Volume 15 Number 7 1987 Nucleic Acids Research

# Unusual structure, evolutionary conservation of non-coding sequences and numerous pseudogenes characterize the human H3.3 histone multigene family

Dan Wells[1], Doug Hoffman and Larry Kedes*

The MEDIGEN Project, Department of Medicine, Stanford University School of Medicine and
Veterans Administration Medical Center, Palo Alto, CA 94304, USA

Received January 13, 1987; Revised and Accepted February 28, 1987 Accession no. Y00108

## ABSTRACT

The genomic organization of the replication-independent, basally expressed, human H3.3 gene is atypical of traditional histone gene organization. The gene contains 3 introns totalling 7.8 kb and unusual direct repeats flank all three intron-exon splice junctions. The transcription initiation site was mapped by S1 nuclease protection analysis and confirms that cDNA clones previously reported (1) were full length. Sequence similarities between regions at the 5′ and 3′ termini of this human gene and a chicken H3.3 gene (2) lead us to propose that either the previous assignments of termini of the chicken gene are in error, or there are alternative transcription start and polyadenylation sites. The 85% base matching of the human and chicken H3.3 3′ UTR sequences for 520 bases is unprecedented among homolog 3′UTR segments, especially considering that these species are separated by over 250 Myr of evolution. We also present the sequence of three related processed human H3.3 pseudogenes and provide evidence demonstrating that most of the 20 to 30 copies of the H3.3 gene within the human genome are in fact processed pseudogenes.

## INTRODUCTION

Traditionally, histone genes share a number of common structural characteristics (3). They are intronless genes that are transcribed into non-polyadenylated mRNAs; they have "typical" RNA polymerase II promoters, bearing the consensus CCAAT and TATA boxes; and they have relatively short 5′ and 3′ untranslated regions. Furthermore the 3'UTRs contain a hyphenated dyad symmetry element known to be important in the 3′ end processing of histone mRNAs (4). In vertebrates, regulation of histone genes is typically cell cycle-dependent, with expression being linked to ongoing DNA replication.

Recent work has shown that histone genes are considerably more versatile in both their structure and regulation than this traditional description would suggest. In contrast to the classical histone genes initially studied, histone proteins and genes have now been characterized which are either partially replication-dependent or completely replication-independent, or even tissue-specific (reviewed in 5, 6). The H3.3 histone, for example, is expressed in a replication-independent manner (6, 7, 8). Our recent work has suggested that this unexpected regulation of H3.3 expression is also accompanied by an unexpected mRNA structure that distinguishes it from the replication-dependent histone variants (1). It contains lengthy 5′ and 3′ leader and trailer sequences, lacks a 3′ hyphenated dyad symmetry element, is polyadenylated and has a unique codon usage pattern (9). Additionally, we described the cDNA of a partially processed H3.3 mRNA which strongly suggested that the gene contains at least one intron (1). In order to confirm that these unexpected features are a reflection of gene structure and to prepare for a study of regulation of the H3.3 gene, we have isolated and sequenced genomic DNA clones containing the human H3.3 gene.

In this paper we describe the atypical genomic organization of the replication-independent, basally expressed human H3.3 gene and define its 5′ and 3′ termini and its intron-exon boundaries. The gene is shown to contain 3 introns totalling 7.8 kb. Unusual direct repeats are shown to flank all three intron-exon splice junctions. In addition we note an interesting region of sequence conservation in the 5′ flanking DNA, demonstrate the extreme conservation of the H3.3 3′ UTR over 250 Myr of evolution and speculate on the significance of these conserved regions. These comparative data suggest that the original assignments of 5′ and 3′ termini to a chick H3.3 gene (2) may have been incorrect. Finally we present the sequence of three related processed pseudogenes and provide evidence demonstrating that most of the 20 to 30 copies of the H3.3 gene within the human genome are in fact processed pseudogenes.

## MATERIALS AND METHODS

### General Methods.

Plasmid DNA preparation, restriction enzyme digestions, agarose gel electrophoresis, DNA blotting to nitrocellulose, nick-translation and sequencing using M13 vectors were performed by standard techniques and as described previously (1). Over 95% of the H3.3 gene sequence presented was sequenced multiple times and on both strands using overlapping fragments.

### Screening and Isolation of Recombinant Phage.

Two human genomic libraries (10) were screened with nick-translated H3.3 cDNA probes using standard *in situ* plaque hybridization techniques (11). One library was prepared using a partial *Eco* RI digestion and the other using a partial *Hae* III-*Alu* I digestion with ligated *Eco* RI linkers. Both were inserted into *Eco* RI-digested Charon 4A arms (10). Plaques which screened positive were plaque purified and isolated phage were grown as plate stocks and purified by CsCl gradient centrifugation (12).

### DNA fragment probes.

Five different fragments were isolated from the H3.3 cDNAs (1), nick-translated, and used as probes. These include: 1) a 500 bp *Nco* I coding region fragment (NC); 2) a 375 bp 5′ intron sequence from cDNA pHH3C-3 (5'C); 3) a 5'UTR 75 bp fragment upstream of the *Sac* I site in pHH3B-2 (5'B); 4) a 5′ coding region fragment from a *Sac* I site in the 5'UTR to a *Bgl* II site in the coding region and 5) a 350 bp 3'UTR sequence from the *Kpn* I site to a *Rsa* I site at the 3′ end of the cDNA (3'KR) (Figure 8).

For use in nuclease S1 mapping of the 5′ end of the H3.3 mRNA, the DNA of the subclone described in Figure 5 was digested with *Nco* I, dephosphorylated with calf intestinal phosphatase (Boehringer Mannheim) and treated with T4 polynucleotide kinase (New England Biolabs) using standard techniques (12).

### Nuclease S1 mapping.

Nuclease S1 mapping of the 5′ end of the H3.3 mRNA was done essentially as detailed by Maniatis et al. (12), with the following modifications. The kinased *Nco* I probe was hybridized to 50 μg of total cytoplasmic RNA at 55°C for 16 hours. Each sample was digested with 300 units of nuclease S1 (Sigma Scientific) at 37°C for 30 min. All buffers used were identical to those described by Maniatis et al. (12). Following digestion, 0.2 volumes of 4.0M ammonium acetate, 20 μg of yeast tRNA and an equal volume of isopropyl alcohol were added to each sample. The precipitates were resuspended in 1 μl of formamide dye mix (37% formamide with .08% xylene cyanol, .08% bromophenol blue, and 20 mM disodium EDTA), electrophoresed on a 6% polyacrylamide urea sequencing gel and autoradiographed.

### Computer analysis.

Sequencing data, including data entry and alignment, were processed using the IntelliGenetics GEL program. Comparisons and alignments between chicken and human genes and between intron-exon borders were performed by the IntelliGenetics IFIND program. Alignment and analysis of pseudogene sequences was performed using MULTAN (13). MULTAN was run on the SUMEX-AIM DEC 2060 computer.

## RESULTS

### Isolation of the H3.3 genomic sequences.

There is a complex human H3.3 multigene family containing 20-30 members with varying degrees of sequence similarity (1). However, since all 20 of the cDNAs isolated appeared to be encoded by the same gene, we proposed that there may be only one or at most a few functional genomic loci. We had previously isolated a cDNA representing an apparently unprocessed H3.3 mRNA precursor containing a putative intron in its 5′ UTR. Since both the H3.3 coding and 3′ UTR regions contain sequences which are present in 20-30 copies in the human genome, we attempted to use the putative 5′ intron as a hybridization probe to distinguish the functional H3.3 gene. A 375 bp fragment (5'C) from the putative 5′ intron of pHH3C-3 (Figure 1C) was isolated and labeled by nick-translation. This probe was then hybridized to blots of size-fractionated genomic DNA that had been digested with either *Eco*RI or *Hin*dIII restriction endonucleases. As can be seen in Figure 1A, this intron probe hybridizes to a single 6.5 kb *Eco*RI fragment and a single 4.0 kb *Hin*dIII fragment indicating that it is probably present in a single copy in the human genome. This probe was then used to examine 21 clones of lambda phage carrying human DNA that had been selected previously on the basis of their ability to hybridize to an H3.3 coding region probe. Of these 21 genomic clones, only HuH3-6 and HuH3-149 hybridized to the 5′ intron probe.

