Volume 13 Number 4 1985 Nucleic Acids Research

Inverted duplication of histone genes in chicken and disposition of regulatory sequences

S-W.Wang[1], A.J.Robins, R.d'Andrea and J.R.E.Wells

Adelaide University Centre for Gene Technology, Department of Biochemistry, University of Adelaide, Adelaide, South Australia, Australia 5000, and [1]Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, 5 Dong Dan San Tiao, Beijing, Peoples' Republic of China

Received 10 December 1984; Accepted 22 January 1985

ABSTRACT

Sequence analysis of an 8.4 kb fragment containing five chicken histone genes shows that an H4-H2A gene pair is duplicated and inverted around a central H3 gene. A left and right region, each of 2.1 kb are 97% homologous and the boundaries of homology coincide with ten base pair repeats. These boundary regions also contain highly conserved gene promoter elements, suggesting that interaction of transcriptional machinery with histone genes may be connected with recombination in promoter regions, resulting in the inverted duplication structure seen in this cluster.

INTRODUCTION

Histone genes of invertebrates such as sea urchin and *Drosophila* are, with minor exceptions, clustered in highly ordered tandem arrays and repeated several hundred-fold in their respective genomes, (reviewed in 1). In contrast, core and H1 genes of vertebrates such as chickens, (2,3) mice, (4) and humans (5) are not present in ordered repeats, while the situation in the frog is intermediate with the majority of genes in tandem arrays and others disordered (6,7). In the chicken there are about ten copies of each core histone gene and six H1 genes, thus the number of genes is approximately balanced despite the diversity of gene arrangement (D'Andrea *et al.*, submitted).

Current data suggests that sequences 5' to polymerase II genes are more important in the control of gene expression than the gross organisation of gene families within the genome. The 5' regions contain well characterised elements such as the ubiquitous TATA box and the CAAT box which are important for the regulation of transcriptional initiation. In addition to these general promoter elements, several gene-specific motifs have been recognised and their importance in transcriptional control has been documented (8).

In the sea urchin histone gene system specific elements have been reported upstream from H2A (9) and H4 (10) genes. We previously reported an

H2B-specific promoter sequence (11) which has now been found as an important element for heavy and light chain immunoglobulin gene expression in lymphoid cells (12) and recently a histone H1-specific element has been noted (Coles et al., submitted). If all such elements are shown to be functionally significant, they may account in part for independent regulation of the dispersed histone genes in vertebrate genomes.

Here we investigate the organisation of five chicken histone genes and their associated regulatory elements present in an unusual array in which H2A and H4 genes form an inverted duplication centred around a single H3 gene. Sequence analysis of an 8.4 kb region allows delineation of the borders of the duplication and we note that regulatory elements are intimately associated with these.

MATERIALS AND METHODS

Restriction endonucleases were purchased from New England Biolabs; Klenow Polymerase and α-labelled ^{32}P-dNTPs were purchased from Biotechnology Research Enterprises S.A. Pty. Limited. All enzymes were used according to the specifications of the manufacturer.

Fragment Sub-Cloning

An 8.4 kb EcoRI fragment isolated from the chicken genomic clone λCH03 (D'Andrea et al., submitted) was subcloned into the EcoRI site of pBR325. This recombinant, pCH8.4E, contains five histone genes arranged in the order H4-H2A-H3-H2A-H4 (Figure 1). To ensure that sequences from the left and right regions of the cluster could be derived independently, a unique SalI site within the coding region of the H3 gene was utilised and independent recombinants pCH3.5E/S and pCH4.5E/S were constructed in pBR322.

Sequencing Strategy

Libraries of randomly sonicated fragments of the pCH3.5E/S and pCH4.5E/S inserts were generated in M13mp8 and sequenced by the dideoxy chain termination sequencing method (13). Briefly, each insert was concatamerised by ligation and sonicated under conditions which gave an average fragment length of 700 bp (data not shown). The sheared DNA was blunt-ended using the Klenow fragment of *E. coli* DNA polymerase and size-fractionated on a LGT agarose gel. DNA in the size range 0.5 - 1.5 kb was isolated and subcloned into SmaI digested M13mp8. A library of approximately 200 recombinants was generated in this manner and sequenced by the dideoxy chain termination method. A computer program was used to align overlaps (14).

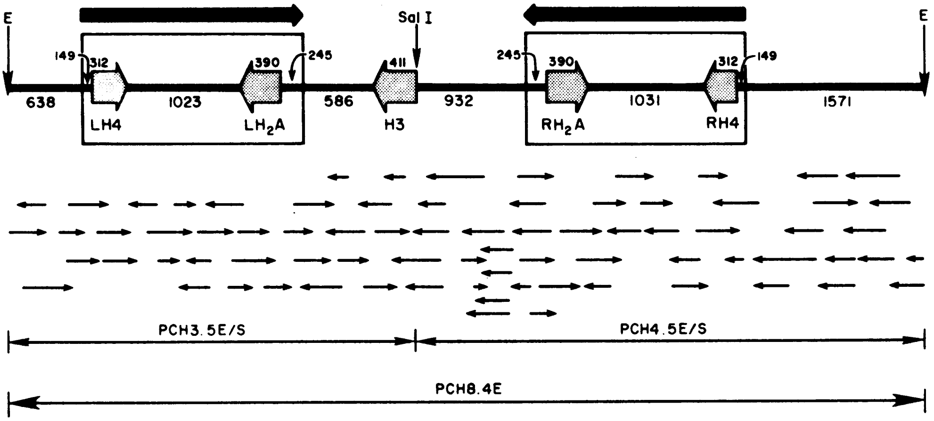

FIGURE 1: Organisation of pCH8.4E. Coding regions of the five histone genes and their direction of transcription are denoted. The LH4-LH2A region and the RH4-RH2A region form part of an inverted duplication shown in boxes. These are referred to in the text and in Figure 3. Numbers represent base-pairs. Small arrows indicate sequences generated from random M13 clones. The subclones pCH3.5E/S and pCH4.5E/S were generated by cloning the two Eco/SalI fragments from pCH8.4E into pBR322.

RESULTS AND DISCUSSION

Complete Sequence of the 8.4 Kb Region

The alignment of the M13 subclones used to generate the whole sequence of the 8.4 kb fragment as well as the relative position and orientation of the five histone genes is shown in Figure 1. The complete nucleotide sequence is presented in Figure 2. Regions of significance are boxed in bold type and designations are explained in the legend.

The H4-H2A Gene-Pairs are the Product of an Inverted Duplication

With complete sequence data, direct comparisons can be made between the left and right H4-H2A gene pairs centred around the single H3 gene. It is expected that each iso-coding region will be highly conserved within the fragment pCH8.4E (and elsewhere in the histone gene locus) whereas non-coding regions may diverge. In the case of the two H4-H2A pairs reported here it is clear that both coding and substantial non-coding regions are highly homologous. Furthermore, the boundaries of homologous and non-homologous bases for the left and right gene pair are easily defined. As shown by the boxes in Figure 1, the homology extends 149 base pairs 5' to the H4 gene initiation codon and 245 base pairs 5' to the H2A gene initiation codon. Omitting the coding regions in each case, the remaining thousand or so base-pairs within each intergene region show 97% homology. If coding regions are now included there is a block of approximately 2.1 kb containing an H4 and H2A gene which has been duplicated in reverse orientation around the H3 gene (Figure 1). The degree of conservation of sequences suggests that the recombination event is relatively recent.

The Boundaries of the Duplication Contain Repeats

The boundaries of the duplication are characterised by a ten base pair direct repeat 5' to the H4 gene and a ten base pair inverted repeat 5' to the H2A gene. The direct and inverted repeats are separated by forty and twenty-one base pairs of DNA respectively and are related to each other sharing the octamer sequence 5' GCCCCGCC 3' (Figure 3). Immediately adjacent to the inverted repeat upstream from the right H2A gene (RH2A - Figure 1) and outside the duplicated area is another inverted repeat of unrelated sequence. These seven base-pair sequences are twenty-one base-pairs apart and are shown in bold type as regions IR(1) at positions, 4,652 and 4,680 in Figure 2. The coincidence of repeated sequences immediately adjacent to the boundaries of the inversion suggests involvement of these elements in recombination events giving rise to the inverted duplication. One mechanism which may account for this is reverse chromatid pairing as suggested by Vitelli and Weinberg (15). They have reported an inverted duplication in an unusual sea urchin histone gene cluster and the presence of short direct repeats at one of the boundaries of the duplication. In the chicken histone gene cluster, the inverted repeats 5' to the H2A genes (IR(L), positions 2,717 and 2,757; IR(R) 4,687 and 4,727, Fig. 2) are themselves made up of a direct repeat (5' CCGCCCCGCC 3'). Although not commented on by Vitelli and Weinberg (op. cit.), the sea urchin inverted duplication also contains an imperfect inverted repeat close to one of the boundaries of the duplication. The repeats may have played a role in the generation of the chromosomal inverted duplications seen in both systems.

