Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 2834–2838, May 1985
Genetics

Structure of a human histone cDNA: Evidence that basally expressed histone genes have intervening sequences and encode polyadenylylated mRNAs

(DNA cloning/cell cycle/histone mRNA)

DAN WELLS AND LARRY KEDES

The MEDIGEN Project, Department of Medicine, Stanford Medical School and Veterans Administration Medical Center, Palo Alto, CA 94305

Communicated by Stanley N. Cohen, December 26, 1984

ABSTRACT **We have isolated and sequenced full-length cDNA clones encoding the human basally expressed H3.3 histone from a human fibroblast cDNA library. Several features of this atypical cDNA distinguish it and its gene from the well-characterized cell-cycle regulated histone genes and their RNA transcripts. The H3.3 mRNA is ≈1200 bases long, contains unusually long 5′ and 3′ untranslated regions, and has a 3′ polyadenylylated terminus. In addition, we have isolated and characterized a cDNA clone that is a precursor to the H3.3 mRNA and contains an intervening sequence interrupting its 5′ untranslated region. Hybridization of subsegments of the cDNA to human genomic DNA reveals a complex multigene family. The differences in the structures of basal and cell-cycle histone genes suggest a model to explain the differences in their expression.**

A number of fundamental features of the histone genes in higher vertebrates are similar to those classically observed in sea urchins, *Drosophila melanogaster*, and lower vertebrates such as the newt [reviewed by Maxson *et al.* (1, 2)]. The genes for each of the five classes of histone proteins in higher vertebrates are repeated tens of times in each genome, contain no intervening sequences, and encode short transcripts that undergo no further processing or polyadenylylation. Unlike the tandemly repeated histone genes of sea urchins, *Drosophila*, and the newt, however, the higher vertebrate gene sets are loosely clustered with no consistency in their linked arrangement (3–6).

Most mammalian histone proteins are synthesized only during S phase as a function of regulation of both RNA synthesis and degradation (7, 8). In addition to these cell-cycle-dependent histones, there are histone variants expressed throughout the cell cycle as well as in quiescent and terminally differentiated cells (9–11). These basal or constitutive histones have a slightly altered amino acid sequence. Although they constitute only 5%–10% of histone protein synthesis during S phase, they are the predominant histones synthesized in nondividing cells. The genes that encode these basal histones somehow escape the cell-cycle regulation.

We report here the cloning, characterization, and nucleotide sequence of a basally expressed human H3.3 histone cDNA. The transcript represented by this cDNA contains several features that make it strikingly different from cell-cycle histone transcripts: it is polyadenylylated, contains lengthy 5′ and 3′ leader and trailer sequences, and does not contain the 3′ hyphenated dyad symmetry segment. Furthermore, we provide evidence that the gene for this basal histone contains at least one intervening sequence. In addition, we have reanalyzed data from the literature and show that some vertebrate genes that share these striking structural features are probably basal histone genes. These observations suggest ways in which the differences between cell cycle and basal histone gene structure might explain the differences in their expression.

MATERIALS AND METHODS

General Methods. Plasmid DNA preparation, restriction enzyme digestions, agarose gel electrophoresis, DNA and RNA blotting to nitrocellulose, and isolations of DNA fragments, total RNA, and polyadenylylated RNA were performed by standard techniques as described (5, 12). Nick-translations were done using the method of Rigby *et al.* (13) and DNAs were labeled to a specific activity of $\approx 10^8$ cpm/μg followed by precipitation of the DNA in 4 mM spermine HCl. Spermine pellets were washed in water and resuspended in 0.5% sodium lauryl sulfate/10 mM EDTA prior to denaturation and hybridization. Hybridizations were carried out between 37°C and 45°C in 50% formamide/10% dextran sulfate/5× SET (1× SET = 0.15 M NaCl/25 mM Tris, pH 8.0/2 mM EDTA)/0.02% polyvinylpyrrolidone/0.02% Ficoll/0.02% bovine serum albumin.

Screening the cDNA Library. A human fibroblast cDNA library (14) was kindly provided by H. Okayama and P. Berg (Stanford Medical School). We selected for cDNAs >1 kilobase (kb) by isolating linearized library DNA (15) >4.1 kb. The DNA was isolated from agarose gels, religated, and used to transform HB101 *Escherichia coli*. Ampicillin-resistant colonies were screened *in situ* by the method of Grunstein and Hogness (16), by using the radiolabeled eukaryotic pseudogene probe.