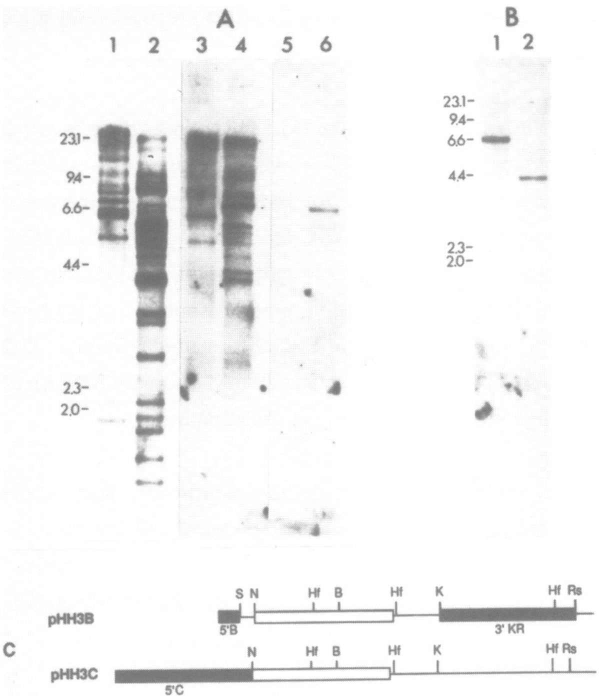

Figure 1. Hybridization of cDNA fragments to human genomic DNA. A) About 8 μg of human genomic DNA isolated from Hela cells was digested with either BamHI (lanes 1, 3, 5) or EcoRI (lanes 2, 4, 6), electrophoresed onto 0.8% agarose gels and blotted onto nitrocellulose filters. Genomic blots were then hybridized to radiolabeled probes illustrated in (C). Lanes 1 and 2 were hybridized to the 3'KR probe. Lanes 3 and 4 were hybridized to the 5'B probe. Lanes 5 and 6 were hybridized to the 5'C probe. B) Same as in (A) except genomic DNA was digested with either EcoRI (lane 1) or HindIII (lane 2) and hybridized to the 5'C probe. C) Restriction maps of the cDNAs pHH3C and pHH3C (Wells and Kedes, 1985). The coding region of the cDNAs are illustrated by open boxes. The regions used as probes are illustrated by solid boxes and are labeled separately. For complete description of probes see the Methods. Restriction enzyme abbreviations are as follows: Hf, HinfI; N, NcoI; Rs, RsaI; B, BglII; K, KpnI; S, SacI.

### Characterization of the genomic clones.

Clone HuH3-6 was originally isolated from a partial *Eco* RI digestion of human DNA. The other, HuH3-149, was isolated from a partial *Alu* I-*Hae* III digest of human genomic DNA. We determined a number of restriction endonuclease cleavage sites of the two clones (Figure 2) and identified the corresponding cDNA regions by hybridization to cDNA restriction fragments. HuH3-6 contains a 17 kb human DNA insert which hybridized to a H3.3 coding region probe (NC) and a 3′ UTR probe (3'KR) but failed to hybridized to the 5′ UTR probe (5'B). HuH3-149 contains an 11 kb insert which hybridized to all regions of the cDNA including the 5′ UTR probe (5'B) (Figure 2). The location of the cDNA hybridizing regions in the clones is shown in Figure 2.

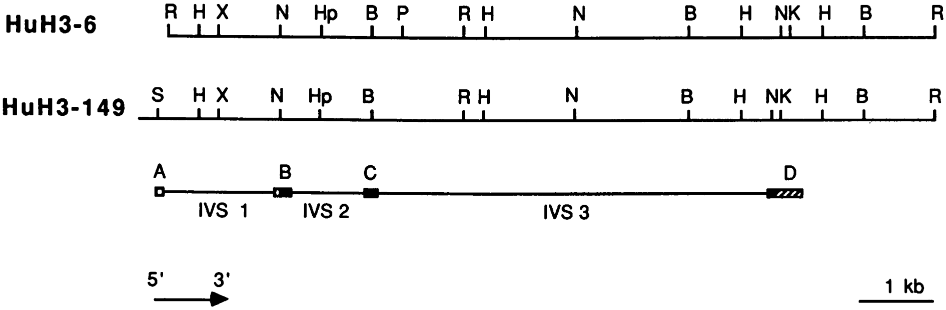

Figure 2. Restriction maps of the human H3.3 gene. Two largely overlapping lambda genomic clones (HuH3-6 and HuH3-149) containing the human H3.3 gene were mapped by restriction enzyme digestion. The exon regions were localized by hybridization to cDNA fragments and specific exon borders were determined by sequence analysis. The exons are labelled A, B, C, and D. The introns are numbered 1, 2, and 3. Open boxes correspond to 5' untranslated regions; black boxes correspond to coding regions; hatched boxes correspond to 3' untranslated regions. Restriction enzyme abbreviations are as follows: R, EcoRI; H, HindIII; X, XhoI; N, NcoI; Hp, HpaI; B, BglII; P, PstI; K, KpnI; S, SacI. Not shown: 6kb of HuH3-6 downstream of the EcoRI site.

Four regions of hybridization were detected in the genomic clones. The sequencing of these regions was performed using isolated restrictions fragments inserted either directionally or randomly into M13 vectors. The genomic sequence of the human H3.3 gene presented in Figure 3 includes the 1040 bp identical to the cDNA sequenced previously. In addition it contains approximately 100 bp of 5′ flanking sequences, 150 bp of 3′ flanking sequences and 1500 bp of intron sequences extending from each intron-exon border. As previously predicted, the H3.3 gene contains an intron in its 5′ UTR. In addition, there are two other introns within the coding region of the gene. These two introns appear at the precise location of the two coding region introns previously observed for the chicken H3.3 genes (2). With respect to the primary transcript, intron 1 begins at +88 and extends ~1.5 kb. Intron 2 begins at +1776 (within amino acid 42) and extends ~1.0 kb, while intron 3 begins at +2933 (between amino acids 93 and 94) and extends ~5.2 kb. The entire primary transcription unit is a remarkable 8.8 kb, more than 16-fold longer than the 0.5 kb primary transcription unit of a typical cell cycle-regulated H3.1 gene.