The Five Gene Coding Regions

None of the five genes contains intervening sequences and each encodes a protein with the same amino acid sequence as the respective calf thymus histone (16).

The DNA sequence of the left H4 gene (LH4 - Figure 1) has been previously reported, (17). We find 17 differences between our sequence and the reported one but none of these changes affect the coding potential of the gene. In addition, Sugarman et al. (17) assigned an EcoRI site just 3' to the end of the gene which is not present in our sequence. These differences may represent polymorphisms although the extra EcoRI site 3' to the gene would

```
ATTCATGGACATTTAAAATCATGTTTGCTGTATTTGAGCCTTTAAAAAACATATATTGTTAAAAATTACTTCTTACG
TAAGTACCTGTAAATTTTAGTACAAACGACATAAACTCGGAAATTTTTTGTATATAACAATTTTTAATGAAGAATGC

                      ↓
GTTTATGCAATTTATCATTTATTCTTAGATATAATGTTGTTTGCAATGGCTTGGGTTGCATATGCAGGCCTAACTGC
CAAATACGTTAAATAGTAAATAAGAATCTATATTACAACAAACGTTACCGAACCCAACGTATACGTCCGGATTGACG

                                              ↓
ATAATAATAATAATAATAATAAAATAATCGAAGGGAACAAAGTCATACTGTTCTTAGGGATCAGTACACTAAACATG
TATTATTATTATTATTATTATTTTATTAGCTTCCCTTGTTTCAGTATGACAAGAATCCCTAGTCATGTGATTTGTAC

                                                                      ↓
CGTAAGGCAACTTATATACTGTATTGACATCTTCGGATTGTACGTACAGTGCCCACCATCAAAAGAGCAAGACATGA
GCATTCCGTTGAATATATGACATAACTGTAGAAGCCTAACATGCATGTCACGGGTGGTAGTTTTCTCGTTCTGTACT

AGAGTTGAAAGACTTGGTGCTACTCTAAATGGAAAGCGAATTAAAGAAAATTCCCTATGGTTTGACACCAGTTAGAG
TCTCAACTTTCTGAACCACGATGAGATTTACCTTTCGCTTAATTTCTTTTAAGGGATACCAAACTGTGGTCAATCTC

              ↓
GAAAAAAAAACCTCACCTTGGAACTACAAGGTTACATATACACAGGAATTGTTACATGACTATTGAAAAATATTACG
CTTTTTTTTTGGAGTGGAACCTTGATGTTCCAATGTATATGTGTCCTTAACAATGTACTGATAACTTTTTATAATGC
                                      500
                                       ↓
ATACACGGAAAGCATTCAAGAAAATAAAGAGGAGCAATACAAAAAGAAAGTGCAATAGATTTCTCAGGTAGTGTACT
TATGTGCCTTTCGTAAGTTCTTTTATTTCTCCTCGTTATGTTTTTCTTTCACGTTATCTAAAGAGTCCATCACATGA

                                                             ↓
TTACACCGAAAAAAAAAAAAAGTGATGAGAGAAACAGACTGTAATGAGGAAAATGAAACGCATTGCAGCCAGGAGAA
AATGTGGCTTTTTTTTTTTTTCACTACTCTCTTTGTCTGACATTACTCCTTTTACTTTGCGTAACGTCGGTCCTCTT
                          DR(L)
                       ──────────►                             H4 I
AAAAAAAAATTAGCTAGGAGGA |GGCCCCGCCC| TCCAGGGAGGAGGAGGCAGTGGCGC |TCCCGCC| CCCGCCTG
TTTTTTTTTAATCGATCCTCCT |CCGGGGCGGG| AGGTCCCTCCTCCGTCACCGCG    |AGGGCGG| GGGCGGAC
   DR(L)
──────────►   ↓     H4 II       H4 III             TATA
|GGCCCCGCCC| CTGGT |TTCA| ATCAG |GTCC| GACCATACGCC |ATAACA| CCCGCGCGCGCCCCGCCACATCCTC
|CCGGGGCGGG| GACCA |AAGT| TAGTC |CAGG| CTGGTATGCGG |TATTGT| GGGCGCGCGCGGGGCGGTGTAGGAG

                                               ↓
ACTGGTGTCGGACGACTCAGGCTCTCGGC ATG TCT GGC AGA GGC AAG GGC GGG AAG GGG CTC GGC
TGACCACAGCCTGCTGAGTCCGAGAGCCG     ser gly arg gly lys gly gly lys gly leu gly

AAG GGC GGC GCC AAG CGC CAC CGC AAG GTG CTG CGC GAC AAT ATC CAG GGC ATC ACC
lys gly gly ala lys arg his arg lys val leu arg asp asn ile gln gly ile thr

                         ↓
AAG CCG GCC ATC CGC CGC CTG GCG CGG CGC GGC GGC GTC AAG CGC ATC TCG GGG CTC
lys pro ala ile arg arg leu ala arg arg gly gly val lys arg ile ser gly leu

ATC TAC GAG GAG ACG CGC GGC GTG CTC AAG GTC TTC CTG GAG AAC GTC ATC CGC GAC
ile tyr glu glu thr arg gly val leu lys val phe leu glu asn val ile arg asp
    1000
     ↓
GCC GTC ACC TAC ACC GAG CAC GCC AAG AGG AAG ACG GTC ACG GCC ATG GAC GTG GTC
ala val thr tyr thr glu his ala lys arg lys thr val thr ala met asp val val
```

