

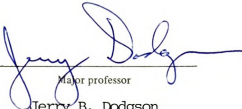




This is to certify that the  
thesis entitled  
CHARACTERIZATION OF A CHICKEN  
H3.3 REPLACEMENT VARIANT  
HISTONE GENE  
presented by  
DAVID CHRISTOPHER BRUSH

has been accepted towards fulfillment  
of the requirements for

M.S. degree in Biochemistry

  
Major professor  
Jerry B. Dodgson

Date 2/18/85





RETURNING MATERIALS:

Place in book drop to  
remove this checkout from  
your record. FINES will  
be charged if book is  
returned after the date  
stamped below.

--	--	--

CHARACTERIZATION OF A CHICKEN H3.3 REPLACEMENT  
VARIANT HISTONE GENE

By

David C. Brush

A THESIS

Submitted to  
Michigan State University  
in partial fulfillment of the requirements  
for the degree of

MASTER OF SCIENCE

Department of Biochemistry

1985

## ABSTRACT

### CHARACTERIZATION OF A CHICKEN H3.3 REPLACEMENT VARIANT HISTONE GENE

By

David C. Brush

The subclone pBH6b-2.6 was restriction mapped and subsequently sequenced by the method of Maxam and Gilbert. This subclone contains a 2.3 kb fragment from the  $\lambda$ Charon 4A recombinant clone,  $\lambda$ CH6b, which hybridized to a probe consisting of sea urchin H3 histone DNA sequences. Nucleotide sequence data reveals that pBH6b-2.6 contains a H3.3 replacement variant histone gene interrupted by three introns. Two of the intervening sequences occur in the coding portion of the gene while the third (revealed by S1 mapping experiments) occurs in the 5'-nontranslated region of the gene. The H3.3 gene described herein is the second example of a H3.3 replacement variant histone gene characterized in detail. It codes for an identical protein sequence to the first such gene (H3.3-1), thus it has been designated H3.3-2. Although the two genes code for the same histone variant, comparison of their non-coding base pairs and intron sizes suggests that the two genes are evolutionarily, only distantly related. In



addition, it was shown that H3.3-2 messenger RNA is post-transcriptionally polyadenylated whereas mRNA of the replication variant, H3.2, is not. The expression of the H3.3-2 gene in different tissues is also described. It is demonstrated that the H3.3-2 gene is expressed at low (basal) levels in dividing and nondividing tissues. The structure and expression of H3.3-2, a replacement variant histone, is compared to that of H3.2, the corresponding replication variant histone.

To My Wife, LeAnne

#### ACKNOWLEDGEMENTS

I would like to thank Dr. Jerry B. Dodgson for his guidance and support throughout this research project.

I would also like to thank Theresa Fillwock for typing of this manuscript.





# TABLE OF CONTENTS

	<u>Page</u>
List of Figures. . . . .	iii
I. Introduction. . . . .	1
II. Materials and Methods . . . . .	22
A. Methods . . . . .	22
1. Subcloning of DNA . . . . .	22
2. Transformation of HB101 with Subclone pBH6b-2.6 . . . . .	22
3. Large Scale Isolation and Purification of Plasmid DNA . . . . .	23
4. Restriction Endonuclease Mapping of pBH6b-2.6 . . . . .	25
5. Purification of Gel Fractionated DNA Fragments. . . . .	28
6. Labeling of Double-Stranded DNA . . . . .	29
7. Maxam and Gilbert DNA Sequencing of Plasmid pBH6b-2.6 . . . . .	31
8. Preparation and Isolation of Chicken RNA. . . . .	34
9. S1 Nuclease Mapping . . . . .	35
10. Primer Extension Analysis . . . . .	36
B. Materials . . . . .	37
III. Results . . . . .	38
IV. Discussion. . . . .	75
V. References. . . . .	86



# LIST OF FIGURES

<u>Figure</u>		<u>Page</u>
1.	Arrangement of the tandem repeat units of sea urchin histone genes and <u>Drosophila</u> histone genes. . . . .	6
2.	Organization of the $\lambda$ Charon 4A recombinant clone $\lambda$ CH6b. . . . .	41
3.	Restriction endonuclease map, Maxam and Gilbert sequencing strategy and relative position of H3.3 histone gene for subclone pBH6b-2.6 . . . . .	43
4.	Complete nucleotide sequence of the 2.3 kb insert of subclone pBH6b-2.6. . . . .	46
5.	Comparison of the protein and nucleotide sequences of histone H3.3-2 to histones H3.3-1 and H3.2. . . .	50
6.	Identification of the leader exon of the H3.3-2 gene . . . . .	56
7.	Comparison of the major chicken H3 histone (H3.2) versus variant chicken H3 histone (H3.3-2) consensus sequences thought to be essential for the proper initiation of transcription . . . . .	62
8.	Bar graph representation of the variation in intron size among the histones known to contain intervening sequences. . . . .	64
9.	Comparison of the intron/exon junctions within the histone genes known to contain intervening sequences. . . . .	66
10.	Expression of variant H3.3-2 and total H3.2 histone genes in different RNA samples . . . . .	70



## INTRODUCTION

Our fundamental knowledge of eukaryotic gene expression has been enhanced over the years by the ever-increasing body of information relating to the histone genes. Histones comprise a group of highly conserved, small, basic proteins which are present in all eukaryotes. In addition, histones complex with DNA to form the basic subunit of chromatin (1). This subunit, the nucleosome, can complex with other nuclear proteins to form even higher orders of chromatin structure. Two sources of variation are known to exist within the histones; these include sequence differences and post-translational modifications (acetylation, phosphorylation, methylation) (2). These sources of variation are known to affect the ways in which histones interact with DNA, as well as with each other. Although the role of these variants is not firmly established, they may indirectly influence gene expression. Nucleosomes with different variant composition could display differences in DNA-binding, thereby accounting for alterations in chromatin structure. Chromatin structure has been linked to transcriptional activity through the use of DNase I (3). Regions upstream of actively transcribed genes are more susceptible to nuclease attack than are similar sequences upstream of inactive genes. Histones are interesting therefore, since they may exert a wide spread influence on gene expression through their fundamental association with chromatin.