These cDNA clones were grown, isolated using standard procedures, and mapped by using single and multiple restriction enzyme digestions. All of the cloning procedures were carried out in accordance with the guidelines for recombinant DNA research issued by the National Institutes of Health.

DNA Fragment Probes. Four different fragments from the H3.3 cDNAs were isolated, nick-translated, and used as probes against DNA and RNA blots. The "coding region" probe used in the experiment described in Fig. 3 is a 500-base-pair fragment from the *Nco* I site at the AUG start codon to the *Nco* I site just 3′ of the TAA termination codon. The 3′ untranslated region (UTR) fragment used in the experiments described in both Figs. 3 and 4 is an *Hin*fI fragment, 400 base pairs long, containing only 3′ UTR sequences. The coding region probe hybridized to genomic blots in the experiment described in Fig. 4 contained all the sequence upstream of the *Bgl* I site within the coding region of pHH3B-2 including the 5′ UTR. The 5′ UTR probe from pHH3C-3 used to hybridize against RNA fractions described in Fig. 3 was a fragment 375 base pairs long containing all

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: kb, kilobase(s); UTR, untranslated region.

sequences upstream from the *Nco* I site at the AUG start codon of pHH3C-3.

DNA Sequence Analysis. All sequencing was done using the dideoxy method of Sanger *et al.* (17) after subcloning the cDNA inserts into the M13 mp8 vector. Sequencing was performed using M13 primers as described by Hu and Messing (18). For the insert in pHH3C-3, the sequence was determined from both strands multiple times over the entire length of the cDNA except for the 3'-most 70 base pairs, which were only sequenced in the 5' to 3' direction because of interference from the long poly(A) tail. Overlapping sequences were determined in all cases to ensure proper alignment. For the insert in pHH3B-2, the 5' half of the cDNA (bases 1–640) was sequenced in both directions by using overlapping fragments. The 3' half (bases 640–1040) was sequenced in only one direction. Only the 5'-most 150 nucleotides of the insert in pHH3A-1 were sequenced. DNA sequencing data were managed by the IntelliGenetics GEL program.

RESULTS

Isolation of Human H3.3 cDNAs. To isolate human basal histone gene sequences, we probed a human genomic λ phage recombinant DNA library (provided by T. Maniatis) with a mouse histone *H3.3* pseudogene that we had previously isolated (4). One of the human λ recombinants isolated from this screen is a reverse-transcribed *H3.3* pseudogene (unpublished observations). The structure of this human pseudogene suggested that transcripts of the authentic gene might be polyadenylylated. Accordingly, we used the subcloned human *H3.3* pseudogene (pHH3-1) to screen a size-selected human fibroblast cDNA library (see *Materials and Methods*). Approximately 150,000 colonies were screened, from which 20 positive clones were isolated.

Characterization of the cDNA Clones. The 20 clones fall into three distinct size classes (A, B, and C; see Fig. 1). Although cDNA inserts of 1100 base pairs and larger had been preselected, 15 of the 20 cDNA inserts are only ≈950 base pairs long (class A cDNAs; Fig. 1). Partial sequencing of the 5' end of one class A cDNA, pHH3A-1, shows that it lacks 145 nucleotides encoding the amino terminus of histone H3.3. Restriction maps obtained for the other 14 class A clones indicate that their 5' termini are located no more than 50 base pairs from the 5' terminus of pHH3A-1. Thus, none of the class A clones was full length cDNA.

Two of the 20 cDNA clones contain 1200-base-pair inserts (class B cDNAs; Fig. 1). Restriction maps of these two clones place their 5' ends ≈100 base pairs upstream from the AUG initiation codon. The three cDNAs making up class C are ≈1500 base pairs long (Fig. 1). Their 5' ends are ≈400 base pairs upstream from the initiation codon.

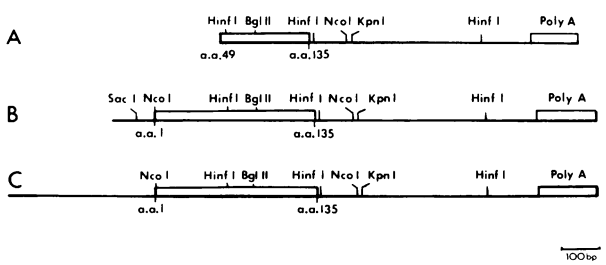

FIG. 1. Restriction maps of the human H3.3 cDNA clones. The cDNA clones were classified A, B, or C based on increasing lengths. Class B and class C cDNAs were homogeneous in length, and class A cDNAs varied only slightly in length at the 5' terminus (see text). The class A, B, and C clones represent pHH3A-1 (*A*). pHH3B-2 (*B*), and pHH3C-3 (*C*), respectively. The limited restriction maps of all cDNA clones were identical, with the exception of the *Sac* I site only present in the 5' UTR of class B cDNAs.