FIGURE 2

```
                                                                 ↓
TAC GCG CTC AAG CGC CAG GGA CGC ACC CTC TAC GGC TTC GGC GGT TAA  ACTCGTCTCCGAT
tyr ala leu lys arg gln gly arg thr leu tyr gly phe gly gly stop TGAGCAGAGGCTA

                                          T                               DSE
TCCGGCCACCCGAACTCGTTTTTAGCA |ACCCAAAGGCTCTTTTCAGAGCCGCCCA| CTTGGTTC |CAACAAAGAGCT|
AGGCCGGTGGGCTTGAGCAAAAATCGT |TGGGTTTCCGAGAAAAGTCTCGGCGGGT| GAACCAAG |GTTGTTTCTCGA|

             ↓
GTGTCACCTCGCCTGATGTGACGGGGCTTTTTCACTTAATAGTTAGGCTCTTTTATCTCCCCCAGCCGTATTTTCAG
CACAGTGGAGCGGACTACACTGCCCCGAAAAAGTGAATTATCAATCCGAGAAAATAGAGGGGGTCGGCATAAAAGTC

                                   ↓
CTCTTCGCTTTCCGTGTCCGGACTGCAGAGCTGTGTAGCACAGCTCTAGCGCCTCGCGGCGTCCCTCCGTTGCCTCG
GAGAAGCGAAAGGCACAGGCCTGACGTCTCGACACATCGTGTCGAGATCGCGGAGCGCCGCAGGGAGGCAAGGGAGC

                                                          ↓
CAAGCGGCTCGCTGCCTCCGCTTCCCCGCAGCAGGCTCTCACCGGGCAGCTTTCGGGCTCCGCTCGCTCCAGGGCGC
GTTCGCCGAGCGACGGAGGCGAAGGGGCGTCGTCCGAGAGTGGCCCGTCGAAAGCCCGAGGCGAGCGAGGTCCCGCG

GCTCTTGACTTCTATTCCCGTTGCTAGCGACCGGGTCAGGCACGTGCGACACAGTGCCCAGGCGCGAATCCAGCCCC
CGAGAACTGAAGATAAGGGCAACGATCGCTGGCCCAGTCCGTGCACGCTGTGTCACGGGTCCGCGCTTAGGTCGGGG
  1500
    ↓
TCCCTTCGGCACCGCGTCCGGAAAGGCCAGGAGAAGGGGGCCGGCCCCATTACTCACAGACCGGGGGAAGGGCGAAG
AGGGAAGCCGTGGCGCAGGCCTTTCCGGTCCTCTTCCCCCGGCCGGGGTAATGAGTGTCTGGCCCCCTTCCCGCTTC

                           ↓
GAGAAGCAGGTTCTGCCCTGACAGGAACCCCCCGAGGTGGGGGGAGAAATGGGGAAGAGGCGGCTATTCCGCTTCCG
CTCTTCGTCCAAGACGGGACTGTCCTTGGGGGGCTCCACCCCCCTCTTTACCCCTTCTCCGCCGATAAGGCGAAGGC

                                                  ↓
CCCGCCCGCAGGCCGCGCACGCCGGGCTGCGCGGCAAGAGGCGGACACCACAACGGACCCCGCGGTGTCCGCTCCTT
GGGCGGGCGTCCGGCGCGTGCGGCCCGACGCGCCGTTCTCCGCCTGTGGTGTTGCCTGGGGCGCCACAGGCGAGGAA

                                                                     ↓
TTCCCCCAACTGGGCGTTTCGCCCGCCCTTACGAAAGGCGCGCGGGGCGCCCACGCGGACGCGGATGCGCGGC
AAGGGGGTTGACCCGCAAAGCGGGCGGGAATGCTTTCCGCGCGCCCCGCGGGTGCGCCCTGCGCCTACGCGCCG

TGCCGCTCGAGAGGGGGACGGGGAGCGGCTTCAGCGCAGCGTTTCGGGTTGCAGTCAGGACAGGGCAGCGACGCCCA
ACGGCGAGCTCTCCCCCTGCCCCTCGCCGAAGTCGCGTCGCAAAGCCCAACGTCAGTCCTGTCCCGTCGCTGCGGGT

                   ↓
GGGGAAAACTCGGGGCTTCCCGAACGAACACAGCGCCTCTCCGGCGGAGCTGCGAGGGCCGCACGGACGGAAAACCC
CCCCTTTTGAGCCCCGAAGGGCTTGCTTGTGTCGCGGAGAGGCCGCCTCGACGCTCCCGGCGTGCCTGCCTTTTGGG
                                        2000
                                          ↓
CCCAGTGCCGCCGTGCGCCCAGCAGTACGGAACCTCCCGCAACAACACACCACACAAGCACTCGTCGAGTCACCGCA
GGGTCACGGCGGCACGCGGGTCGTCATGCCTTGGAGGGCGTTGTTGTGTGGTGTGTTCGTGAGCAGCTCAGTGGCGT

           DSE                            T                               ↓
TTCCG |AGCTCTTTCTCG| AGACTG |TGGGTGGCTCTGAAAAGAGCC TTTGGGT| TTCACTTCGCTGCTCGGCGCG
AAGGC |TCGAGAAAGAGC| TCTGAC |ACCCACCGAGACTTTTCTCGG AAACCCA| AAGTGAAGCGACGAGCCGCGC

CTTCCTCGGGGCCGGC       stop lys ala lys ala lys his ser asp thr lys lys pro
GAAGGAGCCCCGGCCG        GAT GAA TCG GAA CCG GAA CAC CGA CAG CCA GAA GAA CCC
```

```
                                                  ↓
leu leu val ala gln ile asn pro leu val gly gly gln ala ile thr val lys gly
GTC GTC GTG CCG GAC CTA CAA CCC GTC GTG GGG TGG GAC GCG CTA CCA GTG GAA CGG

leu leu lys asn leu glu glu asp asn arg ile ala leu gln leu his arg pro ile
GTC GTC GAA CAA CTC GAG GAG CAG CAA CGC CTA CCG GTC GAC GTC CAC CGC CCC CTA

                                ↓
ile arg thr lys lys asn asp arg ala ala asn gly ala leu glu leu ile glu ala
CTA CGC GCA GAA GAA CAA CAG CGC CCG GCG CAA CGG GCG GTC GAG GTC CTA GAG CCG

thr leu tyr glu leu val ala ala leu tyr val pro ala gly ala gly val arg glu
GCA GTC CAT GAG GTC GTG CCG GCG GTC CAT GTG GCC CCG CGG CCG CGG GTG GGC GAG

             ↓
ala tyr asn gly lys arg leu leu arg his val arg gly val pro phe gln leu gly
GCG CAT CAA CGG GAA CGC GTC GTC GGC CAC GTG CGC CGG GTG CCC CTT GAC GTC GGG
                                                                   2500
                                                                     ↓
ala arg ser ser arg ser lys ala lys ala arg ala lys gly gly gln lys gly arg
CCG GGC GCT GCT CGC GCT GAA CCG GAA CCG CGC GCG GAA GGG CGG GAC GAA GGG CGC

gly ser      CGCCGATCAGTCGCTCAGAAGCTCAACTGCCAACTCAACAGAAATCAAAACACGCAACAGCTGC
GGG GCT GTA  GCGGCTAGTCAGCGAGTCTTCGAGTTGACGGTTGAGTTGTCTTTAGTTTTGTGCGTTGTCGACG

                  TATA↓                                     H2AC2
CCGCGCTCGCCCGCGCC |TTTATAT| CCCTTCCCCGTTGCGCGCCCGCCGGCTCC |TGATAGGC| GGACGTGTCC
GGCGCGAGCGGGCGCGG |AAATATA| GGGAAGGGGCAACGCGCGGGCGGCCGAGG |ACTATCCG| CCTGCACAGG

                    H2AC1                         ↓
TCACGGGGAACGAGCGC |CCAATGGC| GTAGCGAATCTCGGCCCGACCAATAGCGACGGGCGCCCTTCC
AGTGCCCCTTGCTCGCG |GGTTACCG| CATCGCTTAGAGCCGGGCTGGTTATCGCTGCCCGCGGAAGG
   IR(L)
──────────────►        H2A S.S.              IR(L)
|CCGCCCCG |CC| CCTCTCTCTGCTCACAGCA| AT |GGCGGGGCGG| AAGGGGCGGGAGCAAGGGGGCATTCCGCCCA
|GGCGGGGC |GG| GGAGAGAGACGAGTGTCGT| TA |CCGCCCCGCC| TTCCCCGCCCTCGTTCCCCCGTAAGGCGGGT
                                        ◄──────────
           ↓
CTAACTGAGCGTGTGCTGCGGGCGGGTGTGGGGCGGTCAATGAGAGCGGTTACGGCTGCTCCGGCCTTTTTCTCTAC
GATTGACTCGCACACGACGCCCGCCCACACCCCGCCAGTTACTCTCGCCAATGCCGACGAGGCCGGAAAAAGAGATG

                                  ↓