A second consideration involves the mode of histone gene expression itself. Histone gene expression is known to be both cell-cycle and developmentally regulated. Most histone protein synthesis appears to be confined to S phase, with some evidence that transcription is closely coordinated to DNA replication (4). Since control at the transcriptional level is the most frequently utilized mechanism for control of eukaryotic gene expression (5), this "linkage" of histones to DNA replication might be explained by a transcriptional regulatory mechanism. Hereford et al. have shown that in yeast things are not that simple, and that two levels of control are implicated (6). The first is an activation of histone transcription as the cell cycle passes through late G1. The second involves stabilization of histone mRNA by the process of DNA replication itself. Such a control mechanism could be used to regulate expression of any gene whose product was required in late G1 or S phase. A somewhat separate but related area of interest is the developmental regulation of histone genes. Various groups of distinct histone genes are turned on during development; a good example are the early and late histone genes of sea urchin (7). Both sets of genes are under tight transcriptional control relative to the requirement for certain proteins at defined stages of development. Many other genes in the cell also must find a way to meet similar requirements. At this point, it seems reasonable that both the cell cycle and the developmental regulation of histone gene expression are related to similar phenomena that exist for a variety of other eukaryotic gene families.

Four distinct sets of histones are recognized (8). These include the histones unique for spermatogenesis and oogenesis, as well as the



replication and replacement histones. The oocyte-specific histones are present during maturation and in embryos, while the spermatocyte-specific histones are found during meiotic prophase of spermatocytes. The replication histones comprise a set of embryonic histones which are found in rapidly dividing somatic tissue. The replacement histones can be seen in non-dividing tissues in increasing amounts as these tissues age.

The histone proteins fall into five classes originally based on the composition of their basic amino acids (9). These include the arginine-rich histones H3 and H4, the slightly lysine-rich histones H2A and H2B, and the very lysine-rich H1 histones. From the evolutionary standpoint, histones H3 and H4 are among the most highly conserved proteins known. Histones H2A and H2B exhibit a greater degree of evolutionary variability, with histone H1 being the most variable of the five classes. Most of the variability is confined to the N-terminal portion of the protein, with this being the region involved in histone:DNA binding. The C-terminal portion is very highly conserved and it participates in the histone:histone interactions crucial to nucleosome formation. As previously stated, the histones are small ranging in size from 11,000 daltons for H4 to 24,000 daltons for H1.

The nucleosome is composed of a core particle consisting of two copies each of histones H2A, H2B, H3 and H4 to form an oligomer (1). Histone H3 binds to histone H4 to form a  $(H3_2H4_2)$  tetramer, which determines the diameter of the core (9). DNA is wrapped around the core particle twice, thereby binding 146 base pairs of DNA. In between the nucleosomes are the linker regions, which consist of 20-80 base



pairs of DNA connecting the nucleosomes. The appearance of the nucleosome at this point has the characteristic of beads on a string. Histone H1 has been shown to bind to the DNA as well as to the other histones. A monomer of histone H1 is thought to seal the DNA in the nucleosome by binding at the point where it enters and leaves. Whether or not H1 is present determines which of two characteristic appearances chromatin takes on. At low ionic strength without H1, a 10 nm fiber is visible under the electron microscope. This is essentially a continuous string of nucleosomes. At greater ionic strength with H1 present, a 30 nm fiber is visible. The 30 nm fiber can be seen to have an underlying coiled structure that contains approximately six nucleosomes per turn. The 30 nm fiber can also be involved in even more complex types of chromatin structuring. Thus, we can see that histones are central to the most basic level of chromatin structure, and they are essential for the highly complex organization of DNA into the eukaryotic chromosome.

Early attempts to characterize the histones revolved around the search for a unifying theme. The essential use of histones in all eukaryotes coupled with their high degree of sequence conservation encouraged the somewhat naive belief that many structural and functional properties would be similar in most organisms. This unified view was further encouraged when the first histone genes were isolated in various sea urchins. Different species of sea urchin which were a hundred million years apart evolutionarily had almost identical topological gene organizations (7). However, as more organisms have been characterized with respect to their histone genes,

much greater diversity has been revealed. Thus, it can no longer be said that there exists a "typical" arrangement of histone genes.

As stated earlier, analysis of histone gene organization was first attempted in sea urchin (7). Consequently, the largest body of information pertaining to histone gene organization and regulation exists regarding the sea urchin. A logical starting point for any discussion of histone genes therefore involves a review of sea urchin histone gene organization.

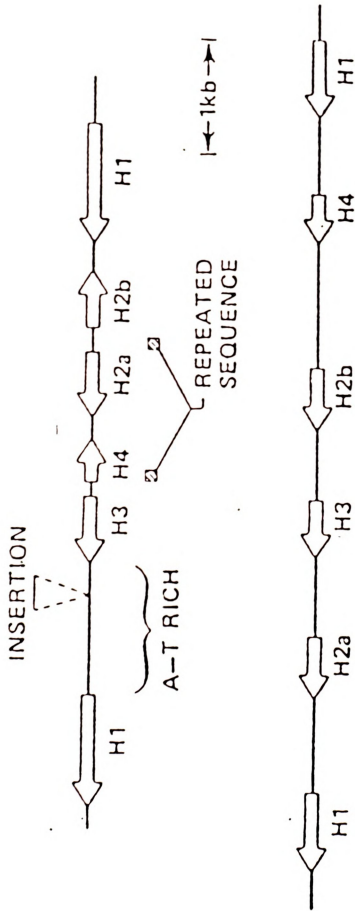
The most common sea urchin histone genes are organized into a series of highly-reiterated tandem repeats. Each of the five histone genes is present in the repeat unit once, with the repetition frequency anywhere from one-hundred to several hundred copies (7). Coding portions are GC-rich and all five genes are transcribed off the same strand. The gene order relative to the transcribed strand is 3' to 5' H1, H4, H2B, H3, H2A (10). Although all the genes are transcribed off the same strand, there is little or no evidence for polycistronic messages suggesting that each gene has its own separate promoter. Hentschel and Birnstiel (10, Table 2) reviewed sequence data for all of the five major classes of histones from P. miliaris and S. purpuratus. Consensus promoter sequences (TATA box) could be demonstrated upstream of the coding portions for each of the sea urchin histone genes. The fact that these sequences have been highly conserved suggests that they must be functional and that each gene is transcribed individually. Spacer regions are composed of AT-rich nontranscribed DNA interspersed between coding regions. An examination of spacer organization reveals several notable structural characteristics. The primary sequences of various spacer regions