Our initial hypothesis was that these three cDNA classes represented increasingly longer reverse transcripts from the same template mRNA. However, there is a *Sac* I endonuclease site in the 5' UTR of the class B clones that is not present in any of the longer class C clones (see Fig. 1). Otherwise our initial overlapping restriction maps of all three classes were identical. We conclude that the B and C classes of cDNAs either represent transcripts from different genes or represent differently processed transcripts from a single gene with an intervening sequence in the 5' UTR. The DNA sequence analysis of these two cDNAs distinguished these two possibilities.

Sequence Analysis of Two H3.3 cDNAs. We determined the complete nucleotide sequences of a class C clone, pHH3C-3, and a class B clone, pHH3B-2, by dideoxy sequencing using M13 vectors. These sequences are presented in Fig. 2 and reveal several unusual features.

As expected from their presence in the cDNA libraries and from restriction maps, both cDNAs have long 3'-terminal poly(A) tails. The total length of pHH3C-3 [excluding its poly(A) tail] is 1305 base pairs, with 374 base pairs of 5' UTR and 520 base pairs of 3' UTR. The cloned inserts of pHH3B-2 and pHH3C-3 are identical from the poly(A) tail to 24 base pairs upstream from the AUG initiation codon, a total of 955 bases. We interpret the absence of *any* nucleotide differences in this 955-base segment to mean that class B and C cDNAs are transcribed from the same gene. An explanation of the differences in the 5' UTRs of these two classes of cDNAs was found by the direct nucleotide comparison.

Comparison of the 5' UTRs of pHH3B-2 and pHH3C-3. The 5' UTRs of pHH3B-2 and pHH3C-3 diverge starting at an A-G dinucleotide 24 base pairs upstream from the initiation codon. This divergence explains the *Sac* I restriction site difference described previously. The presence of an A-G dinucleotide in pHH3C-3, but not in pHH3B-2, is consistent with the possibility that this segment of pHH3C-3 is an intervening sequence. In addition, the 5' UTR of pHH3B-2 is G+C rich (75%) and has a remarkable over-representation of the dinucleotide CpG (18% compared to the genomic average of ≈1%). In contrast, the 5' UTR of pHH3C-3 is G+C poor (33%) and the frequency of CpG is 1.4%, close to the genomic average. This low G+C content of the 5' UTR is consistent with it being part of an intervening sequence (19). Taken together, these observations suggest that pHH3C-3 could be an incompletely reverse-transcribed processing intermediate with an intervening sequence and that the 5' end of pHH3B-2 represents the mature 5' end of the mRNA. The reverse transcript of pHH3C-3 must be incomplete and must have terminated in the 5' intervening sequence, because it lacks the 5' UTR of the mRNA represented by pHH3B-2. The likelihood of these and other possibilities will be considered further below.