## Analysis of the intron borders.

All intron-exon borders agree with the splice junction consensus of Benoist et al. (14). Curiously, all 3 sets have 7-8 bp direct repeats at the splice junctions (Figure 4). The presence of these direct repeats could be fortuitous, especially since splice consensus sequences are already present. However, several points argue against such an explanation. First, the direct repeats are specific to each intron. Secondly, based on consensus data, the 5′ ends of the splice junctions should show more homology to each other than to the 3′ end of the splice junctions, and they do not. Finally, the repeat region is on the 3′ side of the intron-exon border in intron 1 but on the 5′ sides of the intron-exon borders in introns 2 and 3, suggesting that the constraints imposed by

```
          -80                     .                 -40                     .
GCGAGCCTTCCCTCCATTGTGTGTGATTGGCTGCGCGCGGCGGGGGCGGGGCGGCGTGTGTTGGGGGATAGCCTCG
                +1◊ ◊◊                 .                   .
GTGTCAGCCATCTTTCAATTGTGTTCGCAGCCGCCGCCGCGCCGCCGTCGCTCTCCAACGCCAGCGCCGCCTCTCG
.                   .                  100                   .
CTCGCCGAGCTCCAGCCGAAGAGAAGGGG^GTAAGTTTCCCCGTCTGCCCGCTTCCCCGGAACCGAGCCCCGTTGC
     .                   .                              .
CGTCAGCCCCGAGCCGGGGCCTGGCGGTGCGGGAGGGGA  -960 bp- ATATTTTCAGATATTTACAGCAATCC
.                1200                   .                   .
ATCAACAGTCAGACGAAGGGGGGGCAGCCAAAGTGGGGGGAGTAAGACCTTTTTTTTGCTGTAATTTGACTCGAC
    .                   .                1300                    .
CTTCCARTGGTTTGTGCCTTTTTTCTTTTGACTTGTTTGTGGATGGAATGTTTACAGACATTTCTAATTACTGCTT
        .                   .                   .                1400
TAATTAAATAAATTGGATCAAAGGCCGTTCGAGGTATTTTTGTTTTGCCGTTTGTCGCTCAGAATTGGCATTTTGA

            .                   .                   .                   .
GAGGTGATTGATACTGCTAACAATTTTCTAGTACTCTAGTTTGTTTAAGAAGAGATTTTGGGTAGACGTAATCTTC
             1500                   .                   .
ACCTTTCAAATTATATAACAATACGAACATTATTTTTTATACTGATCATAATTTCCAGATTTGGGGAGGGGGTGAT
.                   .                1600                   .
CGTGGCAGGAAAAGTTGTATGTTTGGTAGTTGCATATGGTGATTTTTGATTTTTCAATGCTGGTAG^GTAAGTAAG
     .                       .                          .
GAGGTCTCTGTACC ATG GCT CGT ACA AAG CAG ACT GCC CGC AAA TCG ACC GGT GGT AAA
               MET Ala Arg Thr Lys Gln Thr Ala Arg Lys Ser Thr Gly Gly Lys
     1700                         .                          .
GCA CCC AGG AAG CAA CTG GCT ACA AAA GCC GCT CGC AAG AGT GCG CCC TCT ACT GGA
Ala Pro Arg Lys Gln Leu Ala Thr Lys Ala Ala Arg Lys Ser Ala Pro Ser Thr Gly
            .                            .                1800
GGG GTG AAG AAA CCT CAT CGT TAC AG ^GTATTAAAAAACAGGAAAAAAATGGGACAAGTCTCTCTTGT
Gly Val Lys Lys Pro His Arg Tyr Arg

  .                   .                   .                   .
ATGTATCCATATAATTTAACAAAAAGATGGATAACAGGAAAACTTTTTGCTTTAGAGAAACTTTTTTTTTTCATTTG
  1900                    .                   .                   .
AACACTTAACTACTGCTTAAATAAATGTACTGTATGATCATTTATATATAAAGTTAAGTATTAGGTTTTATTXAAAC
        .                2000                   .                   .
GTTTAACTTTXAAGCCATAATCTTACCTGGAGGTCTAAGGAGACCTCGTATATCACTGATAATGTTAATGGGATATA
           .                            .                2700
TTGACATTTTAGTTAACC -600bp- AAGAATGTTGGGTGGCTTATTTTTGAAAGATTACTGCATTTCTXTGAAGC

   .                   .                   .                    .
TGCCCACTTACCTTTTGTGCTATTTATGTTTTTGGTAACAGTTTCTTTATTAATTTTTTACAG^G CCT GGT ACT
                                                                  Pro Gly Thr
          2800                            .                          .
GTG GCG CTC CGT GAA ATT AGA CGT TAT CAG AAG TCC ACT GAA CTT CTM ATT CGC AAA
Val Ala Leu Arg Glu Ile Arg Arg Tyr Gln Lys Ser Thr Glu Leu Leu Ile Arg Lys
                 .                          .                   2900
CTT CCC TTC CAG CGT CTG GTG CGA GAA ATT GCT CAG GAC TTT AAA ACA GAT CTG CGC
Leu Pro Phe Gln Arg Leu Val Arg Glu Ile Ala Gln Asp Phe Lys Thr Asp Leu Arg
                     .                         .                   .
TTC CAG AGC GCA GCT ATC GGT GCT TTG CAG ^GTAAAATGGTGGGTGGGAAGACTCAGAGTTTGTATT
Phe Gln Ser Ala Ala Ile Gly Ala Leu Gln
        .                3000                    .                   .
CCTGTTGTGTACCAAGAACAGTTCCAAATTGTTGCATGTGCTTATATCATTTAATCACAAGCCTGTCAGGTAMTTGA
             .                   .                3100                    .
TATTGTTACTTCACTGTKAGACTTCAGAAAGGTTAATTGCTCAAGGTCATACACGTAGAAATGGCAAACCATAATTG
                .                             .                   .
ACTATTTGACTCAGGCATTGCTACATTCAG -4.8 kb- CTTTGTTTTTTTCATAAGCTGCTTTTGAGCTTTGTCC
```

```
      8000                      .                   .                   .
CACAGGTTGTAAAATGTAAGCATTTGGTAAAATTGTCAGCATCTTGCCCAGTCATTTTTTAAAGGGTTXAAAAACCT
            .                 8100                  .                   .
TTTTGTTTTAATTCGTATAGTTGGGTCTTAACTATTGGAAATAACATCATCAGTAATTTTTTCCTTCATTCCTTTTG

                  .                          .                          .
CAG^GAG GCA AGT GAG GCC TAT CTG GTT GGC CTT TTT GAA GAC ACC AAC CTG TGT GCT
    Glu Ala Ser Glu Ala Tyr Leu Val Gly Leu Phe Glu Asp Thr Asn Leu Cys Ala

                      .                          .
ATC CAT GCC AAA CGT GTA ACA ATT ATG CCA AAA GAC ATC CAG CTA GCA CGC CGC ATA
Ile His Ala Lys Arg Val Thr Ile MET Pro Lys Asp Ile Gln Leu Ala Arg Arg Ile
.                         .                 8300                   .
CGT GGA GAA CGT GCT TAA GAATCCACTATGATGGGAAACATTTCATTCTCAAAAAAAAAAAAAAAATTTCT
Arg Gly Glu Arg Ala  .
         .                   .                   .                8400
CTTCTTCCTGTTATTGGTAGTTCTGAACGTTAGATATTTTTTTTCCATGGGGTCAAAGGTACCTAAGTATATGATTG

            .                   .                   .                   .
CGAGTGGAAAAATAGGGGACAGAAATCAGGTATTGGCAGTTTTTCCATTTTCATTTGTGTGTGAATTTTTAATATAA
           8500                    .                   .
ATGCGGAGACGTAAAGCATTAATGCAAGTTAAAATGTTTCAGTGAACAAGTTTCAGCGGTTCAACTTTATAATAATT
                  .                8600                  .
ATAAATAAACCTGTTAAATTTTTCTGGACAATGCCAGCATTTGGATTTTTTTAAAACAAGTAAATTTCTTATTGATG
  .                   .                   .                 8700
GCAACTAAATGGTGTTTGTAGCATTTTTATCATACAGTAGATTCCATCCATTCACTATACTTTTCTAACTGAGTTGT

     .                   .                   .                   .
CCTACATGCAAGTACATGTTTTTAATGTTGTCTGTCTTCTGTGCTGTTCCTGTAAGTTTGCTATTAAAATACATTAA
     8800                  .                   .                   .
ACTATACTGCTTTTGGTCTTTATTATAGCCTTGCCCTACAATTATATTCCAAATAATTTCAGTATTAAGTCCTTTAA
     *      .               8900                  .                   .
TAGTCACTTTATATCTGAGACGTGAGACTTATTAGAGAGAAGGAAAGTATAGACTTGGTGGCAAAAGGAACTTGGC

              .
TTTATTTCTAATTTTTAGAAGCTT
```

Figure 3. The nucleotide sequence of the human H3.3 gene. Sequence analysis was performed as described in the Methods. Sequence numbering, above the sequences, begins at the trancription initiation site and continues through introns and exons. Exons are separated from introns by (^). The 5' end of the mature mRNA, as determined by S1 nuclease protection, is indicated by +1. The 5' ends of the cDNAs are indicated by diamonds. The 3' end of the mature mRNA is noted by a (*) at nucleotide 8797. Specific sequences of interest are underlined and are discussed separately throughout the text.

the splice consensus sequence plays an insignificant role in the conservation of these direct repeats. The possible significance of these direct repeat sequences will be discussed later.