Figure 1. Characteristic arrangement of the tandem repeat units of sea urchin histone genes (bottom) and Drosophila histone genes (top).

# Histone Gene Tandemly Repeated Arrays



from Lifton et al. (1978) CSHSQB 41, 1047.

diverge greatly between sea urchin species but overall size and location are fairly constant. Spacers also contain no detectable highly repetitive sequences, but some clustering of AT-rich sequences is observed. Additionally, spacers have been found to contain homocopolymer stretches such as (CT/GA)<sub>27</sub> which may function in the maintenance of repeat homogeneity. Although some microheterogeneity exists, overall the several hundred-fold duplication of tandem repeats results in a fair degree of identity among the repeats.

At this point, it seems appropriate to relate new ideas to classical thinking and thereby put into perspective the ways in which our understanding of histone gene organization and regulation is changing. Initially, it was a commonly held belief that each repeat in sea urchins represented one of several hundred identical copies (7). Sea urchins require massive amounts of histone mRNA during a relatively short timespan of early development. This type of gene arrangement could very easily account for the large number of transcripts sea urchins require at this time. However, recently increased sensitivity in resolving histone variants has begun to paint a somewhat different picture. The use of nonionic detergents such as Triton X-100 in gels by Zwiedler and his co-workers (11) have shown that several distinct variants of histones exist. Restriction endonuclease mapping and hybridization studies also bear out the existence of variants. Variation occurs in both primary sequence and in peptide length. It now appears that the tandem repeats actually represent distinct gene batteries which contain variants to be expressed at certain times during development. Thus the commonly studied sea urchin tandem repeats would be presumably only that

battery of replication histone genes used specifically during rapid cell division early in embryonic development.

Variants, or isohistones, can be shown for a variety of sea urchin species in both mature sperm and early embryos (12, Von Holt et al.). P. angulosus mature sperm cells contain three H2B isohistones (H2B1, H2B2, H2B3) which are found in widely varying amounts from cell to cell. In early embryos of S. purpuratus and P. angulosus a variety of histone mRNAs are detected. These represent distinct mRNAs required at different stages of development and not simply post-translational modifications. Evidence for this comes from placing mRNAs in a heterologous translation system (in which no modifications take place) which yields proteins with electrophoretic mobilities identical to corresponding histone protein variants isolated from the growing embryo. Additionally, stage-specific histones have been isolated in sufficient amounts to allow a partial structural determination. These stage-specific histone variants demonstrate distinct primary structural differences (see below).

Several isohistones display unique sequence variations which could play a role in determining gene expression. Various H2B isohistones upon sequence comparison reveal highly variable N-terminal regions both in sequence and in length (12, Von Holt et al.). At the crux of this variability is a characteristic pentapeptide repeat unit. An example of the pentapeptide repeat for isohistone H2B1 of P. angulosus sperm cells is Pro-Thr-Lys-Arg-Ser. In a typical H2B isohistone from an embryonic cell, the pentapeptide repeat can be absent or present in a single copy. However, in a haploid sperm cell where transcription and replication are completely repressed, up to

three or four intact pentapeptide repeats are present with an additional one or two mutated repeats. A similar phenomenon is also seen with isohistones of H1. N-terminal variability is even more pronounced in isohistones of H1 than H2B. In this case though, a tetrapeptide (Ser-Pro-Arg-Lys) is absent in embryo H1 isohistones and four distinct repeats are reiterated three or four times in H1 isohistones from mature sperm. It has been suggested that both of these repeats bind to DNA due to a highly basic composition. The greater the number of repeats, the stronger the interaction with the DNA and therefore the tighter the isohistones complex with the DNA. If this binding difference does occur, it is interesting that the strongest interaction would occur in transcriptionally inactive tissue (sperm) and the weakest binding would occur in the most transcriptionally active tissues (in embryos). Nucleosomes comprised of different isohistones which are developmentally regulated suggest structural features which could give rise to the major functional forms of the genome. These include actively transcribed, temporarily repressed but inducible, and permanently repressed regions of DNA.

Our understanding of histone gene regulation in sea urchins is changing almost as rapidly as have our structural conceptions: different distinct sets of histones are expressed at different stages during development (10). Early in development until the fifth or sixth cleavage, the cleavage stage histones are expressed. There is also a set of early histone genes and a set of late histone genes which can be identified during sea urchin development. Until recently, transcriptional control appeared to be the most obvious method of regulating these events. However, recent advances in the powerful technique of in situ hybridization are indicating a different

mechanism may be involved (12, Angerer et al.). Initial experiments using a nick-translated early histone repeat (S. purpuratus) demonstrated a high signal in the pronuclei of embryos. Further experimentation involving a series of much more specific probes authenticated the level of hybridization as being due to early histone mRNA localized in the pronuclei. Calculations by two separate groups (12, Angerer et al., 13; Showman et al.) using different methods placed the fraction of localized early histone mRNA in the pronucleus at 95-100%. This high content persists until the first cleavage whereupon the nuclear membrane breaks down. Showman et al. (13) have demonstrated that other abundant maternal mRNAs (tubulin, actin) do not accumulate in the pronucleus. From these studies a striking observation in the developmental regulation of histone genes emerges (12, Angerer et al.). Cleavage stage histone genes are transcribed and mRNAs are transported to the cytoplasm before maturation of the oocyte. However, early variant mRNA transcription is initiated after maturation and the resulting transcripts are sequestered in the pronuclei until the period of development where they are required. Thus, both transcriptional and transport control are necessary to provide for the appearance of the proper histone proteins at their respective times during development. This serves to reinforce two earlier points: histones vary widely in their approaches to solving regulatory problems, and what we learn from histones could be widely applicable to the control of gene expression in general.