ATCTCTTATTTATTTTGTTGATCTGTTTTTTTAAACAGTTGCCAAGGGCCCGAGCACGGCACAGTCAGGGGGAGAGG
TAGAGAATAAATAAAACAACTAGACAAAAAAATTTGTCAACGGTTCCCGGGCTCGTGCCGTGTCAGTCCCCCTCTCC
                                                       3000
                                                         ↓
GATTCATAAAAGAGGGTGTTGAGCGTTTGCGTCATAGCGAGCCGAGGGGTGGGACATTTCGGACGGCCCCTCTGAAA
CTAAGTATTTTCTCCCACAACTCGCAAACGCAGTATCGCTCGGCTCCCCACCCTGTAAAGCCTGCCGGGGAGACTTT

ATTAACCGTTCTAGGTTGCCTTCCTGGAAGTAAACATCACAGTTCTACTTCCATGCCTAAATTAATCACTTGAAACG
TAATTGGCAAGATCCAACGGAAGGACCTTCATTTGTAGTGTCAAGATGAAGGTACGGATTTAATTAGTGAACTTTGC

   ↓
AAAAGCTAATACAGATGGCACTCTGCTAGCGACTCTGTATACATACATAGTACATTGACTAGCGTTTCGTTCTTTGC
TTTTCGATTATGTCTACCGTGAGACGATCGCTGAGACATATGTATGTATCATGTAACTGATCGCAAAGCAAGAAACG
```

```
                                   ↓
TTGTATATAAGAATCTCAGATTGGCTTCGATAATTGTATAAATTTGTGCCTATTTCAGTTACATTGCAGCGTTAC
AACATATATTCTTAGAGTCTAACCGAAGCTATTAACATATTTAAACACGGATAAAGTCAATGTAACGTCGCAATG

      DSE                           T                  ↓
|AGCTCGTTTTTATTG| TAGA |TGGAGTGGCTCTTAAAAGAGCCTTT TGGGT| TGATTTAAGTAGACGTTTAAAGAT
|TCGAGCAAAAATAAC| ATCT |ACCTCACCGAGAATTTTCTCGGAAA ACCCA| ACTAAATTCATCTGCAAATTTCTA

AACTTTCGAGGCTTAACTTTCT stop ala arg glu gly arg ile arg arg ala leu gln ile
TTGAAAGCTCCGAATTGAAAGA  AAT ACG GGA GAG GGG CGC CTA CGC CGC CCG CTC GAC CTA

                          ↓
asp lys pro met ile thr val arg lys ala his ile ala cys leu asn thr asp glu
CAG GAA CCC GTA CTA CCA CTG CGC GAA CCG CAC CTA CCG CGT GTC CAA CCA CAG GAG

phe leu gly val leu tyr ala glu ser ala glu gln leu ala met val ala ser ser
CTT CTC GGG GTG GTC CAT CCG GAG CGA GCG GAG GAC GTC GCG GTA CTG CCG GCT CGA
   3500
    ↓
gln phe arg leu asp thr lys phe asp gln ala ile glu arg val leu arg gln phe
GAC CTT CGC GTC CAG CCA GAA CTT CAG GAC GCG CTA GAG GGC GTG GTC CGC GAC CTT

                                                         ↓
pro leu lys arg ile leu leu glu thr ser lys gln tyr arg arg ile glu arg leu
CCC GTC GAA CGC CTA GTC GTC GAG GCA CCT GAA GAC CAT CGC GGC CTA GAG CGC GTC

ala val thr gly pro arg tyr arg his pro lys lys val gly gly thr ala pro ala
GCG GTG GCA CGG CCC GGC CAT CGC CAC GCC GAA GAA GTG CGG CGG GCA CCG GCC GCG

                                         ↓
ser lys arg ala ala lys thr ala leu gln lys arg pro ala lys gly gly thr ser
CGA GAA CGC CCG GCG GAA CCA GCG CTC GAC GAA CGC CCC GCG GAA GGG CGG GCA GCT

lys arg ala thr gln lys thr arg ala      CGCTTAAGAACGCAAGCAGGCAGGGAGGAACGCCAG
GAA TGC GCG GCA GAC GAA GCA TGC GCG GTA  GCGAATTCTTGCGTTCGTCCGTCCCTCCTTGCGGTC

         ↓                          TATA                H3C5
GAAATAGACGCCGCTCTCCGCTGTGGGGAT |ATTTATA| CTAGCACCGCTC |ATTCTGATTG| GCCGAAACAAACAG
CTTTATCTGCGGCGAGAGGCGACACCCCTA |TAAATAT| GATCGTGGCGAG |TAAGACTAAC| CGGCTTTGTTTGTC

        H3C4                  ars-like          ↓
GCTG |AATCTCATTG| GTCGA |TTCGAAGTTTAAAATAA| CCGCCTTTTGGAACGTACCGGCAACCACCG
CGAC |TTAGAGTAAC| CAGCT |AAGCTTCAAATTTTATT| GGCGGAAAACCTTGCATGGCCGTTGGTGGC

                                                                                  4000
    H3C3             H3C2                H3C1                                      ↓
|TCTCTCATTG| T |TGTCTCATTG| CAG |TGTCTCATTG| AGCACACAGAAACTTTTTTTTTTTTTTTGCTTTGTTTT
|AGAGAGTAAC| A |ACAGAGTAAC| GTC |ACAGAGTAAC| TCGTGTGTCTTTGAAAAAAAAAAAAAAACGAAACAAAA

CTTTTTTATCCTTTTGGTTTCTTTTAAGGTAGAAATTTATTCTCGCATTTCTGGAAAGCACTCATGTACGCTAGATT
GAAAAAATAGGAAAACCAAAGAAAATTCCATCTTTAAATAAGAGCGTAAAGACCTTTCGTGAGTACATGCGATCTAA

                         ↓
CAGTAGAGAGACAATAAGCGGTAGATTGCAAAGCCTTCCATGCTGAAGTCAGTAATGAGAAAAATAAGAGCAAAAAT
GTCATCTCTCTGTTATTCGCCATCTAACGTTTCGGAAGGTACGACTTCAGTCATTACTCTTTTTATTCTCGTTTTTA
```

```
                                            ↓
ATGTTTTGTTAAGGAAAGATATCTAAAATAGGTTCCGATTATGTGCCTATACTGAATTACGAGGAATACATTACCTC
TACAAAACAATTCCTTTCTATAGATTTTATCCAAGGCTAATACACGGATATGACTTAATGCTCCTTATGTAATGGAG

                                                                   ↓
AGGAAATGAGGATCATTAGGTTTTTCATAACTACTGTTGATTTCTTTGAGAAGTGCTGTTTAGAAAGGAGGGTGGGG
TCCTTTACTCCTAGTAATCCAAAAAGTATTGATGACAACTAAAGAAACTCTTCACGACAAATCTTTCCTCCCACCCC

GTCTCTTCATGGTACATGGAAGAGTGGGTGCCCATATAACATGACCGTGTGGCAACCGCGTAAAGTTCCCCTGTGGG
CAGAGAAGTACCATGTACCTTCTCACCCACGGGTATATTGTACTGGCACACCGTTGGCGCATTTCAAGGGGACACCC

            ↓
GAAAAGTCCCAGGACGCCTTACGGGGGCACAGTGAGGGGAGCTCATTATGACAGAAATTCCGTGCACTCAGAGACAC
CTTTTCAGGGTCCTGCGGAATGCCCCCGTGTCACTCCCCTCGAGTAATACTGTCTTTAAGGCACGTGAGTCTCTGTG
                                   4500
                                    ↓
ACGTGCGCACAAGTCCCGTTCTCCCCACAGGCCGCTCATCCTTACAAATAACTCTTTCCCTCAGCTACAGTTTCCCT
TGCACGCGTGTTCAGGGCAAGAGGGGTGTCCGGCGAGTAGGAATGTTTATTGAGAAAGGGAGTCGATGTCAAAGGGA

                                                           ↓
CACTCTGCAGTAGAAGGGCAAGGAAAAAAGAAAGAAAAAGAAAGAAAGAAAAAACGGAAAAAAAGACCCGGTTCAGA
GTGAGACGTCATCTTCCCGTTCCTTTTTTCTTTCTTTTTCTTTCTTTCTTTTTTGCCTTTTTTTCTGGGCCAAGTCT
                                     IR(1)                          IR(1)
                                    ──────→                        ←──────
CCCCCCGGAAGGACCGGGACATGAGACTGCCCGT GAGCTGC TGCCGCCAGTGGGCAACCACC GCAGCTC
GGGGGGCCTTCCTGGCCCTGTACTCTGACGGGCA CTCGACG ACGGCGGTCACCCGTTGGTGG CGTCGAG
   IR (R)          ↓       H2A S.S.          IR (R)
