gene is the consensus TATA sequence (40). This sequence is always of the form TATAAA in chicken H2B genes. About 7 bp in front of the TATA block is a region of sequence that is shared by all the chicken H2B genes (ATTTGCATA, with lesser homology in regions surrounding this core consensus sequence). As shown previously (12,22,31), this sequence is not unique to chicken H2B genes, but indeed is present in sea urchin, X. laevis, and human H2B genes as well, although the sequence is often further upstream of the TATA region in H2B genes of other organisms. Histone gene class-specific 5' flanking sequences seem to be quite common (1,22), although their function remains unknown. A further H2B histone class-specific flanking sequence (GTCATGTGACAAAA) about 90 bp upstream from the mRNA start site (22) is not generally found in the chicken H2B genes, although weakly homologous regions can be found flanking one or two of the chicken genes.

The site where transcription initiates, referred to as the "cap" site, is usually located 25-30 bp downstream of TATA in the histone genes which have been examined (1). Attempts to identify a specific block of sequence, or cap box, associated with this site have not been entirely successful because no strong homology is evident. For the chicken H2B genes, putative cap sites have been indicated based on their distance from TATA in Figure 6. In the case of the pKR1a-1.3 gene, the indicated cap site has been confirmed by S_1 mapping experiments (results not shown).

Chicken H2B Histone Gene 3' Flanking Sequences

A comparison of the flanking sequences 3' of the seven H2B genes again demonstrates that each gene is unique, yet they all have in common stretches of consensus sequence (Fig. 5). A comparison of the consensus 3' sequences for each of the seven H2B genes is found in Figure 7. The 3' data for pRR2e-3.5 was obtained from Harvey et al. (31). The most striking sequence homology that is shared by all seven genes is a 27 bp element located 17-33

```
                    5'                                                    3'
pKR1a-1.3    TAG-23bp-ACCCAAAGGCTCTTTTCGGAGCCACC-8bp-TAATAAAAGG

pRR2e-3.5    TAA-33bp-ACCCAAAGGCTCTTTTCAGAGCCACC-8bp-CAGAGAAAGA

pRR3c-3.5    TAG-24bp-ACCCAAAGGCTCTTTTCAGAGCCACC-8bp-CAATGAAAGA

pPP2d-2.3    TAG-23bp-ACCCAAAGGCTCTTTTCGGAGCCACC-8bp-CAGTAAAAGA

pPP2d-4.0    TAG-23bp-ATCCAAAGGCTCTTTTCAGAGCCATT-8bp-CCATGGAAGG

pBRA-5.4     TAG-23bp-ACCCAAAGGCTCTTTTCAGAGCCACC-8bp-CAATGAAAGA

pBBA-3.0     TAA-17bp-ACCCAAACGGCTCTTTTCGGAGCCCA-10bp-CAAAGGGCTG

Chicken H2B Consensus Sequence:

          TAR-17-33bp-ACCCAAAGGCTCTTTTCRGAGCCACC-8bp-CAATGAAAGA

Consensus Sequence for Histone Genes in General:

                        AACGGCYCTTTTCRGRGCCACC-6-9bp-CAAGAAAGA
```

Figure 7. Consensus sequences and their positions 3' to the chicken H2B genes. Sequences and their arrangements are given for each of the seven H2B genes sequenced. The consensus of these genes is given below along with the more general consensus for most non-polyadenylated histone messages as derived by Hentschel and Birnstiel (1). R = purine, Y = pyrimidine.

nucleotides 3' from the termination codon. Contained within this large sequence is an inverted repeat which is interrupted by a stretch of four T residues followed by a C and then a purine. This sequence is highly conserved in the 3' regions of almost all histone genes examined (1; see, however, 9, for an exception) and is referred to as the hyphenated symmetrical dyad. Another sequence element shared by all of the H2B genes is a purine rich sequence generally of the form CAATGAAGA, located 11-12 bp 3' of the hyphenated symmetrical dyad. The pKR1a-1.3 H2B gene is unusual in that the purine rich sequence for this gene contains the sequence AATAAA which appears to be involved in mRNA polyadenylation (41,42). However, no appreciable amount of the H2B mRNA generated by transcription of this gene appears to be polyadenylated as judged by oligo(dT)-cellulose fractionation and S_1 analysis of mRNA levels (results not shown). A similar AATAAA sequence was found in a chicken H3 histone gene whose mRNA was also not polyadenylated (43).