The second species in which histone genes were extensively characterized was Drosophila melanogaster. Comparisons of Drosophila vs. sea urchin and subsequent comparisons with histone genes of other

organisms only serve to reinforce the variation that exists in the regulation of histone genes.

Drosophila histone gene organization was initially examined by Karp and Hogness when they screened plasmids containing Drosophila DNA with probes made from labelled sea urchin histone mRNA (14). The histone genes are present as highly-reiterated clustered tandem repeats with a repetition frequency of about 110 copies per haploid genome (10). Each repeat consists of one copy of the 5 major types of histone genes. Interspersed between the five coding portions of each repeat are spacer regions. Two major types of repeats are found in Drosophila; these include a 4.8 and 5.0 kilobase repeat. The only detectable difference between the two is an insert of 240 base pairs in the spacer region between the coding portions for histones H1 and H3. The larger of the two repeating units is present in excess of the smaller repeat by a ratio of 3:1.

In many ways the Drosophila histone genes appear to be very similar overall to the major sea urchin histone genes. However, several important differences between the two are known to exist. For instance, the order of the histone genes and the mechanism of transcription are both distinct. The Drosophila histone gene order is H3, H4, H2A, H2B and H1. Histones H3, H2A and H1 are transcribed off one strand while histones H4 and H2B are transcribed off the opposite strand (10). The divergent transcription utilized by Drosophila is similar to yeast, but different from that seen for sea urchin. Divergent transcription would also require that at least two sites for the initiation of transcription be present. Initially it was postulated that all five genes in sea urchin could be transcribed from a single promoter and would give rise to repeat-length RNA. Although



this now appears not to be the case, Drosophila represented the first system where the requirement for multiple sites of initiation could be demonstrated in histone gene expression.

We next wish to consider the multiple levels of gene regulation that exist during Drosophila development. It is important to note that, unlike sea urchins, no stage- or tissue-specific variant histone mRNAs have been shown in Drosophila (12, Anderson et al.). Therefore, elucidation of regulatory pathways do not necessarily have to take into account the expression of a variety of histones required at different stages of development. The three major levels of control of Drosophila histone gene expression include translational efficiency, rates of transcription and rates of mRNA turnover (12, Anderson et al.). All of these regulatory mechanisms combine to produce the appropriate level of histone protein to complex with DNA at various developmental stages. Since very little histone protein is stored in the mature egg, it is the histone mRNA in the embryo which is crucial. Each of the three regulatory mechanisms make a contribution to the control of Drosophila histone protein synthesis. However, each contributes to a different degree and one must look at all three to sort out the overall picture. In examining these mechanisms, it is helpful to compare synthesis and turnover to a cellular standard. In this case total cytoplasmic poly (A)+ mRNA serves as a useful reference point. Take for instance, translational efficiency. The fraction of total poly (A)+ mRNA associated with polysomes during early embryogenesis increases slightly from 55% at one hour to 70% at four hours. This contrasts with recruitment of histone mRNA into polysomes which goes from 25% immediately after oviposition to 90%

four hours later. Transcription also shows increased activity that has been quantitated. In the first six hours of embryogenesis, the rate of synthesis per nucleus of total poly (A)+ mRNA and histone mRNA is roughly parallel. However, between 6 and 13 hours the rate of synthesis per nucleus of total poly (A)+ mRNA remains constant while the rate of histone mRNA synthesis per nucleus drops 20-fold. This decrease almost exactly parallels the rate of DNA replication. Finally, the rate of mRNA turnover of total poly (A)+ mRNA remains constant throughout embryogenesis while histone mRNA stability drops at least 15-fold. Upon careful consideration of each of these mechanisms of regulation, we gain a better understanding of the multiple levels of gene regulation that exist in Drosophila. Translational control represents a relative fine tuning of the system since the fraction of histone messages associated with the polysomes changes only about three-fold. By comparison, histone mRNA stability changes 15-fold and the rate of histone mRNA synthesis per nucleus decreases 60-fold in the first thirteen hours of embryogenesis relative to total poly (A)+ mRNA. Histone mRNA turnover and synthesis therefore must account for the major levels of control of histone gene expression in Drosophila.

Having given a general account of histone genes in sea urchin and Drosophila, it is now possible to examine the histone genes of chicken. The initial characterization of chicken histone genes and their regulation of expression were among the first attempts to expand our knowledge of histone gene organization to vertebrates. As stated earlier, it was initially thought that the high copy number of histone genes in the sea urchin would allow for the large quantity of histone protein required in embryogenesis. DNA replication in vertebrate

development is not nearly as intense as in sea urchin; therefore the demand for extremely rapid histone protein synthesis is also absent.

Histone gene organization in the chicken has indeed turned out to be quite different from Drosophila and sea urchin. Crawford et al. (15) initially observed that each of the chicken histone genes was represented approximately ten times, and that the genes were present in a tandemly duplicated array. The latter observation, however, is now clearly in error. The chicken histone genes are often (but not always) present in clusters but no tandem repeats have been observed. Engel and Dodgson (16) established this by direct isolation and characterization of a variety of genomic chicken histone clones. Harvey et al. (17) also established at about the same time that the chicken histone genes were non-tandemly arranged. They did this by isolating two genomic clones with chicken histone cDNA which were then mapped with several gene specific probes. Additionally, Harvey et al. showed that two chicken H3 histone genes present in one clone were divergently transcribed.