──────────→
CCGCCCCGCC AT TGCTGTGAGCAGAGAGAGG GG CGGGGCGG GGAAGGGCGCCCGTCGCTATTGGTCGGGCCG
GGCGGGGCGG TA ACGACACTCGTCTCTCTCC CC GCCCCGCC CCTTCCCGCGGGCAGCGATAACCAGCCCGGC
                                  ←──────────
             H2AC1                                     H2AC2
                                          ↓
AGATTCGCTAC GCCATTGG GCGCTCGTTCCCCGTGAGGACAGCTCC GCCTATCA GGAGCCGGCGGGCGCGCAA
TCTAAGCGATG CGGTAACC CGCGAGCAAGGGGCACTCCTGTCGAGG CGGATAGT CCTCGGCCGCCCGCGCGTT

             TATA                                                      ↓
CGGGGAAGGG ATATAAA GGCGCGCGGGCGAGCGCGGGCAGCTGTTGCGTGTTTTGATTTCTGTTGAGTTGGCAGT
GCCCCTTCCC TATATTT CCGCGCGCCCGCTCGCGCCCGTCGACAACGCACAAAACTAAAGACAACTCAACCGTCA

TGAGCTTCTGAGCGACTGATCGGCG   ATG TCG GGG CGC GGG AAG CAG GGC GGG AAG GCG CGC
ACTCGAAGACTCGCTGACTAGCCGC       ser gly arg gly lys gln gly gly lys ala arg
                                               5000
                                                ↓
GCC AAG GCC AAG TCG CGC TCG TCG CGG GCC GGG CTG CAG TTC CCC GTG GGC CGC GTG
ala lys ala lys ser arg ser ser arg ala gly leu gln phe pro val gly arg val

CAC CGG CTG CTG CGC AAG GGC AAC TAC GCG GAG CGG GTG GGC GCC GGC GCC CCG GTG
his arg leu leu arg lys gly asn tyr ala glu arg val gly ala gly ala pro val

                        ↓
TAC CTG GCG GCC GTG CTG GAG TAC CTG ACG GCC GAG ATC CTG GAG CTG GCG GGC AAC
tyr leu ala ala val leu glu tyr leu thr ala glu ile leu glu leu ala gly asn

GCG GCC CGC GAC AAC AAG AAG ACG CGC ATC ATC CCC CGC CAC CTG CAG CTG GCC ATC
ala ala arg asp asn lys lys thr arg ile ile pro arg his leu gln leu ala ile
```

```
    ↓
CGC AAC GAC GAG GAG CTC AAC AAG CTG CTG GGC AAG GTG ACC ATC GCG CAG GGC GGG
arg asn asp glu glu leu asn lys leu leu gly lys val thr ile ala gln gly gly

                                                              ↓
GTG CTG CCC AAC ATC CAG GCC GTG CTG CTG CCC AAG AAG ACC GAC AGC CAC AAG GCC
val leu pro asn ile gln ala val leu leu pro lys lys thr asp ser his lys ala

AAG GCT AAG TAG   GCCGGCCCCGAGGAAGCGCGCCGAGCAGCGAAGTGAA
lys ala lys stop  CGGCCGGGGCTCCTTCGCGCGGCTCGTCGCTTCACTT

              T                          DSE↓
|ACCCAAAGGCTCTTTTCAGAGCCACCCA| CAGTCT |CGAGAAAGAGCT| CGGAATGCGGCGACTCGACAGTGCTTGT
|TGGGTTTCCGAGAAAAGTCTCGGTGGGT| GTCAGA |GCTCTTTCTCGA| GCCTTACGCCGCTGAGCTGTCACGAACA
                                                                 5500
                                                                   ↓
GTGGTGTGTTGTTGCCGCGGGAGGTTCCGTACTGCTGGGCGCACGGCGGCACTGGGGGGTTTTTCCGTCCGTGCGGG
CACCACACAACAACGGCGCCCTCCAAGGCATGACGACCCGCGTGCCGCCGTGACCCCCCAAAAAGGCAGGCACGCCC

CCTCGCAGCTCCGCCGGAGAGGCGCTGTGTTCGTTCGGGAAGCCCCGAGTTTTCCCCTGGGCGTCGCTGCCCTGTCC
GGAGCGTCGAGGCGGCCTCTCCGCGACACAAGCAAGCCCTTCGGGGCTCAAAAGGGGACCCGCAGCGACGGGACAGG

             ↓
TGACTGCAACCCGAAACGCTGCGCTGAAGCCGCTCCCCGTCCCCCTCTCGAGCGGCAGCCGCGCATCCGCGTGCCGC
ACTGACGTTGGGCTTTGCGACGCGACTTCGGCGAGGGGCAGGGGGAGAGCTCGCCGTCGGCGCGTAGGCGCACGGCG

                                  ↓
GTGGCGGCCCCGCGCGCCTTTCGTAAGGGCGGGCGAAACGCCCGGAAGGGGGAAAAGGAGCGGACACCGCGGGG
CACCGCCGGGGCGCGCGGAAAGCATTCCCGCCCGCTTTGCGGGCCTTCCCCCTTTTCCTCGCCTGTGGCGCCCC

                                                           ↓
TCCGTTGTGGTGTCCGCCTCTTGCCGCGCAGCCCGGCGTGCGCGGCCTGCGGGCGGGCGGAACGGGAATAGCCGCCT
AGGCAACACCACAGGCGGAGAACGGCGCGTCGGGCCGCACGCGCCGGACGCCCGCCCGCCTTGCCCTTATCGGCGGA

CTTCCCCATTTCTCCCCCCACCTAGAGAGGGTTCCTGTCAGGGCAGAACCTGCTTCTGCCTTCGCCCTTCCCCCGGT
GAAGGGGTAAAGAGGGGGGTGGATCTCTCCCAAGGACAGTCCCGTCTTGGACGAAGACGGAAGCGGGAAGGGGGCCA

     ↓
CTGTGAGTAATGGGGCCGGCCCCCTTCTCCCGGCCTTTCCGGACGCGGTGCCGAAGGGAGGGGCTGGATTCGCGCCT
GACACTCATTACCCCGGCCGGGGGAAGAGGGCCGGAAAGGCCTGCGCCACGGCTTCCCTCCCCGACCTAAGCGCGGA
                          6000
                            ↓
GGGCACTGTGTCGCACGTGCCTGACCCGGTCGCTAGCAACGGGAATAGAAGTCAAGAGCGCGCCCTGGAGCGAGGGC
CCCGTGACACAGCGTGCACGGACTGGGCCAGCGATCGTTGCCCTTATCTTCAGTTCTCGCGCGGGACCTCGCTCCCG

                                                  ↓
AGCCCGAAAGCTGCCCGGTGAGAGCCTGCTGCGGGAAGCGGAGGCAGCGAGCCGCTTGCGAGGGAACGGAGGGACG
TCGGGCTTTCGACGGGCCACTCTCGGACGACGCCCCTTCGCCTCCGTCGCTCGGCGAACGCTCCCTTGCCTCCCTGC

                                                                          ↓
CCGCGAGGCGCTAGAGCTGAGCTACACAGCTCTGCAGTCCGGACACGGAAAGCAAAGAACTGAAAATGCGGCTGGGG
GGCGCTCCGCGATCTCGACTCGATGTGTCGAGACGTCAGGCCTGTGCCTTTCGTTTCTTGACTTTTACGCCGACCCC

                                                                   DSE
GAGATAAAAGAGCCTAACTATTAAGTAAAAGAAAAGCCCCGTCACATCAGGCGAGGTGACAC |AGCTCTTTGTTG| G
CTCTATTTTCTCGGATTGATAATTCATTTTCTTTTCGGGGCAGTGTAGTCCGCTCCACTGTG |TCGAGAAACAAC| C
```

```
                          T  ↓
AACCAAG TGGGCGGCTCTGAAAAGAGCCTTTGGGTT GCTAAACACGAGTTCGGGTGGCCGGAATCGGAGACGAGT
TTGGTTC ACCCGCCGAGACTTTTCTCGGAAACCCAA CGATTTGTGCTCAAGCCCACCGGCCTTAGCCTCTGCTCA

                                                                          ↓
   stop gly gly phe gly tyr leu thr arg gly gln arg lys leu ala tyr val val
    AAT TGG CGG CTT CGG CAT CTC CCA CGC AGG GAC CGC GAA CTC GCG CAT CTG GTG

asp met ala thr val thr lys arg lys ala his glu thr tyr thr val ala asp arg
CAG GTA CCG GCA CTG GCA GAA GGA GAA CCG CAC GAG CCA CAT CCA CTG CCG CAG CGC
                                                6500
                                                 ↓