### Analysis of 5' UTR and upstream promoter sequences.

The start of transcription of the H3.3 mRNA (the cap site) was tentatively identified by the comparison of two full length cDNAs and three processed pseudogenes. All five of these sequences have their 5′ terminus within three nucleotides of each other. We suspected that the transcription start site must be very close to this location. The actual cap site was formally identified by S1 analysis (Figure 6). Since there was an intron in the 5′ UTR and since exon A did not contain any restriction sites that allowed simple 5′ end labeling, we engineered a new plasmid construct to provide restriction sites at more convenient locations (Figure 5). We excised a DNA segment, starting in exon A and ending in exon C, and replaced it with the

```
  EXON A ...AGAAGGGG  GTAAGTTT... INTRON 1
                   *  ******
INTRON 1 ...GCTGGTAG  GTAAGTAA... EXON B

  EXON B ...CGTTACAG  GTATTAAA... INTRON 2
              ******  *
INTRON 2 ...TTTTACAG  GCCTGGTA... EXON C

  EXON C ...CTTTGCAG  GTAAAATG... INTRON 3
             *******  *
INTRON 3 ...TTTTGCAG  GAGGCAAG... EXON D
```

Figure 4. Comparison of Exon-Intron splice borders. Sequences colinear with the mRNA are underlined. The direct repeats at the splice borders are aligned and repeated sequences are noted by asterisks.

corresponding region of the cDNA. This reconstructed the gene, except for deletion of introns 1 and 2, and allowed us to use restriction sites in exon B, 3′ of intron 1, to create a short DNA fragment useful for end-labeling. In the experiment shown in Figure 6, the gene construct illustrated in Figure 5 was digested with *Nco* I and end-labeled with polynucleotide kinase. Fifty µg of total cytoplasmic RNA from either chicken liver, mouse Ltk- cells, or human MRC-5 cells was hybridized to the labeled probe and treated with nuclease S1 as described in Methods. The protected fragment was electrophoresed on a sequencing gel (lanes 1 through 4), along with a known sequence for assistance in sizing. The protected fragment was 114 bases long, placing the 5′ end of the mRNA at the adenosine labeled +1 in Figure 3. This corresponds well to the location of the 5′ ends of the cDNAs and pseudogenes. The location of the transcription initiation site by primer extension analysis (data not shown) gave the same result. It is interesting to note that total cytoplasmic RNA from mouse L cells also generates a 114 nucleotide protected fragment (lane 3) suggesting an unusual extensive conservation of the 5' UTR between mouse

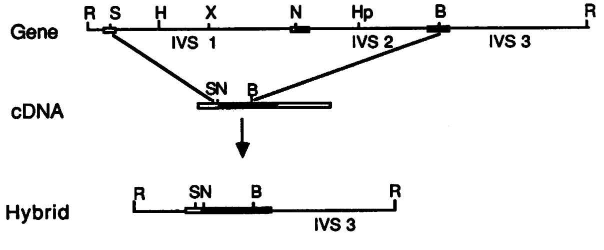

Figure 5. Construction of the hybrid gene used as a probe for S1 nuclease protection assays. A SacI / BglII fragment from the gene was removed and replaced with the corresponding fragment from the cDNA. The hybrid lacks introns 1 and 2, and the NcoI site of exon 2 is now near the 5' end of the gene. Restriction enzyme abbreviations are as follows: R, EcoRI; H, HindIII; X, XhoI; N, NcoI; Hp, HpaI; B, BglII; P, PstI; K, KpnI; S, SacI.

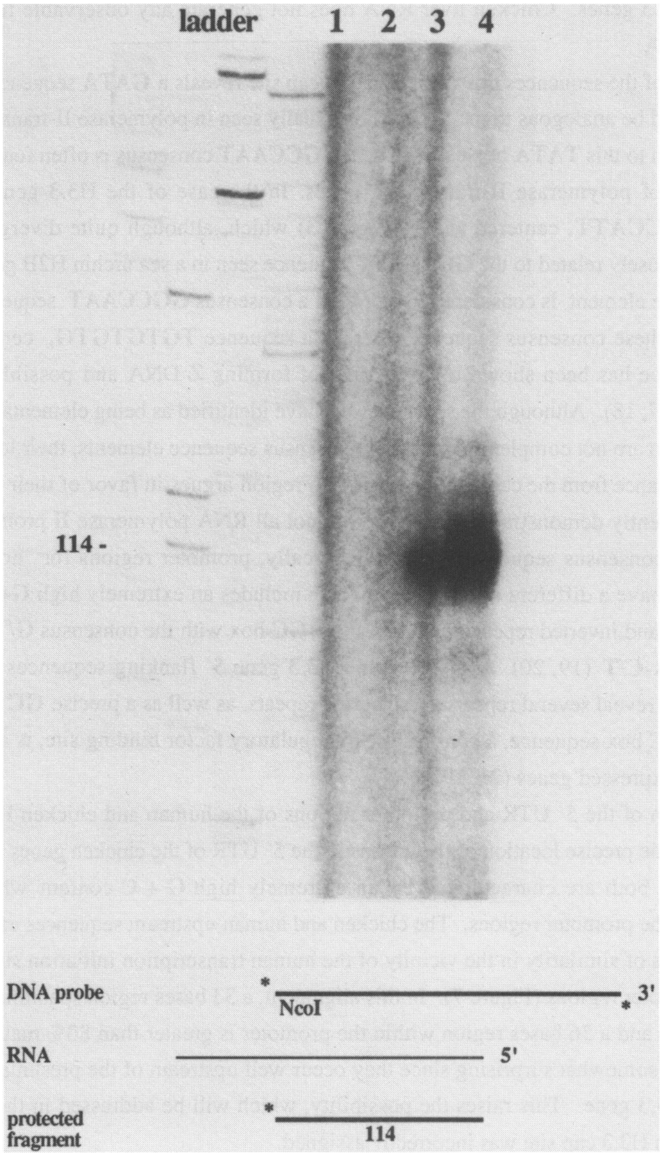

Figure 6. S1 mapping of the 5' cap site. *Above left:* The M13 sequencing ladder of a known sequence was used as a sizing standard. *Above right:* 50 ug of yeast tRNA (lane 1) or 50 ug of total cytoplasmic RNA from chicken liver (lane 2), mouse Ltk- cells (lane 3) and human MRC-5 cells (lane 4) was hybridized overnight to the NcoI-digested, end-labelled probe illustrated in Figure 5 and then digested with S1 nuclease as described in Methods and Materials. The signal at 114 nt in lane 3 has been demonstrated in independent experiments (not shown), and is not due to spillover from lane 4. *Below:* full length RNAs will protect an end-labelled DNA probe to a length of 114 nucleotides.

and human H3.3 genes. Chicken liver RNA does not generate any observable fragment, nor does yeast tRNA.