Additional clues to the gross overall organization of the chicken histone genes has come from two sources. Sugarman et al. (18) characterized in detail 15 lambda Charon 4A recombinant bacteriophage containing chicken histone genes. Initially 50 lambda recombinants had been isolated (16) due to hybridization to sea urchin H2A and H3 histone genes, and 15 unique recombinants were selected for further experimentation. Sugarman et al. extensively mapped all 15 lambda recombinants with probes constructed of each of the five chicken histone genes H1, H2A, H2B, H3, H4. J.R.E. Wells and colleagues took a somewhat different approach to look at the overall topology. Wells and colleagues (12, Engel) were able to walk down the chromosome and

look at the organization of 36 kilobases of contiguous chromosomal DNA containing chicken histone genes. Both of these studies reveal a clustering of the chicken histone genes, but once again there are no tandem repeats.

The existence of histone variants in chicken is well documented. A good example of tissue-specific variation is histone H5 which replaces histone H1 in the condensed nuclei of adult avian erythrocytes (12, Harvey and Wells). The chicken histone H5 gene has recently been isolated by two groups working independently as well as in our own laboratory. Krieg et al. (19) utilized the known protein sequence available for H5 to choose a region in the gene from which to construct a unique 11-base deoxynucleotide. This sequence was then used to prime cDNA synthesis and thereby generate a probe. The extended primer was used to screen a cDNA library made from reticulocyte RNA. The final step in isolating the gene involved using the H5 cDNA to isolate the H5 gene from a lambda Charon 4A recombinant library. Ruiz-Vasquez and Ruiz-Carillo (20) chose a different route. They used a specific antibody to identify unique cDNA clones containing H5 sequences. As a result of these efforts, the chicken H5 histone gene has been sequenced and characterized. Several interesting observations regarding histone H5 can be made from this data. Briefly, H5 is present as a single-copy gene, it codes for the expected protein sequence (it is not a pseudogene) and it is not linked to any of the other histone genes (13, Harvey, Wells). Probably the most surprising finding was that only one copy of H5 is present per haploid genome, even though it replaces histone H1 which is present approximately 10 times per haploid genome. Furthermore,

the histone H5 gene was found to contain no intervening sequences and, as was previously demonstrated, the H5 mRNA is polyadenylated (despite the lack of a AAUAAA sequence). So far, histone H5 is the only proven example of a tissue-specific histone in chicken.

In the course of characterizing the chicken histone gene family, other variant chicken histones have come to our attention. Engel, Sugarman, and Dodgson (21) have isolated a chicken histone variant during analysis of their initial 50 lambda Charon 4A histone gene clones. While attempting to locate chicken H5 histone genomic sequences, they characterized a clone which contained a gene coding for a known protein variant H3 histone (22). This gene, designated H3.3 shows only four amino acid differences (135 total) when compared to a normal H3 histone (H3.2) gene, however, there is a 19% primary sequence difference. The most unique feature of the H3.3 variant is that the coding portion of the gene is interrupted by two intervening sequences. For a long time histones have been recognized as exceptions to the rule that most eukaryotic genes contain intervening sequences (23). This was the first example of any histone gene in any organism which contained intervening sequences. Since this report, Woudt et al. (24) have reported intervening sequences in the genes coding for histones H3 and H4 of Neurospora crassa. Wells and his collaborators have also isolated a chicken histone variant while screening a cDNA recombinant library with core histone probes to identify clones with large inserts that showed weak hybridization (12, Harvey, Wells). Several H2A and H2B clones were isolated and a cDNA that weakly hybridized to the H2A probe was characterized. Sequencing

data revealed an extremely variant H2A-like protein. Therefore, Wells has labeled this gene H2A.F. H2A.F contains a unique nonapeptide sequence highly conserved in all H2A histones sequenced so far, and yet it shows a 40% divergence from the amino acid sequence of the most abundant H2A histone in chicken erythrocytes. To put this in perspective, calf and chicken H2A histones only differ by 4%. Additionally, H2A.F shows no hybridization to mouse, human or sea urchin DNAs and considerable variation in level of expression is apparent between different tissues.

From a structural examination of these variants, we can see that differences do exist even in the relatively small numbers of the vertebrate histone genes. This leads to the main thrust of this investigation. Do the variants which have been characterized so far represent isolated mutational events or the existence of a yet to be discovered class of histone genes? Engel (12, Engel) concluded that all previous histone studies have relied on embryonic histone genes. Childs et al. (25) have isolated and characterized late-stage histone H3 and H4 genes from the sea urchin Lytechinus pictus. The late-stage histone genes are different from the early-stage histone genes which are tandemly repeated and highly reiterated several hundred times. Late-stage histone genes are present in a greatly reduced number so that isolation of these genes has lagged until the development of a positive cloning procedure that specifically excluded tandemly repeated early histone genes from recombinant DNA libraries. Comparison of early-stage histone H3 to late-stage histone H3 reveals identical proteins but a 19% primary sequence difference. The question has arisen whether or not all of the histone genes isolated

so far simply represent genes coding for embryonic histone proteins. If so, then where are the other histones which must replace the embryonic histones in adult tissue. A clue may come from the report of Wu and Bonner (26) who noted that some types of variant histone biosynthesis are not exclusively restricted to S phase. Several variants were shown to be synthesized at a "basal" level throughout the cell cycle in Chinese hamster ovary cells. Included among the variants was the histone H3.3.

One of the most pressing questions at this time regarding histone gene expression is whether or not separate classes of histones exist which have been overlooked in the past. Due to the sequence differences of the two variants isolated so far it is possible that classical methods of isolating histone genes with heterologous sea urchin probes may have selected for only one class. The majority of genes isolated so far would fall under the heading of replication histones since they are present in rapidly dividing tissue. The class yet to be explored (variants like H3.3, H2A.F) would be that present in non-dividing tissue at low levels throughout the cell cycle as demonstrated by Wu and Bonner. This group probably constitutes the replacement histones briefly mentioned earlier. If these two classes exist, it should be possible to demonstrate the existence of other variant chicken histone genes. If their existence can not be demonstrated, variants such as H3.3 might be considered artifacts or peculiarities rather than representatives of a separate distinct histone class. Taking into account the possible role which has been postulated for variation in nucleosome structure with regards to gene



expression, these variant classes of histone genes may play a fundamental role in regulation of overall gene expression.