ile val asn glu leu phe val lys leu val gly arg thr glu glu tyr ile leu gly
CTA CTG CAA GAG GTC CTT CTG GAA CTC GTG CGG CGC GCA GAG GAG CAT CTA CTC GGG

ser ile arg lys val gly gly arg arg ala leu arg arg ile ala pro lys thr ile
GCT CTA CGC GAA CTG CGG CGG CGC GGC GCG GTC CGC CGC CTA CCG GCC GAA CCA CTA

                             ↓
gly gln ile asn asp arg leu val lys arg his arg lys ala gly gly lys gly leu
CGG GAC CTA CAA CAG CGC GTC GTG GAA CGC CAC CGC GAA CCG CGG CGG GAA CGG CTC

                                                                                     ↓
gly lys gly gly lys gly arg gly ser      GCCGAGAGCCTGAGTCGTCCGACACCAGTGAGGATG
GGG GAA GGG CGG GAA CGG AGA CGG TCT GTA  CGGCTCTCGGACTCAGCAGGCTGTGGTCACTCCTAC

                     TATA                 H4 III       H4 II          DR (R)
TGGCGGGGCGCGCGCGGG TGTTAT GGCGTATGGTC GGAC CTGAT TGAA ACCAG GGGCGGGGCC
ACCGCCCCGCGCGCGCCC ACAATA CCGCATACCAG CCTG GACTA ACTT TGGTC CCCGCCCCGG
                                                             ←

         H4 I                             ↓       DR (R)
CAGGCGGG GGCGGGA GCGCCACTGCCTCCTCCTCCCTGGT GGGCGGGACC TTCGGCTCTCGCCGCTCTCTGTC
GTCCGCCC CCGCCCT CGCGGTGACGGAGGAGGAGGGACCA CCCGCCCTGG AAGCCGAGAGCGGCGAGAGACAG
                                                  ←
                                    IR (2)
                                      →                             ↓
ACCCAGGGCCTCAGCCTGGTTCGCCCTCCCCATCCCG CCTCCTCCC TTAAGTTCCCTTTCCCAGCATCTCTCGCT
TGGGTCCCGGAGTCGGACCAAGCGGGAGGGGTAGGGC GGAGGAGGG AATTCAAGGGAAAGGGTCGTAGAGAGCGA

      IR (2)
CTCGT GGGAGGAGG TGCTTCCGTCCTCATTCCTCTCCAACTCAGTGTCTCCCCGTGAGGATTAGTGGTGTTTTTT
GAGCA CCCTCCTCC ACGAAGGCAGGAGTAAGGAGAGGTTGAGTCACAGAGGGGCACTCCTAATCACCACAAAAAA
          ←
           7000
             ↓
GTTTCTTTTTTTTACGGTAGTAAAATGATACGTTGATATACTTTATTGCTCTTACTATTTACATTTCTATGTTATTT
CAAAGAAAAAAAATGCCATCATTTTACTATGCAACTATATGAAATAACGAGAATGATAAATGTAAAGATACAATAAA

                                       ↓
TTCTATTTAGTTAATATTTATGTCGCATACCTTTGTTACTCTTATCCAGCTTGTTTAATTTCTGGGAATTTTGTGAA
AAGATAAATCAATTATAAATACAGCGTATGGAAACAATGAGAATAGGTCGAACAAATTAAAGACCCTTAAAACACTT

                                                            ↓
ATTTTTTGTACTTTATTTTCAAAATCTCAAAAAATGAAGTGGACAGAAAGGCCAACCTAATGTACTTTTTAGGCAGC
TAAAAAACATGAAATAAAAGTTTTAGAGTTTTTTACTTCACCTGTCTTTCCGGTTGGATTACATGAAAAATCCGTCG

CTAATCTTGAAAAAATAACATTTGCTTTGTATGGAATTAGGGCTATGGTCTGATTTCATATTACACTGAAGAAAACA
GATTAGAACTTTTTTATTGTAAACGAAACATACCTTAATCCCGATACCAGACTAAAGTATAATGTGACTTCTTTTGT
```

```
     ↓
GAACCCCTGGAGATTTCTATGTCATTTATTGCCCAAATTATTTGATTTTTAGTAGAATCTATTGACTAAGCTTTAAA
CTTGGGGACCTCTAAAGATACAGTAAATAACGGGTTTAATAAACTAAAAATCATCTTAGATAACTGATTCGAAATTT

                            ↓
TGATTTCATAATTATATGACCCCTTTGTTGTAGGACATAAATTTGCTGCTTCTTGTGGACTGTGGTGTAGGTGACAC
ACTAAAGTATTAATATACTGGGGAAACAACATCCTGTATTTAAACGACGAAGAACACCTGACACCACATCCACTGTG
                                                 7500
                                                   ↓
TTGTTCCTATTTATTATGTGCACTGTTTTTCACGTGGGTACTCATCGAGACAGGGAGAAGGGTTGTAGGTGAAAGCA
AACAAGGATAAATAATACACGTGACAAAAAGTGCACCCATGAGTAGCTCTGTCCCTCTTCCCAACATCCACTTTCGT

                                                                          ↓
ATCTGGCTTTGTGCACCTGGAACTAAAGATCTGAAATGTTGTCTACACTGATCTTGCATTTGTTTGTAAAAATAACT
TAGACCGAAACACGTGGACCTTGATTTCTAGACTTTACAACAGATGTGACTAGAACGTAAACAAACATTTTTATTGA

AACTACATACTAATAATAACCTGCCCATTAATTTGGGTACATACTCATAATGAGTGCTACATGTTTTTGGTGTTCAT
TTGATGTATGATTATTATTGGACGGGTAATTAAACCCATGTATGAGTATTACTCACGATGTACAAAAACCACAAGTA

                    ↓
GCACACACCTCTAAATCATTGAGTCCGAGCTCTGCTCGCCTCATAAATGAAAGCAGCAAGTTAAAAAAAAAAAAAAA
CGTGTGTGGAGATTTAGTAACTCAGGCTCGAGACGAGCGGAGTATTTACTTTCGTCGTTCAATTTTTTTTTTTTTTT

                                           ↓
AAAAAAAAAAAAGCTTGGCTGTTTTAATGTACACAGTGTGCATCCTACAAGCAGTCTCTGGAAGGTCACTTCCCACT
TTTTTTTTTTTTCGAACCGACAAAATTACATGTGTCACACGTAGGATGTTCGTCAGAGACCTTCCAGTGAAGGGTGA

                                                                  ↓
GAGCCACTGTATTTCCAGGTGACAGTTTCTCTGAAACACTGCATTCAAAACATCCTGCATGCTGAAGTAGGCCAGTG
CTCGGTGACATAAAGGTCCACTGTCAAAGAGACTTTGTGACGTAAGTTTTGTAGGACGTACGACTTCATCCGGTCAC

TGGGACAGCCATTTTCCTCCAGTCTGTTGCTGTAACTTTTTGTCATTCATGCCATGGTGCTTGTCTGCATCAAGGAT
ACCCTGTCGGTAAAAGGAGGTCAGACAACGACATTGAAAAACAGTAAGTACGGTACCACGAACAGACGTAGTTCCTA
          8000
            ↓
AGCCCATTTCCCTGTGTGAGTGACAATGCTACAATACTTTGATGCATGTGAGCCTGAGAGAATTAACTTGGGGTAAC
TCGGGTAAAGGGACACACTCACTGTTACGATGTTATGAAACTACGTACACTCGGACTCTCTTAATTGAACCCCATTG

                                   ↓
CTCCATGGACTTATGCTCACATATTGCTCACACTTTGAAGAATCAACAAACCCAGTTTCATGCCTTTATCTCAGACC
GAGGTACCTGAATACGAGTGTATAACGAGTGTGAAACTTCTTAGTTGTTTGGGTCAAAGTACGGAAATAGAGTCTGG

                                                          ↓
AGGGAGAAGTGTGAATCTTCTCAGCCATGTGTCCAACCCGCTAGATTTATATTTCTGTTGCCTTCATTTTTTTAATA
TCCCTCTTCACACTTAGAAGAGTCGGTACACAGGTTGGGCGATCTAAATATAAAGACAACGGAAGTAAAAAAATTAT

TATATTTTAATAACCAATATATACAAATTAGGAAAGCTTTGTTACATACATTAACAATATTATACATTTTTTATGTT
ATATAAAATTATTGGTTATATATGTTTAATCCTTTCGAAACAATGTATGTAATTGTTATAATATGTAAAAAATACAA

    ↓
GTTCAAGACAACATTCCTCTTTACACAACCCGATTCTCAGACTGTAGCATCACTGTCATGATGAGCATCATCAGCAT
CAAGTTCTGTTGTAAGGAGAAATGTGTTGGGCTAAGAGTCTGACATCGTAGTGACAGTACTACTCGTAGTAGTCGTA

AGAGCATGAATT
TCTCGTACTTAA
```