Analysis of the sequences upstream from the cap site reveals a GATA sequence centered at -25 which could be analogous to the TATA box usually seen in polymerase II-transcribed genes (15). In addition to this TATA box equivalent, a GGCCAAT consensus is often found in the -70 to -80 region of polymerase II-transcribed genes. In the case of the H3.3 gene, there is a sequence, CTCCATT, centered at -77 (Figure 3) which, although quite diverged from the consensus, is closely related to the GCTCATT sequence seen in a sea urchin H2B gene. The sea urchin sequence element is considered to represent a consensus GGCCAAT sequence (14, 16). In addition to these consensus sequences there is a sequence TGTGTGTG, centered at -73. Such a sequence has been shown to be capable of forming Z DNA and possibly enhancing transcription (17, 18). Although the sequences we have identified as being elements of a putative promoter region are not completely typical of consensus sequence elements, their location at the appropriate distance from the cap site in a G+C rich region argues in favor of their significance. It has been recently demonstrated, however, that not all RNA polymerase II promoter regions contain these consensus sequences (19). Specifically, promoter regions for "housekeeping" proteins often have a different arrangement, which includes an extremely high G+C content, a series of direct and inverted repeat structures, and a GC box with the consensus G/T-G-G-G-C-G-G-G/A-G/A-C/T (19, 20). Analysis of the H3.3 gene 5′ flanking sequences for repeated structures does reveal several repeats and inverted repeats, as well as a precise GC box centered at -48. This GC box sequence, a consensus Spl regulatory factor binding site, is also found in other basally expressed genes (21, 19).

Comparison of the 5′ UTR and promoter regions of the human and chicken H3.3 genes is difficult since the precise location of the introns in the 5′ UTR of the chicken genes are uncertain (2); however, both are characterized by an extremely high G + C content which extends upstream into the promoter regions. The chicken and human upstream sequences can be aligned to show regions of similarity in the vicinity of the human transcription initiation site, as well as upstream promoter regions (Figure 7). In this alignment, a 33 bases region spanning the cap site is 94% similar, and a 56 bases region within the promoter is greater than 80% matched. These alignments are somewhat surprising since they occur well upstream of the presumed cap site of the chicken H3.3 gene. This raises the possibility, which will be addressed in the discussion, that the chicken H3.3 cap site was incorrectly assigned.

### Analysis of the 3' UTR and downstream sequences.

The 3′ end of the mRNA of the human H3.3 gene has been established previously by the isolation of 20 cDNA clones all containing the same 3′ termini. This point has been marked with an asterisk (*) in Figure 3, below base +8797, and immediately follows two ATTAAA sequences centered at +8780 and +8790. This 3′ terminus is also confirmed by three reverse transcribed pseudogene sequences shown in Figure 10.

```
Ch      GCGGGGCCTTCCCCCCATTGTGCCCCATTGGCCG---GCGGGCGGGGCGG
        :::  :::::::: ::::::::    :::::: :   :::::::::::::
Hu      GCGA-GCCTTCCCTCCATTGTGTGTGATTGGCTGCGCGCGGCGGGGGCGG
         -90        -80       -70       -60       -50

Ch      GGGCGCC---------GACCCGGCGGCGGCCCRCCCATCTTTCAATTGTG
        ::  ::          :    : :   :   :: ::::::::::::::::
Hu      GGCGGCGTGTGTTGGGGGATAGCCTCGGTGTCAGCCATCTTTCAATTGTG
          -40       -30       -20       -10        +1

Ch      CTCGCAGTCCGCCGCCGC--CGCCAGCAGCAGCCGCCG-TAGCGCCGGCT
         :::::: ::::::::::  ::::  :     ::  ::  ::::::: ::
Hu      TTCGCAG-CCGCCGCCGCGCCGCCGTCGCTCTCCAACGCCAGCGCCGCCT
          10         20        30        40        50

Ch      CCTCCTCGCGCCCGCCTCGGCCCGCCCTC
        :   :::::      :  : :
Hu      CTCGCTCGCCGAGCTCCAGCCGAAGAGAA
          60         70        80
```

Figure 7. Sequence comparisons between human and chicken 5' sequences. Sequence numbering corresponds to the human H3.3 gene (figure 3). Ch, Chicken H3.3 gene; Hu, Human H3.3 gene. Sequence number +1 indicates the cap site of the human gene. The conserved consensus sequences in the human H3.3 promoter are underlined.

Brush et al. suggested from S1 nuclease protection experiments that the 3'UTR of the chicken H3.3 mRNA extended 250 nucleotides past the termination codon even though no established polyadenylation signal immediately precedes this location. The conclusion of these authors may be incorrect, based on the comparison between the chicken and human sequences shown in Figure 8. The human 3′ UTR is 520 bases in length while the chicken 3′ UTR as assigned by Brush et al. is only 250 bases in length. Comparison of the first 250 bp of the human and chick 3′UTRs reveals 84% similarity of the sequences, allowing one large insert in the human gene for a polyadenosine tract centered at +8318. Strikingly, this similarity continues 270 bases downstream of the assigned end of the chicken gene and stops precisely at the 3′ end of the human gene. The similarity of this 270 bases region is 87.5%, again allowing for one large insert, this time for a polyadenosine tract in the chicken gene. Thus the region of similarity within the 3'UTRs of the chicken and human genes extends a total of 520 bases and shows 85% base pair matching overall. Both 3′ regions contain a long polyadenosine stretch, notably in different locations. The final 158 bases of this region shows the most similarity between the two homologs (93%) (Figure 8).

The sequences beyond the putative end of the chicken mRNA may be conserved for non-message related purposes; however, this type of conservation is unprecedented. A second possibility is that there are two potential mRNA stop sites in the chicken gene and the downstream region of similarity actually represents an alternative mRNA sequence. More likely is that the original assignment of the 3′ termini of the chicken gene is in error and that the actual terminus is at the end of the region of similarity, 520 bases downstream from the translation stop. No cDNA or processed pseudogene has been reported for the chicken H3.3 gene which

```
                           .                 8300
Hu  ATACGTGGAGAACGTGCTTAAGAATCCACTATGATGGGAAACATTTCATTCTCAAAAAAA
    :: ::::: :: ::::: :::   : :::  :: ::::     : :::::::: :  :::
Ch  ATTCGTGGGGAGCGTGCCTAAACTT-CACCTTGGTGGGT---TTGTCATTCTCGAGCAAA

       .                   .                   .
Hu  AAAAAAAAATTTCTCTTCTTCCTGTTATTGGTAGTTCTGAACGTTAGATATTTTTTTTCC
                  :::::::: ::::::::::::  :::::::::::::::::::  ::
Ch  --------------CTTCTTCC-GTTATTGGTAGTAATGAACGTTAGATATTTTTT--CC

       .                 8400                  .
Hu  ATGGGGTC-AAAGGTACCTAAGTATATGATTGCGAGTGGAAAAATAGGGGACAGAAATCA
    :::::::  ::::::::::::::::::: :::: : :::::::  ::::: ::: :::
Ch  ATGGGGTTGAAAGGTACCTAAGTATATGGTTGCAAATGGAAAARAAGGGGTCAG-AATTG

        .                       .                   .
Hu  GGTATTGGCAG----TTTTTCCATTTTCATTTGTGTGTGAATTTTTAATATAAATGCGGA
    ::::::::::     ::::::::::::::::::: :::: :::::::::: ::: : ::
Ch  GGTATTGGCAAGTTTTTTTTCCATTTTCATTTGT-TGTGGATTTTTAATAGAAACGAGGG

         8500                     .                   .
Hu  GACGTAAAGCATTAATGCAAGTTAAAA--TGTTTCAGTGAACAAGTTTCAGCGGTTCAAC
    :::  ::: : :::::::      :::   :: :  :::::::::::::: :  ::::::
Ch  GAC--AAAACGTTAATGCGGTCACAAACAAGTCTGGGTGAACAAGTTTCAACAATTCAAC
                                                      *

              .                                    .
Hu  TTTATAATAAT----------------TATAAATAAA-CCTGTTAAATTTTTCTGGACAA
    ::::::: ::                 :::::::::: :::::::: :::::::::::::
Ch  TTTATAAAAAAAAAAAGAAAAAAAAAGTATAAATAAAGCCTGTCAA-TTTTTCTGGACAA

        8600                   .                       .
Hu  TGCCAGCATTTGGATTTCTTTAAAACAAGTAAATTTCTT---ATTGATGGCAACTAAATG
    ::::::::::::::::: :::: :::  ::::: : :::   ::::::   ::: :::::
Ch  TGCCAGCATTTGGATTTTTTTA-AACGGGTAAA-TTCTTCCAATTGGTT--AAC-AAATG

               .                   .                8700
Hu  GTGTTTGTAGCATTTTTATCATACAGTAGATTCCATCCATTCACTATACTTTT-CTAACT
    :::::::::::::::: :::::::::::: ::::::::::::::: ::::::: ::::::
Ch  GTGTTTGTAGCATTTT-ATCATACAGTAGTTTCCATCCATTCACTCTACTTTTTCTAACT

                .                   .                    .
Hu  GAGTTGTCCTACATGCAAGTACATGTTTTTAATGTTGTCTGTCTTCTGTGCTGTTCCTGT
    ::::::::::::::::::::::::::::::::::::::::: ::::::::::::::::::
Ch  GAGTTGTCCTACATGCAAGTACATGTTTTTAATGTTGTCTGCCTTCTGTGCTGTTCCTGT

                  .                8800                   .
Hu  AAGT--TTGCTATTAAAATACATTAAACTATACTGCTTTTGGTCTTTATTATAGCCTTGC
    :  :  :::  :::::: :::::::::  ::◆ ::   : : : :    :    : :
Ch  AGTTAGTTGTCATTAAA-TACATTAAATGATGATGTCCTGGTTNTAGTCTGATTCTTAAA
```