The chicken H3 histone genes appeared to be an attractive system in which to further this investigation. One of the few variants characterized so far was an H3 histone gene. As it turned out this variant also is the only known vertebrate histone gene to contain intervening sequences. Whether or not this will turn out to be the characteristic of replacement histones is unknown. Only the continuing structural elucidation of variants will give us the answers.

Since the characterization of the chicken histone genes was first initiated, several other vertebrates have been examined for the organization and regulation of their histone genes.

In Xenopus, each of the core histones are reiterated 45-50-fold with the exact number of H1 histone genes yet to be determined (12, Van Dongen et al.). The Xenopus histone genes are clustered and approximately 30 clusters show nearly identical restriction maps while 20 others are unique. The 30 identical clusters show a tandem arrangement. More than one gene order has been found for the Xenopus histone genes, each one associated with a different variant H1 histone gene (10). Expression of the Xenopus histone genes has been shown to be differential in development and transcription takes place off both strands. Regulation of histone gene expression in Xenopus is different from any of the systems examined so far. Large amounts of maternal histone protein and histone mRNA are stored in the oocyte for use in early embryogenesis. During this time, histone protein synthesis is not coordinated with DNA synthesis. Use of the stored



histone protein and mRNAs ceases in the early gastrula stage, normal zygotic transcription takes over and transcription becomes coordinated with DNA replication. This mechanism allows Xenopus to meet the requirement for large amounts of histone protein during early embryogenesis, despite possessing a moderate number of histone genes.

The five major types of histone genes appear to be present in mouse at a gene copy number of 10-20 copies (10). No simple repeating structure of the genes coding for particular histones is obvious (12, Marzluff and Graves). Several variants of histone H1 have been characterized and extensive primary sequence variation is evident. Analysis of a clone containing mouse histone DNA sequences has revealed that transcription occurs off opposite strands of the DNA. Regulation of mouse histone protein levels is accomplished by regulating mRNA transcription and the rate of mRNA turnover.

The human histone genes are moderately reiterated with the gene copy number approximately 40 (12, Stein et al.). The human histone genes appear to be clustered, but no simple tandem repeat is apparent. Anywhere from 4-7 characteristic arrangements of the histone genes can be observed by restriction mapping. Interspersed between the histone genes are several members of the Alu family of highly repetitive DNA sequences. Variants can be detected for each of the five major types of histone genes. An abundance of evidence exists which suggests that human histone gene expression is temporally coupled to DNA replication in the cell.

## MATERIALS AND METHODS

### A. METHODS

Subcloning of DNA. The  $\lambda$ CH6b phage was originally isolated by Engel and Dodgson (16) and further characterized in detail by Sugarman et al. (18). The DNA fragment of  $\lambda$ CH6b which strongly hybridized to a histone H3 probe (see Results) was flanked by a Hind III and Bam HI site. This fragment was isolated by electroelution (see below) and subcloned using standard techniques (27) into the plasmid vector pBR322.

Transformation of HB101 with Subclone pBH6b-2.6. Plasmid pBR322 containing the insert subcloned from  $\lambda$ CH6b phage (see above) was used to transform the E. coli strain HB101 by the RbCl<sub>2</sub> transformation technique (D. Hanahan, 28).

A tube containing 200  $\mu$ l of HB101 at 3.0-A<sub>600</sub> units/ml in 40 mM KOAc, 15% sucrose, 60 mM CaCl<sub>2</sub>, 45 mM MnCl<sub>2</sub> and 100 mM RbCl<sub>2</sub>, pH 5.9 (stored as stock at -70° until needed) was quick-thawed and put on ice for 30 minutes. Next, 7  $\mu$ l of dimethylsulfoxide was added, the tube slightly agitated and put on ice for 10 minutes. The competent HB101 bacteria were added directly to the suspended plasmid (ligated DNA in 10  $\mu$ l TE) and the mixture was placed on ice for 10 minutes. At this point, the bacteria were quick frozen by immersion in a CO<sub>2</sub>/ethanol bath. The bacteria were allowed to remain in the

CO<sub>2</sub>/ethanol bath for 2 minutes, whereupon they were quick thawed. When the thaw was complete, the tube was immediately placed on ice for 30 minutes. The bacteria were then allowed to stand for 2 minutes in a 37° water bath, after which 0.2-0.8 ml of LB broth medium was added (no antibiotic) and the bacteria incubated at 37° for 30 minutes without shaking. Transformed bacteria were spread on LB agar plates containing the antibiotic ampicillin.

The presence of the appropriate insert in the subclone was confirmed in transformants by growing up plasmid mini-preps. The mini-prep procedure used was the alkaline-lysis protocol outlined in reference 27. A 10 µl aliquot of the mini-prep was used in a Hind III/Bam HI double-digest. The DNA from each digestion was run out on a 0.8% agarose gel, stained with ethidium bromide (0.2 µg/ml) and visualized under an ultraviolet lamp. From this it was determined which of the plasmids from the amp<sup>r</sup> colonies contained the appropriate 2.6 kb Bam HI/Hind III insert and one of these strains was selected for large-scale plasmid DNA isolation.

The subclone consisting of the histone H3 hybridizing region of the λCH6b phage and pBR322 vector sequences was designated as pBH6b-2.6.

Large Scale Isolation and Purification of Plasmid DNA. The protocol used to isolate and purify large amounts of pBH6b-2.6 plasmid DNA is a slightly modified version of the alkaline lysis procedure found in reference 27.

An overnight that consisted of 7 ml LB broth medium and ampicillin (50 µg/ml) was inoculated with the pBH6b-2.6-containing HB101 bacteria (see above) and incubated at 37° in a shaker overnight.