give a different genomal blot pattern to that which has been previously reported for the H4 genes (18). The coding region of the RH4 gene differs from the LH4 gene by one conservative third base change in the 26th codon (LH4; AAT → RH4; AAC). Similarly, the coding region of the RH2A differs from the LH2A gene by one conservative third base change in the 106th codon (LH2A; GGT → RH2A; GGC).

The startpoint of transcription of the LH4 gene has been reported (17) while the other startpoints have been inferred by their proximity to the TATA box. As with other histone genes, the putative 5' untranslated regions of the five genes in pCH8.4E are characteristically short and pyrimidine rich.

Correct 3' processing of histone gene transcripts requires a dyad symmetry element found in the 3' untranslated region of most histone mRNAs as well as an adjacent conserved downstream element (19). All the genes shown here contain these elements. The second homology block contains a three base pair extension (GCT) which we have previously noted for other chicken histone genes (11). In the case of the H3 gene, the second element contains the sequence AATAAA. While this motif has been shown to be important in the generation of correct 3'-termini of mRNAs which are polyadenylated (20,21), we do not find polyadenylated forms of mRNA from this H3 gene in chicken.

Non-Coding Sequences

(a) The H3 Gene

The H3 gene leader sequence contains a canonical TATA box (5' TATAAAT 3', position 3,827, Fig. 2) and 5 potential CAAT boxes, 23, 51, 113, 124 and 137 base pairs upstream from the TATA box with the three most 5' of these elements almost fused head to tail (Figure 4). These five elements all share a 10 base pair consensus sequence (5' CAATGAGAN$^{A}_{T}$ 3'). Whether these multiple motifs modulate gene expression awaits further experimentation.

FIGURE 2: Total sequence of pCH8.4E. A vertical arrow marks each 100 base-pairs. Important elements are printed in bold type and boxed. Direct repeats delineating one boundary of the 2.1 kb inverted repeat element are marked with arrows (DR(L) and DR(R)) as are the inverted repeats at the other boundary (IR(L) and IR(R)). Two other inverted repeats just upstream (IR(1)), and downstream (IR(2)) from the right 2.1 kb repeat element are also marked with arrows. TATA boxes, as well as histone gene terminators (T) and associated downstream elements (DSE), are shown for each gene sequence. Upstream from the H4 genes three elements, which may be important for the transcriptional regulation of these genes, are marked (H4I, H4II and H4III). The H2A 5' sequences contain two potential CAT boxes (H2AC1 and H2AC2) and another element shown to be important for the expression of sea urchin H2A genes (H2A SS) while the H3 gene 5' region contains five potential CAT boxes (H3C1 - H3C5) and a sequence (ars-like) with a high degree of homology with a human sequence which acts as an ars element in yeast.

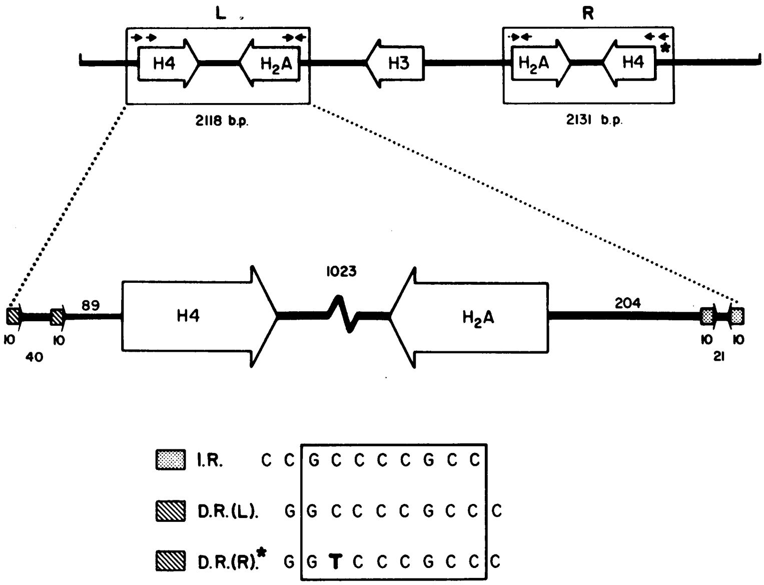

FIGURE 3: The inverted duplication. The area covered by the inverted duplication is boxed. The position and orientation of the direct and inverted repeats are shown relative to the histone genes and the boundaries of the duplication. An octamer sequence is shared between the direct repeats (DR) and inverted repeats (IR).

Within the H3 gene promoter, between the third and fourth CAAT boxes, there is an element 5' TTATTTTAAACTTCGAA 3' which has a high degree of homology with two human elements 5' $\text{TATT}^{\text{T}}_{\text{C}}\text{TAAATTTAGT}^{\text{T}}_{\text{A}}$ 3' which can act as autonomously replicating sequences (*ars*) in yeast (22). Osley and Hereford (23) have shown that a DNA sequence in the 3'-flanking region of a yeast H2B gene is necessary for S-phase transcriptional regulation of an adjacent H2A gene. This motif appears also to act as an *ars* sequence. Experiments to determine whether the element in the chicken H3 gene promoter region can act as a yeast *ars* sequence are in progress. This sequence may also be involved in S-phase transcriptional control of the H3 gene and other histone genes in the cluster.

(b) The H4 Genes

The two H4 gene leader sequences are identical for almost 150 base pairs upstream from the initiation codon except for one base change. Each TATA box (728 and 6,724, Fig. 2) is slightly unusual in sequence (ATAACA) although it still contains the highly conserved tri-nucleotide (ATA).

```
                          C1              C2              C3
AGTTTCTGTGTGCT |CAATGAGACA| CTG |CAATGAGACA| A |CAATGAGAGA|

                                                ars-like
CGGTGGTTGCCGGTACGTTCCAAAAGGCGG |TTATTTTAAACTTCGAA| T

                C4                           C5
CGAC |CAATGAGATT| CAGCCTGTTTGTTTCGGC |CAATCAGAAT| GAGCGG

                TATA
TGCTAG |TATAAATAT| CCCCACAGCGGAGAGCGGCGTCTATTTCCTGGC

                                           Initiation codon H3 gene
GTTCCTCCCTGCCTGCTTGCGTTCTTAAGCG    ATG
```

C1	CAATGAGACA	Human *ars*	TATT$^{T}_{C}$TAAATTTAGT$^{T}_{A}$
C2	CAATGAGACA	*ars*-like	TATTTTAAACTTCGAA
C3	CAATGAGAGA		
C4	CAATGAGATT		
C5	CAATCAGAAT		
Consensus			
	CAATGAGAN$^{A}_{T}$		

FIGURE 4: H3 gene leader sequence. The region immediately 5' to the H3 structural gene contains five potential CAAT boxes (C1-C5), with the consensus CAATGAGAN$^{A}_{T}$, and a TATA box. Also within this region is an element with strong homology to a human sequence which acts as an *ars* in yeast (*ars*-like).

Clerc *et al.* (10) have shown that regions just upstream from the TATA box of a *Xenopus* H4 gene are important in promoting transcription in an homologous oocyte transcription system. Subsequently, they showed these regions contained three conserved motifs present in all other H4 genes sequenced. All these elements are present in the H4 genes reported in this study (Figure 5; see also regions marked H4, I, II, III at positions near 700 and 6,750 in Fig. 2). The most proximal element 5'GTCC 3' is 15 base pairs upstream from the TATA box. Clerc *et al.* (op. cit.) have postulated that this element is equivalent to the 5' GATCC 3' motif found just upstream from the TATA box of most sea urchin histone genes and thus represents a general histone gene promoter element rather than a specific H4 gene element. However, we find no evidence for this element or any related element close to the TATA box of other chicken histone genes or indeed the histone genes of several other higher eukaryotes. We conclude that, at least for these cases, this element is in fact H4 gene-specific. The next element 5' TTCA 3' is 24

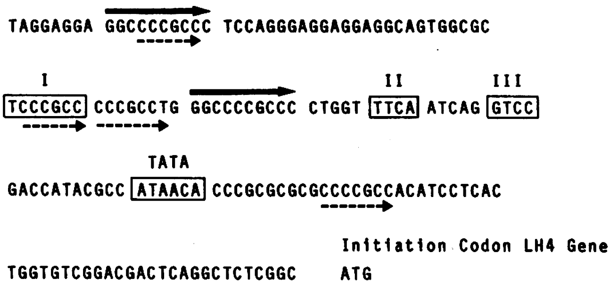
TAGGAGGA GGCCCCGCCC TCCAGGGAGGAGGAGGCAGTGGCGC
I
II
III
TCCCGCC CCCGCCTG GGCCCCGCCC CTGGT TTCA ATCAG GTCC
TATA
GACCATACGCC ATAACA CCCGCGCGCGCCCCGCCACATCCTCAC
Initiation Codon LH4 Gene
TGGTGTCGGACGACTCAGGCTCTCGGC ATG

FIGURE 5: LH4 gene leader sequence. The regions 5' to the LH4 and RH4 structural genes contain three putative regulatory elements (I, II, II) which have homology to sequences important for the transcription of a Xenopus H4 gene (10). The most distal of these elements (I) is partially repeated four times in the promoter region (indicated by broken arrows). A 10 base pair direct repeat is indicated by solid arrows. The 5' boundary of the direct repeat delineates the end of homology between the LH4 and RH4 leader sequence.

base pairs upstream from the TATA box and has the same sequence as the mouse H4 gene in this region. This is slightly different from the consensus sequence 5' GTCA 3'. The third and most distal element is 52 base pairs upstream from the TATA box and conforms to the consensus 5' CCGC 3'. We have noticed that this consensus can be extended for most higher eukaryotic H4 genes to 5' TCCCGC^{C}_{A} 3' and have boxed in this extended sequence in Figure 5. Interestingly this distal element lies between the direct repeats which delineate one end of the 2.1 kb inverted repeat element shown in Figure 3. In fact, this distal element is repeated imperfectly 4 times within the H4 gene promoter region (dotted arrows, Figure 5). As well as being located between the direct repeats, the element is also present once in each of the direct repeats and once just 3' to the TATA box. Thus there is an intimate relationship between part of a gene promoter and the ends of this inverted duplication within the chicken genome.