In summary, we have determined the sequence of seven of the apparent eight major H2B histone genes of chicken. Southern analysis suggests that there may be one or two other H2B genes which, if they exist, appear to be considerably different in nucleotide sequence from typical H2B genes. The sequence analysis shows that the chicken H2B gene family codes for at least three different histone proteins. Six of the seven genes are very similar

in nucleotide sequence even in positions which are silent with respect to amino acid sequence, and even the remaining H2B gene is over 90% homologous to the other six. Sequence analysis of H2B histone gene-flanking regions shows that these are mostly unique to each individual gene, but that some consensus sequence blocks do exist. The most striking of these is an H2B-specific sequence about 7 bp 5' to the TATA sequence in the promoter.

It would be very interesting to know the evolutionary pathway by which the 8 chicken H2B histone genes were generated. For example, is the striking identity of the H2B coding sequences examined due to recent gene duplication and/or gene conversion? Unfortunately, given the very high level of selective pressure on histone coding sequences, it is difficult to identify potential products of gene conversion events. This is especially true since equivalent, comprehensive analyses of other histone gene subfamilies in other vertebrates are not yet available.

It does appear clear that the H2B histone genes (and presumably other chicken histone genes) do not undergo the high levels of gene conversion one might expect in tandemly duplicated histone gene clusters such as those in Drosophila and sea urchin (1). If so, one would expect more sequence conversion in the H2B gene-flanking regions than is observed (Fig. 5). The two H2B genes on pPP2d-2.3 and pPP2d-4.0 (IV and V, respectively, in Figs. 2A and 3), however, may be related by a relatively recent conversion or duplication event. This might be expected since they are symmetrically related in an inverted duplication (Fig. 2A). While the available coding sequence of the pPP2d-4.0 gene is no more similar to the pPP2d-2.3 gene than it is, for example, to the unlinked pKR1a-1.3 gene (I in Fig. 2B), the 3' untranslated regions of the two symmetric genes are more similar to each other than they are to any of the other H2B genes (Fig. 5). Even this similarity in nucleotide sequence does not extend much beyond the functional 3' sequence elements discussed above (to about +460 in Fig. 5). Other than this, there is very little relationship of the location of H2B genes within the histone gene clusters to their sequence similarity. The two linked genes on pRR2e-3.5 and pRR3c-3.5 (II and III, respectively, in Fig. 2A) show some similarity in their 3' untranslated regions but essentially none in their 5' flanking regions beyond generic H2B gene elements. As pointed out above, the most unusual H2B gene, the H2B.2 gene on pBBA-3.0 (VII in Fig. 2B) is most closely linked to the H2B gene (on pBRA-5.4; VI on Fig. 2B) that is most nearly identical in coding sequence to the unlinked H2B gene, pKR1a-1.3 (I in Fig. 2A).

Thus our preliminary conclusion from this work is that the histone gene clusters of chickens are rather stable and rarely subject to duplication and conversion events. Analysis of the histone genes of other avians would be of interest in this regard. Further, it seems unlikely that gene conversion events can explain the high level of sequence identity seen in the coding region of the 7 H2B genes examined, since the primary conserved portion occupies a specific region (codons 61 to 111) within the H2B sequence rather than extending evenly across the gene as a whole.

ACKNOWLEDGEMENTS

We thank Seung-Yeol Son and David Browne of our lab for access to unpublished results. We also thank Barry Sugarman and J. D. Engel (Northwestern U.) for sharing ideas and results prior to publication. This research was supported by grant GM 29687 and by a Research Career Development Award to J.B.D., both from the National Institutes of Health. Initial stages of the work were supported by Grant 1-0759 from the March of Dimes Birth Defects Foundation. Journal Article No. 11949 from the Michigan Agricultural Experiment Station.

+Present address: Institute for Advanced Biomedical Research, Oregon Health Sciences University, 3181 SW Sam Jackson Park Road, Portland, OR 97201, USA

REFERENCES

1. Hentschel, C. C. and Birnstiel, M. L. (1981) Cell 25, 310-313.
2. Stein, G. S., Stein, J. L. and Marzluff, W. F. (1984) Histone Genes, J. Wiley & Sons, New York.
3. Krieg, P. A., Robins, A. J., D'Andrea, R. and Wells, J.R.E. (1983) Nucl. Acids Res. 11, 619-627.
4. Ruiz-Carrillo, A., Affolter, M. and Renaud, J. (1983) J. Mol. Biol. 170, 843-859.
5. Zweidler, A. (1984) In Stein, G. S., Stein, J. L. and Marzluff, W. F. (eds), Histone Genes. J. Wiley & Sons, New York, pp. 339-371.
6. Groppi, V. E. and Coffino, P. (1980) Cell 21, 195-204.
7. Wu, R. S. and Bonner, W. M. (1981) Cell 27, 321-330.
8. Wu, R. S., Tsai, S. and Bonner, W. M. (1982) Cell 31, 367-374.
9. Brush, D., Dodgson, J. B., Choi, O.-R., Stevens, P. W. and Engel, J. D. (1985) Mol. Cell. Biol. 5, 1307-1317.
10. Woudt, L. P., Pastink, A., Kempers-Veenstra, A. E., Jansen, A.E.M., Mager, W. H. and Planta, R. J. (1983) Nucl. Acids Res. 11, 5347-5360.
11. Hereford, L., Fahrner, K., Woolford, J., Jr., Rosbash, M. and Kaback, D. (1979) Cell 18, 1261-1271.
12. Grandy, D. K., Engel, J. D. and Dodgson, J. B. (1983) In Hamer, D. and Rosenberg, M. (eds), Gene Expression, A. R. Liss, Inc., New York, pp. 445-455.