Figure 8. Sequence comparisons between human and chicken 3' sequences. Sequence numbering corresponds to the human H3.3 gene (figure 3). Ch, Chicken H3.3 gene; Hu, Human H3.3 gene. The termination codons are indicated in **Bold Type.** The A residue in the chicken sequence just above the (*) indicates the 3' terminus of the chicken mRNA assigned by Brush et al. (1985). The A residue in the human sequence just above the ♦ indicates 3' the terminus of the human mRNA.

could confirm the 3′ end of the gene. Although the S1 nuclease data presented by Brush et al. (2) appears to be unambiguous, a potential source of error should be noted. Just downstream from the S1 mapped 3′ end of the message is a long stretch of 45 nucleotides which contains over 93% A+T residues (42 of 45). This region could produce an artificial 3′ end due to "breathing" of the RNA-DNA hybrid during S1 digestion.

The H3.3 multigene family.

We have previously observed (1) that when probes derived from either the 5′ end or the 3′ end of the H3.3 cDNAs were hybridized to human genomic DNA digested with a variety of restriction enzymes, the patterns of restriction fragments hybridizing were almost indistinguishable (also see Figure 1). This indicates a lack of restriction sites for these enzymes within the genomic segments that separate these probes. Since these two probes are separated in the expressed gene by over 9 kb of DNA and since its introns appear to be single copy, we predicted that the other H3.3 hybridizing fragments represent either expressed genes with small, highly diverged introns, or processed pseudogenes. In order to distinguish between these possibilities and to further characterize the H3.3 multigene family in human DNA, the 19 recombinant genomic clones which did not hybridize to the intron probe were partially restriction mapped and hybridized to 3′ and coding region probes derived from the cDNA. This analysis

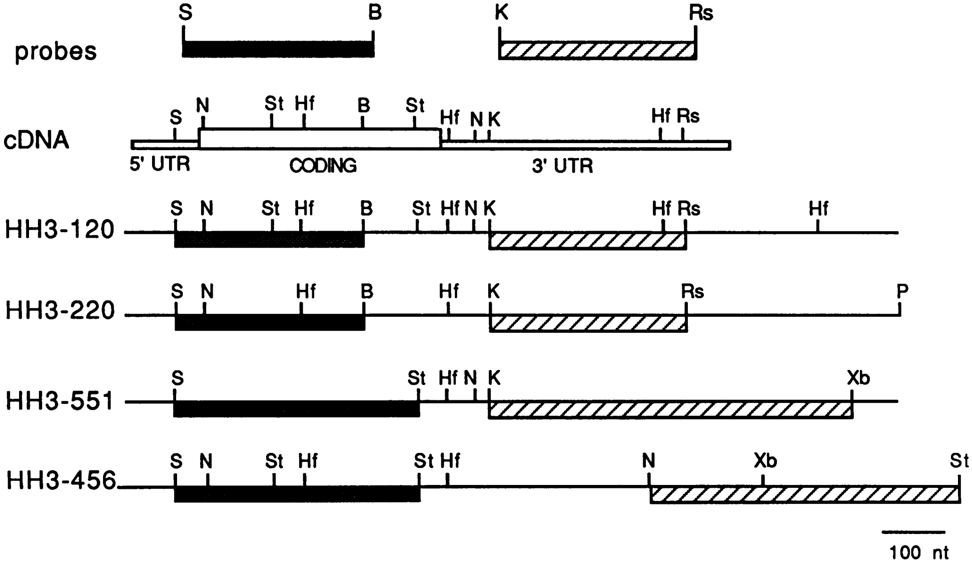

Figure 9. Restriction maps of the H3.3 cDNA and four H3.3 pseudogenes. Pseudogenes were restriction mapped and restriction fragments were hybridized to the nick-translated probes from the cDNA as indicated here and described in the Methods. Restriction maps were aligned and regions hybridizing to the 5'SB (stippled boxes) and the 3'KR (stripped boxes) probes are indicated. Restriction enzyme abbreviations are as follows: S, SacI; N, NcoI; St, StuI; Hf, HinfI; B, BglII; K, KpnI; Rs, RsaI; P, PstI; Xb, XbaI.

```
cDNA      GTGTTCGCAGCCGCCGCCGC         GCCGCCGTCGCTCTCCAACGCCAGCGCCGCCTCTC
 120       ...................         ...................................
 456      ....................CACCGCCAT......A..A.................T.......
 551    TT.....T..........TGC.         .G.................................

cDNA    GCTCGCCGAGCTCCAGCCGAAGAGAAGGGGGGTAAGTAAGGAGGTCTCTGTACCATGGCTCGTACAAA
 120    ......................G.........................N...................
 456    .................T..G..T.....A.....A......A.C....A..........T.......
 551    ................T...G.....N...A...........A......A....T.............

cDNA     GCAGACTGCCCGCAAATCGACCGGTGGTAAAGCACCCAGGAAGCAACTGGCTACAAAAGCCGCTCGCAA
 120    ...................N.................................................
 456    ..G.......T............----....A..G.......AG......G..............A...
 551    ..CA.G..........CT.G.-............-..........................A.......

cDNA    GAGTGCGCCCTCTACTGGAGGGGTGAAGAAACCTCATCGTTACAGGCCTGGTACTGTGGCGCTCCGTG
 120    .................................................A..................
 456    ....................N.................A.....................A.......
 551    ..C.....................................A.T.................A.......

cDNA    AAATTAGACGTTATCAGAAGTCCACTGAACTTCTGATTCGCAAACTTCCCTTCCAGCGTCTGGTGCGA
 120    ................................................................C...
 456    ....C...............G...........................G...T.............A.
 551    .......G.....CT.....G..........---A...................G.............

cDNA    GAAATTGCTCAGGACTTTAAAACAGATCTGCGCTTCCAGAGCGCAGCTATCGGTGCTTTGCAGGAGGC
 120    .............................NN.....................................
 456    .................G......A....A...........T........T..A..............
 551    .................G.........T..T........-----......................T.

cDNA    AAGTGAGGCCTATCTGGTTGGCCTTTTTGAAGACACCAACCTGTGTGCTATCCATGCCAAACGTGTAA
 120    ........................................................G...........
 456    ............C......................G...........................GC...
 551    ...............T...............A............................G.A.....

cDNA    CAATTATGCCAAAAGACATCCAGCTAGCACGCCGCATACGTGGAGAACGTGCTTAAGAATCCACTATG
 120    ....................................................................
 456    .T..............A......................AA...A.......................
 551    ..G.............T.............T.....C...............................

cDNA    ATGGGAAACATTTCATTCTCAAAAAAAAAAAAAAAATTTCTCTTCTTCCTGTTATTGGTAGTTCTGAA
 120    ....................................................................
 456    .........................G............C....................G...A....
 551    ...............................------..A....A.......................

cDNA    CGTTAGATATTTTTTTTCCATGGGGTCAAAGGTACCTAAGTATATGATTGCGAGTGGAAAAATAGGGG
 120    .................................................N..................
 456    G.............................C..........G....N....--C...C.........T
 551    .A.............CC.................................A......C..........
```

```
cDNA  ACAGAAATCAGGTATTGGCAGTTTTTCCATTTTCATTTGTGTGTGAATTTTTAATATAAATGCGGAGA
 120  ...............................................................TT...
 456  .G.C..........G.....C...............................--........T...C.
 551  ..................................G............................A.G..

cDNA  CGTAAAGCATTAATGCAAGTTAAAATGTTTCAGTGAACAAGTTTCAGCGGTTCAACTTTATAATAATTA
 120  ...............................G.....................................
 456  T......................................G........A....................
 55   G..............................C.....................................

cDNA  TAAATAAACCTGTTAAATTTTTCTGGACAATGCCAGCATTTGGATTTCTTTAAAACAAGTAAATTTCTT
 120  ...............................................T.....................
 456  ........................A............................................
 551  ...............................................T.....................

cDNA  ATTGATGGCAACTAAATGGTGTTTGTAGCATTTTTATCATACAGTAGATTCCATCCATTCACTATACTT
 120  ...................................................T.................
 456  .....CA..............................N.C...........-.............GT..
 551  .....CA..........A....G.................T.........T..............T..

cDNA  TTCTAACTGAGTTGTCCTACATGCAAGTACATGTTTTTAATGTTGTCTGTCTTCTGTGCTGTTCCTGTA
 120  .....................................-...............................
 456  ..............................G.....................G................
 551  ............................C....G...................C..............

cDNA  AGTTTGCTATTAAAATACATTAAACTATAAAAAAAAAA
 120  ..........C...........................
 456  ...A.............T..........T.........
 551  ........................-..A........
```