Large scale growth was initiated by infecting 500 ml of M-9 enriched medium with 3 ml of overnight and vigorously shaking at 37° until an OD<sub>600</sub> of 0.4-0.5 was reached. Once the proper density was achieved, 2.5 ml of chloramphenicol (34 mg/ml in ethanol) was added and incubation was continued at 37° with vigorous shaking for 12-16 hours. Bacterial cells were harvested by centrifugation, and the supernatant was discarded. The pelleted bacteria from a 500 ml culture were resuspended in 6 ml of 50 mM glucose, 25 mM Tris·Cl, pH 8.0; 10 mM EDTA; 5 mg/ml lysozyme, transferred to 50 ml SS34 plastic tubes and let stand for 5 minutes at room temperature. Next, 12 ml of 0.2 N NaOH, 1% SDS was added and the contents of the tube were mixed by gently inverting the tube. This mixture was let stand 10 minutes on ice, whereupon 9 ml of ice cold 5 M KOAc (pH 4.8) was added, the contents vigorously mixed, and the tube put back on ice for 10 minutes. Cellular DNA and bacterial debris were pelleted out by centrifugation at 12K for 10 minutes. The supernatant was transferred to a 50 ml SS34 plastic tube, 0.6 volumes of isopropyl alcohol was added, and the plasmid DNA was allowed to precipitate at room temperature for 15 minutes. Plasmid DNA was pelleted by centrifugation at 10-12 K for 15 minutes. Pelleted plasmid DNA was resuspended in 5 ml of 0.2 M NaCl, 0.01 M Tris·Cl, 1 mM EDTA, extracted once with phenol:chloroform (1:1) and precipitated with 5-6 ml of isopropyl alcohol at -20° for 1 hour.

Once the plasmid DNA had been isolated, it had to be purified away from chromosomal DNA sequences. This was accomplished by centrifugation through a two-step cesium chloride-ethidium bromide step gradient (27). Plasmid DNA was spun down at 10K for 15 minutes



at 4°, and the pellet dried. Pelleted plasmid DNA was resuspended in 2.4 ml of 0.01 M Tris·Cl, 1 mM EDTA (TE); 4.3 grams of CsCl was dissolved in the DNA solution and 0.3 ml (10 mg/ml) ethidium bromide was added (in the dark if possible). This solution was mixed thoroughly and immediately underlaid below 5 ml of 1.47 density CsCl in TE in a Ti70 centrifugation tube. Plasmid DNA was banded by centrifugation at 40K overnight using a Ti70 Beckman rotor, and the plasmid DNA was visualized under ultraviolet light. Two bands are apparent with the lower band representing the purified plasmid DNA. This band was removed with a Pasteur pipette, and ethidium bromide was removed from the plasmid DNA by a series of 3-4 extractions with CsCl-saturated isobutanol. The final step in the purification involved extensive dialysis of the plasmid DNA with at least three changes of TE buffer. Concentrations of plasmid DNA were determined using a Beckman spectrophotometer with  $1 A_{260} = 50 \mu\text{g/ml DNA}$ .

Restriction Endonuclease Mapping of pBH6b-2.6. All reactions utilizing restriction endonucleases were carried out under the manufacturer's recommended assay conditions. In the case of double digests where two restriction enzymes had similar assay conditions (ionic strength, temperature) both enzymes were added to the reaction at the same time. In double digests where two restriction enzymes required different assay conditions, one enzyme was added and the reaction was allowed to proceed for several hours. At the end of this time, the conditions were optimized (e.g., salts added, temperature lowered) for the second enzyme and the digestion continued for several more hours. Routinely, approximately 1  $\mu\text{g}$  of plasmid DNA was incubated with 1-2 units of enzyme for 6-8 hours. Restriction



endonuclease reactions were terminated with one-tenth volume of 1 M NaCl, 0.25 M EDTA. DNA was precipitated by the addition of 2.5 volumes of ethanol at -20° overnight (or -70° for 1 hr). The precipitated DNA was recovered by centrifugation in a microfuge for 15 minutes at 4°. Pelleted DNA was drained, dried, and resuspended in 25  $\mu$ l of 100 mM Tris-borate, 2 mM EDTA, 5% Ficoll, 0.05% bromophenol blue and 0.05% xylene cyanol in preparation for polyacrylamide gel electrophoresis.

DNA fragments which had been subjected to restriction endonuclease digestion were electrophoresed vertically on 5% polyacrylamide gels (2 mm, 4 mm thickness) for approximately 300 volt-hours in 100 mM Tris-borate, 2 mM EDTA, pH 8.3. Molecular weight size standards consisting of Hinf I, Hind III or Taq I digested pBR322 were run in lanes next to the restricted DNA fragments. Electrophoresed DNA fragments were visualized by soaking the gel in 0.2  $\mu$ g/ml ethidium bromide in water for approximately 30 minutes at room temperature. Once stained, the DNA fragments were visualized by exposing the gel to an ultraviolet light source. Each gel was subsequently photographed using Polaroid 667 film and a red filter. From the photograph, one could measure the distance traveled by the molecular weight size standards and construct a standard curve for each gel. Once the standard curve had been generated, the actual sizes of the various restricted DNA fragments could be deduced. By examining the sizes of the DNA fragments generated by single- and double-digests, one could construct a physical map of the restriction endonuclease sites within plasmid pBH6b-2.6.