Coding Region of pHH3B-2 and pHH3C-3. The nucleotide sequences of pHH3B-2 and pHH3C-3 are identical over the entire 411 base pairs of coding region. This sequence correctly predicts the amino acid sequence of an H3.3 protein. Specifically diagnostic of this H3 variant are the presence of isoleucine and glycine residues at positions 89 and 90, respectively, and a serine residue at position 96 (20). In addition to these three changes that distinguish H3.3 from H3.1 histones, the human H3.3 protein encoded by these cDNAs has two additional amino acid changes that distinguish it from H3.1. The alanine at position 31 of H3.1 is changed to a serine, and the serine at position 87 of H3.1 is changed to an alanine. These additional changes confirm those predicted by Ohe and Iwai from a peptide analysis of the H3.3 protein (21).

```
pHH3C-3     .....TTTTCTTTTGACTTGTTTGTGGATGGAATGTTTACAGACATTTCTAATTACTGCTTTAATTAAATAAATTGGATCAAAGGCCGTTCGAGGTATTTTTGTTTTGCCGTTTGT

pHH3C-3     CGCTCAGAATTGGCATTTTGAGAGGTGATTGATACTGCTAACAATTTTCTAGTACTCTAGTTTGTTTCAAGAAGAGATTTTGGGTAGACGTAATCTTCACCTTTCAAATTATATAACAAT

                                                        10        20        30        40        50        60        70        80
pHH3B-2                                     .....TGTTCGCAGCCGCCGCCGCGCCGCCGTCGCTCTCCAACGCCAGCGCCGCCTCTCGCTCGCCGAGCTCCAGCCGAAGAGAAGGGGG
pHH3C-3     ACGAACATTATTTTTTATACTGATCATAATTTCCAGATTTGGGGAGGGGGTGATCGTGGCAGGAAAAGTTGTATGTTTGTTAGTTGCATATGGTGATTTTTGATTTTTCAATGCTGGTAG

    90       100                                 128       138               153               168               183
GTAAGTAAGGAGGTCTCTGTACC ATG GCT CGT ACA AAG CAG ACT GCC CGC AAA TCG ACC GGT GGT AAA GCA CCC AGG AAG CAA CTG GCT ACA AAA GCC
                        Met Ala Arg Thr Lys Gln Thr Ala Arg Lys Ser Thr Gly Gly Lys Ala Pro Arg Lys Gln Leu Ala Thr Lys Ala

              198               213               228               243               258
GCT CGC AAG AGT GCG CCC TCT ACT GGA GGG GTG AAG AAA CCT CAT CGT TAC AGG CCT GGT ACT GTG GCG CTC CGT GAA ATT AGA
Ala Arg Lys Ser Ala Pro Ser Thr Gly Gly Val Lys Lys Pro His Arg Tyr Arg Pro Gly Thr Val Ala Leu Arg Glu Ile Arg

    273               288               303               318               333               348
CGT TAT CAG AAG TCC ACT GAA CTT CTG ATT CGC AAA CTT CCC TTC CAG CGT CTG GTG CGA GAA ATT GCT CAG GAC TTT AAA ACA
Arg Tyr Gln Lys Ser Thr Glu Leu Leu Ile Arg Lys Leu Pro Phe Gln Arg Leu Val Arg Glu Ile Ala Gln Asp Phe Lys Thr

          363               378               393               408               423
GAT CTG CGC TTC CAG AGC GCA GCT ATC GGT GCT TTG CAG GAG GCA AGT GAG GCC TAT CTG GTT GGC CTT TTT GAA GAC ACC AAC
Asp Leu Arg Phe Gln Ser Ala Ala Ile Gly Ala Leu Gln Glu Ala Ser Glu Ala Tyr Leu Val Gly Leu Phe Glu Asp Thr Asn

438               453               468               483               498               513
CTG TGT GCT ATC CAT GCC AAA CGT GTA ACA ATT ATG CCA AAA GAC ATC CAG CTA GCA CGC CGC ATA CGT GGA GAA CGT GCT TAA
Leu Cys Ala Ile His Ala Lys Arg Val Thr Ile Met Pro Lys Asp Ile Gln Leu Ala Arg Arg Ile Arg Gly Glu Arg Ala

      529       539       549       559       569       579       589       599       609       619       629       639
GAATCCACTA TGATGGGAAA CATTTCATTC TCAAAAAAAA AAAAAAAATT TCTCTTCTTC CTGTTATTGG TAGTTCTGAA CGTTAGATAT TTTTTTTCCA TGGGGTCAAA GGTACCTAAG

      649       659       669       679       689       699       709       719       729       739       749       759
TATATGATTG CGAGTGGAAA AATAGGGGAC AGAAATCAGG TATTGGCAGT TTTTCCATTT TCATTTGTGT GTGAATTTTT AATATAAATG CGGAGACGTA AAGCATTAAT GCAAGTTAAA

      769       779       789       799       809       819       829       839       849       859       869       879
ATGTTTCAGT GAACAAGTTT CAGCGGTTCA ACTTTATAAT AATTATAAAT AAACCTGTTA AATTTTTCTG GACAATGCCA GCATTTGGAT TTCTTTAAAA CAAGTAAATT TCTTATTGAT

      889       899       909       919       929       939       949       959       969       979       989       999
GGCAACTAAA TGGTGTTTGT AGCATTTTTA TCATACAGTA GATTCCATCC ATTCACTATA CTTTTCTAAC TGAGTTGTCC TACATGCAAG TACATGTTTT TAATGTTGTC TGTCTTCTGT

     1009      1019      1029      1039
GCTGTTCCTG TAAGTTTGCT ATTAAAATAC ATTAAACTAT AAA.....
```

FIG. 2. Nucleotide sequences of cDNAs pHH3B-2 and pHH3C-3. Only the nucleotides in the cDNA pHH3B-2 are numbered to emphasize the 5′ terminus of the mature mRNA. Since the two cDNA sequences are identical from nucleotides 86–1040, only one of the sequences is presented. The amino acids are inferred from the nucleotide sequence and represent the H3.3 histone, including the initiator methionine and terminator codons.