(c) The H2A Genes

The H2A gene leader sequences are almost identical for some 245 base-pairs upstream from the initiation codon. The promoter contains a canonical TATA box (5' ATATAAA 3') and two potential CAAT boxes (5' GCCTATCA 3'; 5' GCCATTGG 3') 37 and 72 base pairs upstream from it. (See regions near 2,650 and 4,800 in Fig. 2.)

Using a sea urchin H2A gene, Grosschedl et al. (9) have found that a region -165 to -111 upstream from the H2A cap site is important for maximal transcription in the Xenopus oocyte system. As pointed out by these authors,

```
                                          H2A SS
CAGCTCCCGCCCCGCCAT [TGCTGTGAGCAGAGAGAGGG] GCGGGGCGG

GGAAGGGCGCCCGTCGCTATTGGTCGGGCCGAGATTCGCTACGCCATTG

GGCGCTCGTTCCCCGTGAGGACAGCTCCGCCTATCAGGAGCCGGCGGGC

                        TATA
GCGCAACGGGGAAGGG [ATATAAA] GGCGCGCGGGCGAGCGCGGGCAGC

TGTTGCGTGTTTTGATTTCTGTTGAGTTGGCAGTTGAGCTTCTGAGCGA

                Initiation Codon RH2A Gene
CTGATCGGCG      ATG

                                     G     6 bp
Sea Urchin Consensus         GCTGCTCTA  ——  CACCAACAGATGG
                               ||||:||         |  | |||  ||
                H2A SS         TGCTGTA  ——    GCAGAGAGAGGG
                                    G
```

FIGURE 6: RH2A gene leader sequence. The regions 5' to the RH2A and LH2A structural genes contain an element H2ASS with homology to motifs important for the transcription of sea urchin H2A genes (9). This element is located between an inverted repeat shown by arrows in the figure. The 5' boundary of the inverted repeat delineates the end of the homology between the RH2A and LH2A gene.

this area contains two conserved regions. One of these has homology with the Moloney murine sarcoma virus enhancer as well as the 5' LTR sequences of the Simian sarcoma virus and the murine Friend spleen focus forming virus.

The two conserved regions in the sea urchin genes are also present in the chicken H2A genes reported in this study although, in the chicken genes, these regions (denoted H2ASS in Figs. 2 and 6) have been fused into one element (Figure 6; see also H2ASS elements near 2,750 and 4,700 in Fig. 2). In the chicken sequences, however, the homology with viral enhancers is less obvious. Of particular note, is the fact that the conserved H2A promoter element is found between the inverted repeat elements which mark one boundary of the 2.1 kb inverted duplication (Figure 3, Figure 6).

We have already noted (Fig. 5) that part of the H4 gene promoter region is found between the direct repeats delineating one end of the 2.1 kb inverted duplication. Thus, both boundaries of the duplicated inversion not only contain related repeats, but within these repeats are conserved elements with strong homology to known modulators of transcription.

Transcription and DNA Rearrangements in Histone Genes

The two features which are common to the inverted duplication seen in pCH8.4E are, firstly, the presence of direct or inverted repeats at the boundaries of the rearrangements and secondly, the intimate association between the ends of the rearrangements and gene promoter elements. The presence of repeated motifs within promoter regions may be important in the transcriptional activation of genes as these elements have been postulated as binding sites for trans-acting regulatory factors. In some cases there is direct evidence for such interactions (21). It is possible that interaction between regulatory factors either directly, or mediated through other molecules such as RNA polymerase, may bring promoter sequences into juxtaposition so that recombination can occur. Ohtsubo and Ohtsubo (25) have postulated that RNA polymerase may play an important role in site-specific recombination while Vitelli and Weinberg (15) have speculated that the basis of many eukaryotic rearrangements may be the fortuitous apposition of small regions of homology which have particular secondary structure due to interaction with protein. It is not possible to tell in the chicken cluster which of the inverted duplication elements is the original and which is the derived, but a prediction is that the derived element is likely to reside in a gene promoter region. If this occurred, the gene would be lost during the reciprocal rearrangement. We note that immediately upstream from the RH2A (IR(1) at 4,652 and 4,680, Fig. 2) gene and just downstream from the RH4 (IR(2) at 6,874 and 6,917, Fig. 2) gene outside the 2.1 kb inverted repeat there are inverted repeats and these may be the remnants of gene promoters. Similar elements are not found close the boundaries of the left inverted duplication.

The complete sequence of a region of chicken histone genes containing an inverted duplication has allowed us to mark the boundaries of the presumed recombination event and to note the features at these boundaries. Considered on its own, the fact that an H3 gene is found at the centre of symmetry in this cluster does not seem significant. However, we find two other examples of symmetrically ordered genes, neither of them related to each other or to pCH8.4E, but both containing central H3 genes (D'Andrea et al., submitted). The significance of these arrangements is not known, but they may confer a selective advantage for co-ordinated expression of blocks of histone genes during S-phase.

REFERENCES

1. Maxson, R., Cohn, R. and Kedes, L. (1983). Ann. Rev. Genet. 17, 239-277.
2. Harvey, R.P., Krieg, P.A., Robins, A.J., Coles, L.S. and Wells, J.R.E. (1981). Nature 294, 49-53.

3. Engel, J.D. and Dodgson, J.B. (1981). Proc. Natl. Acad. Sci. USA 78, 2856-2860.
4. 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.
5. Heintz, N., Zernik, M. and Roeder, R.G. (1981). Cell 24, 661-668.
6. Zernik, M., Heintz, N. and Roeder, R.G. (1980). Cell 22, 807-815.
7. Van Dongen, W., De Laaf, L., Zaal, R., Moorman, A. and Destree, O. (1981). Nuc. Acids Res. 9, 2297-2311.
8. Davidson, E.H., Jacobs, H.T. and Britten, R.H. (1983). Nature 301, 468-470.
9. Grosschedl, R., Machler, M., Rohrer, U. and Birnstiel, M.L. (1983). Nuc. Acids Res. 11, 8123-8136.
10. Clerc, R.G., Bucher, P., Strub, K. and Birnstiel, M.L. (1983). Nuc. Acids Res. 11 8641-8657.
11. Harvey, R.P., Robins, A.J. and Wells, J.R.E. (1982). Nuc. Acids Res. 10, 7851-7863.
12. Falkner, F.G. and Zachau, H.G. (1984). Nature 310, 71-74.
13. Sanger, F., Nicklen, S. and Coulson, A.R. (1977). Proc. Natl. Acad. Sci. USA 74, 5463-5467.
14. Staden, R. (1980). Nuc. Acids Res. 8, 3673-3694.
15. Vitelli, L. and Weinberg, E.S. (1983). Nuc. Acids Res. 11, 2135-2153.
16. Isenberg, I. (1979). Ann. Rev. Biochem. 48, 159-191.
17. Sugarman, B.J., Dogson, J.B. and Engel, J.D. (1983). J. Biol. Chem. 258, 9005-9016.
18. Ruiz-Carrillo, A., Affolter, M. and Renaud, J. (1983). J. Mol. Biol. 170, 843-859.
19. Birchmeier, C., Shumperli, D., Sconzo, G. and Birnstiel, M.L. (1984). Proc. Natl. Acad. Sci. USA 81, 1057-1061.
20. Fitzgerald, M. and Shenk, T. (1981). Cell 24, 251-260.
21. Montell, C., Fisher, E.F., Caruthers, M.H. and Berk, A.J. (1983). Nature 305, 600-605.
22. Montiel, J.F., Norbury, C.J., Tuite, M.F., Dobson, M.J., Mills, J.S., Kingsman, A.J. and Kingsman, S.M. (1984). Nuc. Acids Res. 12, 1049-1068.
23. Osley, M.A. and Hereford, L. (1982). Proc. Natl. Acad. Sci. USA 79, 7689-7693.
24. Parker, C.S. and Topol, J. (1984). Cell 37, 273-283.
25. Ohtsubo, H. and Ohtsubo, E. (1978). Proc. Natl. Acad. Sci. USA 75, 615-619.