13. D'Andrea, R. J., Coles, L. S., Lesnikowski, C., Tabe, L. and Wells, J.R.E. (1985) Mol. Cell. Biol. 5, 3108-3115.
14. Engel, J. D. and Dodgson, J. B. (1981) Proc. Natl. Acad. Sci. USA 78, 2856-2860.
15. Harvey, R. P., Krieg, P. A., Robins, A. J., Coles, L. S. and Wells, J.R.E. (1981) Nature 294, 49-52.
16. Heintz, N., Zernik, M. and Roeder, R. G. (1981) Cell 24, 661-668.
17. Sittman, D. B., Chiu, I. M., Pan, C. J., Cohn, R. H., Kedes, L. H. and Marzluff, W. F. (1981) Proc. Natl. Acad. Sci. USA 78, 4078-4082.
18. Zernik, M., Heintz, N., Boime, I. and Roeder, R. G. (1980) Cell 22, 807-815.
19. Childs, G., Nocente-McGrath, C., Lieber, T., Holt, C. and Knowles, J. A. (1982) Cell 31, 383-393.
20. Engel, J. D., Sugarman, B. J. and Dodgson, J. B. (1982) Nature 297, 434-436.
21. Harvey, R. P., Whiting, J. A., Coles, L. S., Krieg, P. A. and Wells, J.R.E. (1983) Proc. Natl. Acad. Sci. USA 80, 2819-2823.
22. Perry, M., Thomsen, G. H. and Roeder, R. G. (1985) J. Mol. Biol. 185, 479-499.
23. Dodgson, J. B., Strommer, J. and Engel, J. D. (1979) Cell 17, 879-887.
24. Grandy, D. K., Engel, J. D. and Dodgson, J. B. (1982) J. Biol. Chem. 257, 8577-8580.
25. Dolan, M., Sugarman, B. J., Dodgson, J. B. and Engel, J. D. (1981) Cell 24, 669-677.
26. Maniatis, T., Fritsch, E. F. and Sambrook, J. (1982) Molecular Cloning, A Laboratory Manual. Cold Spring Harbor, New York.
27. Southern, E. (1975) J. Mol. Biol. 98, 503-518.
28. Maxam, A. M. and Gilbert, W. (1980) Methods Enzymol. 65, 499-560.
29. Smith, D. R. and Calvo, J. M. (1980) Nucl. Acids Res. 8, 2255-2274.
30. Simoncsits, A. and Torok, I. (1982) Nucl. Acids Res. 10, 7959-7964.
31. Harvey, R. P., Robins, A. J. and Wells, J.R.E. (1982) Nucl. Acids Res. 10, 7851-7863.
32. Sugarman, B. J., Dodgson, J. B. and Engel, J. D. (1983) J. Biol. Chem. 258, 9005-9016.
33. Salser, W. (1978) Cold Spring Harbor Symp. Quant. Biol. 42, 985-1002.
34. van Helden, P. D., Strickland, W. N., Strickland, M. and von Holt, C. (1982) Biochim. Biophys. Acta 703, 17-20.
35. Iwai, K., Ishikawa, K. and Hayashi, H. (1970) Nature 226, 1056-1058.
36. Urban, M. K., Franklin, S. G. and Zweidler, A. (1979) Biochemistry 18, 3952-3960.
37. Urban, M. K. and Zweidler, A. (1983) Dev. Biol. 95, 421-428.
38. DeLisle, A. J., Graves, R. A., Marzluff, W. F. and Johnson, L. F. (1983) Mol. Cell. Biol. 3, 1920-1929.
39. Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DeRiel, J. K., Forget, B. G., Weissman, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C. and Proudfoot, N. J. (1980) Cell 21, 653-668.
40. Goldberg, M. L. (1979) Ph.D. Thesis, Stanford University, Palo Alto, California.
41. Proudfoot, N. and Brownlee, G. G. (1976) Nature 263, 211-214.
42. Fitzgerald, M. and Shenk, T. (1981) Cell 24, 251-260.
43. Wang, S.-W., Robins, A. J., D'Andrea, R. and Wells, J.R.E. (1985) Nucl. Acids Res. 13, 1369-1387.