The encoded protein is 135 amino acids long, a length characteristic of all H3 proteins. Although the amino acid sequence is 94% similar to human H3.1, the coding-region nucleotide sequences of H3.1 (22) and H3.3 are only 78% similar (data not shown). The differences at isocoding positions approach maximal random divergence and suggest that the two sequences arose long ago.

3′ UTRs of pHH3B-2 and pHH3C-3. The 3′ UTRs of pHH3B-2 and pHH3C-3 are identical over the entire length of 520 base pairs [discounting the length of the poly(A) tail], making this 3′ UTR the longest reported for a histone gene transcript. The region is highly A+T rich (69%) and shows no obvious sequence similarity to the 3′ UTR of any other known H3 gene. Most other histone transcripts contain short 3′ UTRs (<50 base pairs) and possess a characteristic hyphenated dyad symmetry that functions in transcription termination and 3′ processing (23, 24). These characteristic features are not present in the H3.3 transcript described here, suggesting that its transcription may be terminated by a very different mechanism. The remote position of the only A-A-U-A-A-A sequence in the 3′ UTR of H3.3, ≈200 base pairs from the end of the mature mRNA, makes it unlikely that it functions as a processing signal. However, two A-U-U-A-A-A sequences near the 3′ end of the transcript (Fig. 2) could well function as processing sites. Although rare, such signal sequences have been observed in several mammalian RNAs (25).

In summary, the sequencing and mapping data unexpectedly suggest that a transcribed human histone *H3.3* gene has at least one intervening sequence in the 5′ UTR that is processed post-transcriptionally and generates a long polyadenylylated mRNA.

Characterization of RNA Transcripts. If pHH3C-3 does represent an alternatively spliced or promoted version of pHH3B-2 rather than a precursor, two transcripts of ≈1200 bases and 1500 bases and differing in size by ≈300 nucleotides should be detected in electrophoretically separated mRNAs. This is not the case.

Total or poly(A)$^+$ RNAs from HeLa cells were electrophoretically separated on agarose gels, blotted onto nitrocellulose, and probed with a radiolabeled 400-base-pair long *Hinf* I fragment from the 3′ UTR of pHH3B-2. The resulting autoradiograph (Fig. 3) demonstrates that only one transcript is detected. The size of this transcript, ≈1200 nucleotides, corresponds well with the size of pHH3B-2 including the poly(A) tail.

Additional evidence that the 1200-base H3.3 RNA does not correspond to pHH3C-3 was obtained by hybridizing a probe containing the putative 5′ UTR intervening sequences from pHH3C-3 to the same set of RNAs shown in Fig. 3. Although this probe and the 3′ UTR probe were similar in specific activity, length and G+C content, no hybridization was seen with the 5′ UTR probe, even with long exposure times (data not shown). This result demonstrates that pHH3C-3 type transcripts are present in very low abundance, at least in HeLa cells. Taken together, these data suggest that the pHH3C-3 cDNA is a copy of an intervening sequence bearing partially processed RNA precursor.

The human H3.3 cDNA also hybridizes to 1200-base transcripts in mouse L-cell RNA and both the coding regions and 3′ UTRs hybridize, but to different degrees (Fig. 3). Under conditions of high stringency, no hybridization is detected between a 3′ UTR probe from the human H3.3 cDNA and total mouse RNA, whereas the coding-region probe continues to hybridize (compare lanes *A*1 and *B*1 with lanes *C*1 and *D*1). Thus, the human and mouse H3.3 3′ termini appear to be highly conserved.

Evidence for an *H3.3* Multigene Family in Humans. Cell-cycle histone genes are present in multiple copies in human and other genomes (for a review see refs. 1 and 2). By contrast, the genomic representation of genes for the basally expressed H3.3 variants is unknown. To determine the copy number for *H3.3* genes, total genomic DNA was digested with either *Eco*RI or *Bam*HI, electrophoresed through agarose gels, blotted onto nitrocellulose, and hybridized to radiolabeled probes derived from different fragments of H3.3 cDNA (Fig. 4). Both a coding-region probe and a 3′ UTR probe hybridize to 20–30 *Eco*RI fragments in these genomic blots. These hybridization signals melt off differentially under increasingly stringent wash conditions (data not