Figure 10. The nucleotide sequence of three H3.3 pseudogenes. Sequence analysis was performed as described in the Methods. The sequence of the cDNA is presented on the top line. Positions where the pseudogenes are identical with the cDNA are marked with a (·). Positions where the pseudogene sequence differs from the cDNA are noted. Dashs (-) indicate deletions . N's indicate an undetermined base. Note one insertion within pseudogene 456 occurs within the first part of the gene.

revealed 10 non-coincident maps and that each clone hybridized at high stringency to both cDNA probes. The hybridizing DNA fragments of each clone totaled less than 4 kb in length (data not shown). Four cloned DNA inserts with different restriction maps were selected for additional study. Their restriction endonuclease maps are presented in Figure 9 along with the H3.3 cDNA map. It is apparent from this analysis that the 5′ and 3′ ends of at least two of the four clones (HuH3-120 and HuH3-220) were too close to contain introns and thus were likely to be processed pseudogenes. This was ultimately confirmed for HuH3-120 by direct sequence analysis (Figure 10 and Discussion). The topological maps of the other two clones (HuH3-456 and HuH3-551) did not exclude the presence of introns and the map of HuH3-456 appeared to be too long to be a processed pseudogene. However sequence analysis of HuH3-456 and

HuH3-551 demonstrates that they too lack introns and contain a poly A stretch immediately adjacent to the 3′ end of the gene (Figure 10). The extra length of HuH3-456 is accounted for by the presence of an Alu sequence insert. In addition, all three pseudogenes were flanked by target site duplications. Thus the three genes that were sequenced (and probably HuH3-220) are processed pseudogenes. This data strongly supports our previous suggestion that most of the 20-30 members of the H3.3 multigene family are processed pseudogenes.

### Characterization of the pseudogenes.

The HuH3-456 pseudogene contains two Alu I repeats inserted at different locations in the 3′ UTR in opposite orientations. This discovery is especially significant since these Alu inserts must represent recently transposed sequences, since they are inserted in a recently generated pseudogene (manuscript in preparation). The locations of these Alu I segments are noted in Figure 10 but the sequences are not shown so as to not disrupt the alignment.

One of the pseudogenes (HuH3-120) is only 1% diverged from the cDNA, including only 2 amino acid changes at positions 44 and 114. Pseudogenes HuH3-456 and HuH3-551 are both diverged approximately 7% at the nucleotide level. The rate of accumulation of neutral point mutations in pseudogenes during evolution has been estimated at ~1.5% divergence per million years (22). This suggests that the HuH3-120 pseudogene arose less than 1.0 million years ago, while both the HuH3-551 and HuH3-456 pseudogenes arose about 5 million years ago.

## DISCUSSION

We have isolated and sequenced the human gene that encodes the basally expressed H3.3 histone. Its primary transcript is 8.8 kb in length and contains 3 introns totalling almost 8 kb in length. The 5′ promoter region contains unusual equivalents to the TATA and CCAAT regions at -25 and -77 respectively. In addition, the 5′ flanking region contains a high G+C content and at least one GC box. These GC boxes are often seen in the promoters of constitutively expressed "housekeeping" genes. Dynan and Tjian (21) have suggested that the interaction of Sp1 with the GC box could provide a basal level of transcription and that this basal level could then be subject to modulation by other regulatory factors. The presence of such a potential Sp1 binding site in the basally expressed H3.3 gene promoter is consistent with this hypothesis.

Particularly surprising are the intron-exon splice borders of the H3.3 gene. Not only do they contain the standard splice consensus sequences, but in all cases the introns are flanked by 7-8 base pair direct repeats. The function, if any, of these repeats is unclear, since the repeats include both intron and exon bases. One functional difference between these introns can be inferred from the structures of the previously isolated cDNAs (1). Three of the cDNAs were shown to contain an unspliced intron 1, but did not carry introns 2 and 3. This could reflect the preferential splicing out of introns 2 and 3 before the splicing out of intron 1. If there is a tendency toward 5′ to 3′ splicing, the unusual splice junctions seen for the H3.3 gene could act to supersede this tendency. The advantage to the organism to remove intron 1 last is unclear but could point to some as yet undetermined function for this intron. In support of this, we have found that a DNA probe derived from intron 1 hybridizes to a single fragment in a Southern blot of total mouse genomic DNA indicating that the sequences in this intron may be conserved, whereas a DNA probe derived from intron 2 does not hybridize (data not shown).

One of the most interesting aspects of the H3.3 gene is the extreme conservation of its entire 3′ UTR over 250 Myr of evolution between *Homo sapiens* and *Gallus domesticus*. This is particularly interesting since this homology extends 270 bases downstream from the mapped end of the chicken mRNA. The simplest explanation for this ambiguity is that an error was made in assigning the 3′ end of the chicken gene. Several points support this explanation in addition to the evidence provided by the continuous region of sequence matching between the human and chicken genes downstream from the putative 3′ terminus of the chicken transcript. First, no consensus poly A addition signal immediately precedes the mapped end of the chicken H3.3 gene. Secondly, the chicken gene does contain the same polyadenylation signal sequence as the human gene, **ATTAAA,** at the end of the region of similarity between the human and chicken genes.

The possibility remains, however, that there are two 3′ termini in the chicken H3.3 gene. A second potential polyadenylation signal is present in both the chicken and human genes and is centered at +8567 (Figure 7). This sequence, **AATAAA,** is the most commonly utilized and efficient polyadenylation-termination signal (23). If this site is not utilized it would strongly support the concept that additional cis-acting elements are required for proper 3′ processing (24,25).