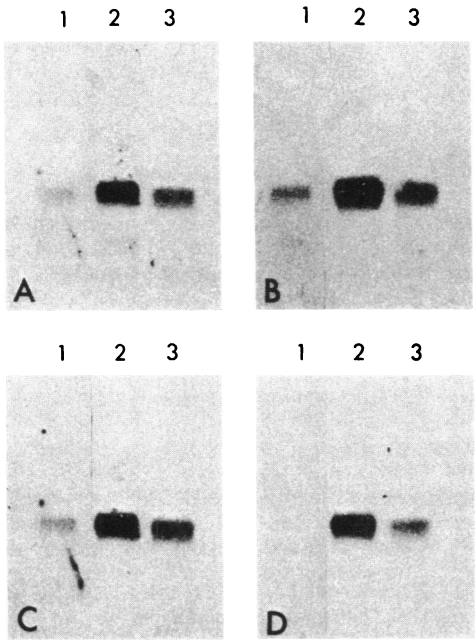

FIG. 3. Hybridization of cDNA segments to human and mouse RNA. Human and mouse RNAs were electrophoresed in 1% agarose gels in the presence of formaldehyde and blotted onto nitrocellulose paper. These filters were hybridized with radiolabeled fragments from either the 500-base-pair *Nco* I fragment from the coding region of the cDNA (*A* and *C*) or the 400-base-pair *Hinf*I 3′ UTR fragment (*B* and *D*). All hybridizations were done at 42°C in 50% formamide. The filters were then washed at 60°C in SET solution for 1 hr (*A* and *B*). After autoradiography, these same filters were rewashed for 1 hr at 60°C in 0.3× SET (*C* and *D*) and were exposed again to x-ray film. Lane 1 in each panel contains 5 μg of total RNA from mouse L-cells. Lane 2 in each panel contains 10 μg of total HeLa cell RNA. Lane 3 in each panel contains 0.1 μg of poly(A)$^+$ RNA isolated from the total HeLa RNA. A faint hybridizing band of 500 nucleotides can be seen in the original autoradiograms in *A* and *C*. We believe these represent the H3.1 histone mRNA.

shown). These results suggest that there is a complex *H3.3* multigene family in humans.

DISCUSSION

We have isolated, characterized, and sequenced cDNAs for the basally expressed human histone variant H3.3. We conclude that the H3.3 mRNA is an atypical histone transcript and the gene that encodes it has at least one intervening sequence that is located in the 5′ UTR. Our interpretation that the *H3.3* gene contains an intervening sequence is based on the following observations. First, the identity of the coding and 3′ UTR sequences of the class B and C cDNAs strongly suggests that they are transcribed from the same gene. Second, the nonidentical sequences at the 5′ end of these two cDNAs are therefore compatible with either different transcription initiation points for the two cDNAs or with the presence of a 5′ UTR intervening sequence. We favor the latter explanation because we detect only one size class of mRNA and the nucleotide sequences are characteristic of an intervening sequence. In either case, there must be processing of the 5′ UTR sequences during production of the mature mRNA. Recently, we have been able to distinguish these two possibilities by using a DNA segment from the putative intervening sequence to clone the human *H3.3* gene. Our sequence analysis and mapping of this gene confirm that it has an intervening sequence in the 5′ UTR and that pHH3C-3 is a truncated partially processed precursor to H3.3 mRNA represented by cDNA clone pHH3B-2 (unpublished data).

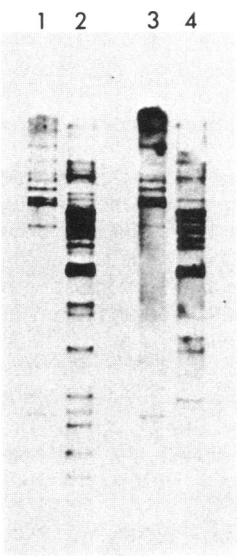

FIG. 4. Hybridization of cDNA fragments to human genomic DNA. About 8 μg of human DNA from HeLa cells was digested with either *Bam*HI (lanes 1 and 3) or *Eco*RI (lanes 2 and 4), electrophoresed onto 0.8% agarose gels, and blotted onto nitrocellulose filters. Genomic blots were then hybridized to radiolabeled probes from either the 5′ end of the coding region of the H3.3 cDNA (lanes 1 and 2) or the 3′ UTR (lanes 3 and 4). Hybridizations were done at 37°C in 50% formamide and washed in SET at 60°C for 1 hr.

The H3.3 mRNA is ≈1200 nucleotides long and contains long 5′ and 3′ UTRs and a poly(A) tail. The H3.3 cDNA represents a complex multigene family that hybridizes to 20–30 *Eco*RI genomic fragments. The *Eco*RI fragments show various degrees of homology to the cDNA probes. We believe it likely that most of these genes are reversed-transcribed pseudogenes and that there are probably one or, at most, a few expressed genes (unpublished).

The 20 cDNA clones that we isolated fell into three specific size classes. There is remarkably little or no size variation within each class. Class A clones represent cDNAs that were interrupted in the reverse-transcription process during cloning.

Class B cDNAs appear to represent the H3.3 mRNA present in both human and mouse cell lines (Fig. 3), and their presence in the cDNA library is expected. However, an explanation for the presence of three clones of the class C type is less straightforward, because they appear to represent precursors of the H3.3 mRNA. The presence of precursor RNAs within the Okayama–Berg human fibroblast library has been observed by others (26). On the other hand, transcripts representing class C cDNAs were not detectable in either HeLa or L-cell total cellular RNA and must be present in low abundance. One possible explanation for their presence among our isolated cDNA clones is that prior rounds of bacteriological amplification of the human fibroblast cDNA library may have skewed the representation of transcripts.

In trying to delineate the critical structural differences between the cell-cycle histone genes and the basally expressed histone gene variants, comparative data are critical. Unfortunately, very little is known about the primary nucleotide structure of variant histone gene transcripts. Three previous examples of cloned vertebrate histone gene variants have been reported, two in chicken (12, 27) and one in *Xenopus laevis* (28). Although the published data do not allow conclusions about the identity or functionality of the cloned sequences, our analysis of these data suggests that there are critical features in common with these sequences and the human H3.3 cDNAs described here.

Harvey *et al.* (27) have reported an extremely variant H2A nucleotide sequence found in a chicken cDNA library. Although the coding regions between this chicken H2A variant and the human H3.3 variant reported here cannot be compared directly, several interesting similarities exist between the two genes. Both have unusually long transcripts that are polyadenylylated. In addition, the results of blotting experiments have been interpreted (27) to infer the presence

of intervening sequences in the gene that gave rise to this chicken H2A variant transcript.

We compared the coding region of the human H3.3 cDNA with a chicken DNA segment (12) that hybridizes to H3.3-like chicken histone H3 RNAs. The hybridizing chicken RNAs probably are polyadenylylated because they bind to oligo(dT) cellulose (12). Of additional importance is that this chicken H3 variant also contains two intervening sequences within the amino acid coding sequences. The nucleotide sequence for the coding region of the chicken *H3* gene and the human H3.3 cDNA reported here show 88% similarity compared to only 78% sequence similarity between the coding regions of the human H3.3 sequence and the human *H3.1* gene. The 3′ halves of the two H3.3 coding sequences are 92% similar, and one 150-base stretch of the UTR is 95% similar (data not shown). The highly conserved nature of these regions of the chicken and human genes suggests that they are homologs and that the conserved segments have a strongly selected function.

Our analysis of published histone H3 sequences of other species has revealed a previously unrecognized H3.3 variant among the *X. laevis* histone cDNAs described by Ruberti *et al.* (28). This sequence was isolated from a cDNA library prepared from *X. laevis* oocyte polyadenylylated RNA. The sequence of this truncated cDNA contains most of the 3′ end of the coding region and includes amino acid changes at residues 87, 89, 90, and 96, diagnostic of an *H3.3* gene.

The conservation of (*i*) intervening sequences, (*ii*) polyadenylylation, and (*iii*) longer mRNA size as common structural features among histone variant genes and transcripts suggests that these features are selected and represent a common set of functional structures. In addition, the H3.3 nucleotide sequences, including the sequences of the 3′ UTR segments, are also highly conserved, suggesting that they are as highly selected as are the basal histone H3.3 peptides.