It would be remarkable if, as we suspect, the 3′UTR of the chicken H3.3A mRNA proved to extend over the entire 520 base segment. Sequence conservation of this extent in 3′UTR segments of homologous genes is unprecedented. In addition to the 85% sequence matching, the presence of internal poly A stretches in both sequences is also intriguing. Both human and chicken H3.3 genes have this internal stretch of A residues albeit not in the same location. That these poly A tracts are located in a region of 85% similarity implies that they may have evolutionary and/or functional significance. Also of interest is a 124 bases stretch at the 3′ terminus with 96% similarity between the chicken and human genes. The closest analogy to this remarkable 3′ conservation comes from actin genes, which also show isotype-specific 3′ UTR regions which are conserved over long evolutionary periods (26, 27, 28). Specifically, the β-actin gene is 60% conserved over its entire 3′ UTR with short localized regions of greater than 95% similarity. The functions of the 3′ UTRs in the β-actin and the H3.3 mRNAs are unknown. Extensive sequence matching of this extent is generally not associated with binding sites for regulatory factors, which generally require less specificity over a smaller region.

There are also indications that the H3.3 5' UTR is conserved over long evolutionary periods. The comigration of human and mouse 5' S1-protected fragments (Figure 6) suggests that considerable sequence similarity exists between the mouse and human H3.3 5' UTRs.

Between human and chicken sequences, the sequence alignments generated for the 5′ upstream regions reveal only one significant region of similarity (Figure 7). This region includes the distal end of the 5'UTR and extends more than 90 bases upstream into the promoter region of the human gene. This region of similarity does not correspond to the presumptive 5′ UTR and proximal promoter region in the chicken gene identified by Brush et al. (2), but, rather, is 600 bp upstream. However, as stated by Brush et al., the assignment of the 5′ cap site in the chicken H3.3 gene was largely based upon sequence matching to other polymerase II transcribed genes since their result using S1 nuclease analysis was ambiguous. We propose that the 5′ region of similarity in the chicken gene shown in Figure 7 corresponds to the 5'UTR and promoter region of the chicken H3.3A gene. Thus the cap site assignment by Brush et al., who did not have the advantage of a second H3.3 gene to use for comparison, is either incorrect or represents an alternative transcription initiation site.

Analysis of the H3.3 multigene family reveals at least 20 Eco RI fragments in the human genome which hybridize to a 350 bp region of the 3′ UTR and only one which hybridizes to a 350 bp intron probe (Reference 1 and Figure 1). Eleven of these have been isolated and partially characterized. One (HuH3-149) represents the 8.8 kb expressed gene encoding the cDNA isolated previously. Of the other ten, none hybridize to the intron probes generated from the expressed gene. All, however, do contain within at most 4 kb, sequences hybridizing to both 5′ and 3′ probes. This indicates that these 10 randomly isolated clones contain either no introns or at most small introns. Four of these ten were subcloned and demonstrated to be intronless. Three of these were shown by direct sequence analysis to be processed pseudogenes. We interpret these results to imply that most of the members of the human H3.3 multigene family are processed pseudogenes and that there are at most a few expressed genes. The possibility that there is more than one H3.3 gene is supported by the discovery in chickens of two expressed genes that encode an H3.3 protein (2). The chicken H3.3A gene is closely related to the human H3.3 gene as described here. The chicken H3.3B gene, however, does not show sequence similarity in noncoding segments with either the H3.3A or the human H3.3 gene. Thus a second human H3.3 gene, the homolog of the chicken H3.3B gene, may yet remain to be isolated.

## ACKNOWLEDGEMENTS

We thank the NIH and the Veterans Administration for support. Doug Hoffman was supported by PHF Grant # CA09302, awarded by the National Cancer Institute DHHS. Lena Arcinas provided expert technical assistance. We thank our colleagues at the MEDIGEN Project for their advice as well as their unsolicited comments. Computer facilities for part of this work were provided by the NIH-supported SUMEX-AIM facility at Stanford University.

[1] Present address: Department of Biology, University of Houston, Houston, TX 77004, USA

*To whom correspondence should be addressed at: Veterans Administration Medical Center, (151M), Miranda Avenue, Palo Alto, CA 94304, USA

## REFERENCES

1. Wells, D. E., and L. Kedes. (1985). *Proc. Natl. Acad. Sci. U.S.A.* 82:2834-2838.
2. Brush, D., J. B. Dodgson, O. Choi, P. W. Stevens, and J. D. Engel. (1985). *Mol. Cell. Biol.* 5:1307-1317.
3. Wells, D. E. (1986). *Nucleic Acids Res.* 14(Supplement): r119-r149.
4. Birchmeier, C., D. Schumperli, G. Sconzo, and M. L. Birnstiel. (1984). *Proc. Natl. Acad. Sci. U.S.A.* 81:1057-1061.
5. Old, R. W. and H. R. Woodland. (1984). *Cell* 38:624-626.
6. Zweidler, A. (1984). In G. S. Stein, J. L. Stein, and W. F. Marzluff (ed.), *Histone genes*. John Wiley & Sons, Inc., New York, p. 339-371.
7. Wu, R.S. and W.M. Bonner. (1981). *Cell* 27:321-330.
8. Wu, R.S., S. Tsai and W.M. Bonner. (1983). *Biochemistry,* 22:3868-3872.
9. Wells, D., W. Bains, and L. Kedes. (1986). *J. Mol. Evol.*, in press.
10. Maniatis, T., R. C. Hardison, E. Lacy, J. Laver, C. O'Connel, D. Quon, G. K. Sim, and A. Efstradiatis. (1978). *Cell* 15:687-701.
11. Benton, W. D., and R. W. Davis. (1977). *Science* 196:180-182.
12. Maniatis, T., E. F. Fritsch, and J. Sambrook. 1982. *Molecular cloning: a laboratory manual.* Cold Spring Harbor Laboratory.
13. Bains, W. (1986). *Nucleic Acids Res.* 14:159-178.
14. Benoist, C., K. O'Hare, R. Breathnach, and P. Chambon. (1980). *Nucleic Acids Res.* 8:127-142.
15. Breathnach, R., and P. Chambon. (1981). *Ann. Rev. Biochem.* 50:349-383.
16. Sures, I., J. Lowry, and L.H. Kedes. (1978). *Cell* 15:1033-1044.
17. Hamada, H., M. Seidman, B. H. Howard, and C. M. Gorman. (1984). *Mol. Cell. Biol.* 4:2622-2630.
18. Nordheim, A., and A. Rich. (1983). *Nature* 303:674-678.
19. Singer-Sam, J., D. H. Keith, K. Tani, R. L. Simmer, L. Shively, S. Lindsay, A. Yoshida, and A. D. Riggs. (1984). *Gene* 32:409-417.
20. Stone, E. M., K. N. Rothblum, M. C. Alvey, T. M. Kuo, and R. J. Schwartz. (1985). *Proc. Natl. Acad. Sci. U. S. A.* 82:1628-1632.
21. Dynan, W. S., and R. Tjian. (1985). *Nature* 316:774-778.
22. Miyata, T., and T. Yasunaga. (1981). *Proc. Natl. Acad. Sci. U. S. A.* 78:450-453.
23. Wickens, M., and P. Stephenson. 1984. Science 226:1045-1051.
24. McLauchlan, J., D. Gaffney, J.L. Whitten and J.B. Clements. (1985). *Nuc. Acids Res.* 13:1347-1368.
25. Sadofsky, M., S. Connelly, J.L. Manley, J.C. Alwine. (1985). *Mol. Cell Biol.* 5:2713-2719.
26. Ng, S., P. Gunning, R. Eddy, P. Ponte, J. Leavitt, T. Shows, and L. Kedes. (1985). *Mol. Cell. Biol.* 5:2720-2732.
27. Ponte, P., S. Ng, J. Engel, P. Gunning, and L. Kedes. (1984). *Nucleic Acids Res.* 12:1687-1696.
28. Yaffe, D., U. Nudel, Y. Mayer, and S. Neuman. (1985). *Nucleic Acids Res.* 10:3723-3737.