The differential accumulation of cell-cycle histones during S phase is a function of both RNA synthesis and degradation (8). Cell-cycle histone genes encode short (400–600 nucleotides) transcripts that are not polyadenylylated. These transcripts typically have very short 5′ and 3′ UTRs and their transcription is probably terminated in a manner very different from polyadenylylated genes. Finally, cell-cycle histone genes do not contain intervening sequences and seem to be uniquely adapted to rapid transport of their RNAs to the cytoplasm. In contrast, the *H3.3* basal histone gene contains at least one intervening sequence and produces long transcripts that are processed and polyadenylylated at their 3′ termini. These features could slow the transcription–translation process in addition to providing potential noncoding information. These distinguishing features might well account for some of the observed differences in the transcriptional and post-transcriptional regulation of the cell cycle versus basal histone genes and their transcripts. For example, proteins or other molecules might stabilize the basal transcripts preferentially over the cell-cycle transcripts and allow the basal mRNAs to escape the cell-cycle-dependent degradation of histone mRNAs. In addition, the 5′ untranslated region of the H3.3 basal histone mRNA, with a dramatically elevated G+C and CpG content may also play a significant role in the post-transcriptional regulation of these genes. We hypothesize that these critical structural features are characteristic of basally regulated histone genes and will prove important in explaining their basal, as opposed to cell-cycle, mode of regulation.

We thank Ron Cohn for helpful advice and Peter Evans and Lena Asterias for excellent technical contributions. This work was supported by Grant GM17995 from the National Institutes of Health and in part by a grant from the Veterans Administration. D.W. is a postdoctoral fellow of the American Cancer Society.

1. Maxson, R., Cohn, R., Kedes, L. & Mohun, T. (1983) *Annu. Rev. Genetics* **17**, 239–278.
2. Maxson, R., Mohun, T. & Kedes, L. (1983) in *Eukaryotic Genes: Their Structure, Activity and Regulation*, eds. MaxLean, N., Gregory, S. & Flavell, R. (Butterworths, London), pp. 277–298.
3. Heintz, N., Zernik, M. & Roeder, R. (1981) *Cell* **24**, 661–668.
4. Sittman, D., Chiu, I., Pan, C., Cohn, R., Kedes, L. & Marzluff, W. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 4078–4082.
5. Engel, J. & Dodgeson, J. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 2856–2860.
6. Old, R. W. & Woodland, H. R. (1984) *Cell* **38**, 624–626.
7. Pederson, T. (1976) in *Protein Synthesis*, ed. McConkey, E. H. (Dekker, New York), Vol. 2, pp. 69–123.
8. Heintz, N., Sive, H. L. & Roeder, R. G. (1983) *Mol. Cell. Biol.* **3**, 539–550.
9. Wu, R. S. & Bonner, W. M. (1981) *Cell* **27**, 321–330.
10. Wu, R. S., Tsai, S. & Bonner, W. M. (1982) *Cell* **31**, 367–374.
11. Wu, R. S., Tsai, S. & Bonner, W. M. (1983) *Biochemistry* **22**, 3868–3872.
12. Engel, J., Sugarman, B. & Dodgeson, J. (1982) *Nature (London)* **297**, 434–436.
13. Rigby, P. W., Dieckmann, M., Rhodes, C. & Berg, P. (1977) *J. Mol. Biol.* **113**, 237–251.
14. Okayama, H. & Berg, P. (1983) *Mol. Cell. Biol.* **3**, 280–289.
15. Gunning, P., Ponte, P., Okayama, H., Engel, J., Blau, H. & Kedes, L. (1983) *Mol. Cell. Biol.* **3**, 787–795.
16. Grunstein, M. & Hogness, D. S. (1975) *Proc. Nat. Acad. Sci. USA* **72**, 3961–3965.
17. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
18. Hu, N. & Messing, J. (1982) *Gene* **17**, 271–277.
19. McClelland, M. & Ivarie, R. (1982) *Nucleic Acids Res.* **10**, 7865–7877.
20. Franklin, S. G. & Zweidler, A. (1977) *Nature (London)* **266**, 273–275.
21. Ohe, Y. & Iwai, K. (1981) *J. Biochem.* **90**, 1205–1211.
22. Zhong, R., Roeder, R. G. & Heintz, N. (1983) *Nucleic Acids Res.* **11**, 7409–7425.
23. Birchmeier, C., Grosschedl, R. & Birnstiel, M. (1982) *Cell* **28**, 739–745.
24. Krieg, P. A. & Melton, D. A. (1984) *Nature (London)* **308**, 203–206.
25. Hagenbuchle, O., Bovey, R. & Young, R. A. (1980) *Cell* **21**, 179–187.
26. Verde, P., Stoppelli, M. P., Galeffi, P., Di Nocera, P. & Blasi, F. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 4727–4731.
27. Harvey, R. P., Whiting, J. A., Coles, L. S., Krieg, P. A. & Wells, J. R. E. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 2819–2823.
28. Ruberti, I., Fragapane, P., Pierandrei-Amaldi, P., Beccari, E., Amaldi, F. & Bozzoni, I. (1982) *Nucleic Acids Res.* **10**, 7543–7559.