Several ambiguities in the restriction map remained after the above procedure; therefore it was also necessary to use the Smith-Birnstiel procedure (27) to map restriction endonuclease sites. Plasmid DNA was linearized using either Hind III or Bam HI, the protruding 5' ends were labeled with [ $\gamma$ - $^{32}$ P]ATP (see below), and the plasmid was digested a second time with whichever of the two enzymes that had not been used to initially linearize the plasmid. The labeled Hind III/Bam HI insert fragment was then isolated by electroelution (see below) for Smith Birnstiel mapping. Approximately  $10^4$  cpm of end-labeled insert fragment, 1  $\mu$ g salmon sperm carrier DNA, 1  $\mu$ l 10X restriction enzyme buffer, water up to 10  $\mu$ l and 1-2 units of restriction endonuclease was placed in a tube. The reaction was incubated at the appropriate temperature and 1.8  $\mu$ l aliquots were withdrawn at time intervals of 2, 5, 10, 15 and 30 minutes into corresponding aliquots of 1  $\mu$ l of 0.5 M EDTA. All samples were combined, 2  $\mu$ l of gel-loading dye (5% Ficoll, 0.05% bromophenol blue, 0.05% xylene cyanol) was added and the reaction was run out on a 5% polyacrylamide gel with  $10^4$  cpm of end-labeled molecular weight size standards. The gel was dried and exposed to Kodak X-Omat AR film for 8-12 hours without intensifying screens. A ladder of digested DNA fragments was visible on the film. Each fragment's length indicated a given restriction site that distance from the labeled end of the insert.

Using a combination of these two techniques, a reasonably accurate restriction endonuclease map was generated for plasmid pBH6b-2.6 (see Figure 3).



Purification of Gel Fractionated DNA Fragments. DNA fragments digested with restriction endonucleases that were purified from gels fell into two categories. Fragments which were to be labeled for DNA sequencing and fragments which had been labeled, but were cut with a second enzyme to generate a singly end-labeled fragment. The procedure outlined below (Girvitz, 29) was used for both types of fragment isolated from either polyacrylamide or agarose gels.

DNA fragments were run out on a 5% polyacrylamide gel (see above), stained with ethidium bromide and visualized under an ultra-violet light score. The portion of the polyacrylamide gel containing the desired DNA fragment was excised as a block with a scalpel. The block of polyacrylamide gel was then placed in an empty minigel pouring form, and molten 0.7% agarose (in 40 mM Tris-acetate, 2 mM EDTA, pH 7.5) was poured around the fragment. Once the agarose had hardened, a slice was made lengthwise with a scalpel at the boundary between the polyacrylamide gel section and the hardened agarose. A piece of Whatman 3 MM paper backed by dialysis membrane was cut to the approximate size of the polyacrylamide section and inserted into the incision. The 3 MM paper was between the dialysis membrane and the DNA fragment of interest. Current was then applied for 15-30 minutes so that the DNA fragment migrated out of the polyacrylamide section and into the 3 MM paper. DNA is recovered from the 3 MM paper by placing both the 3 MM paper and dialysis membrane in separate Eppendorf tubes which had been punctured through the bottom with a hot 25 ga. syringe needle. These tubes were placed in 12 x 75 mm plastic test tubes and both the 3 MM paper and dialysis membrane were washed with 200  $\mu$ l of 0.2 M NaCl, 50 mM Tris-Cl, pH 7.5, 1 mM EDTA. Wash was

collected in the bottom of the tubes by centrifugation at medium speed for 20-40 seconds in a table top centrifuge. Each sample received 3-4 washes, with the membrane wash used to wash the paper in the subsequent wash, and the contents of the collecting tubes were precipitated by the addition of two volumes of isopropyl alcohol overnight at  $-20^{\circ}$ . This procedure also works very well for agarose gels, however it is not necessary to excise the desired DNA fragment as a block with a scalpel. An incision is simply made in front of the DNA fragment in the horizontal preparative agarose gel and the 3 MM paper and dialysis membrane inserted as above. Electroelution is then performed as described.

Labeling of Double-Stranded DNA. DNA fragments which had been generated by restriction endonuclease digestion and purified by electroelution were labeled according to the ends of the DNA which had been generated by enzymatic cleavage. A slightly modified procedure was used for DNA fragments with blunt ends versus fragments with 5'-protruding ends.

Approximately 5-10  $\mu\text{g}$  of DNA with 5'-protruding ends in 50  $\mu\text{l}$  of distilled water was combined with 50  $\mu\text{l}$  of 100 mM Tris-Cl, pH 9.0 and 1  $\mu\text{l}$  of calf alkaline phosphatase (ca. 100 U/ml). This reaction was incubated for 1 hour at  $37^{\circ}$  followed by the addition of 10  $\mu\text{l}$  of 1 M NaCl, 0.25 M EDTA. The DNA was extracted with one volume of phenol:chloroform (1:1) twice, extracted with one volume of ether twice, and the contents precipitated by the addition of 55  $\mu\text{l}$  7.5 M  $\text{NH}_4\text{OAc}$  and 500  $\mu\text{l}$  ethanol for 15 minutes in a  $\text{CO}_2$ /ethanol bath. The DNA was pelleted by centrifugation in a microfuge for 15 minutes at  $4^{\circ}$ , drained, washed with 1 ml ice cold ethanol, respun for 5

minutes, drained and dried. Pelleted DNA was resuspended in distilled water, 5  $\mu$ l 10X protruding kinase salts (0.5 M Tris·Cl, pH 7.6, 0.1 M MgCl<sub>2</sub>, 50 mM DTT, 1 mM spermidine, 1 mM EDTA), 0.2-0.5 mCi of [ $\gamma$ -<sup>32</sup>P]ATP ( 3000 C/mole; ICN) and 3-5 units of T4 polynucleotide kinase (50  $\mu$ l total volume). The kinase reaction was incubated at 37° for 60 minutes, whereupon the reaction was terminated with 50  $\mu$ l of 0.3 M NaOAc. Labeled DNA was precipitated with 250  $\mu$ l ethanol, incubated in a CO<sub>2</sub>/ethanol bath for 5 minutes, spun in a microfuge at 4° for 10 minutes, in some cases reprecipitated, drained and dried. Labeled plasmid DNA was then taken up in water and used either for Maxam and Gilbert DNA sequencing, Smith-Birnsteil mapping or S1 nuclease mapping.

DNA fragments with blunt ends required a slight modified of the above protocol. Approximately 10-20  $\mu$ g of plasmid DNA with blunt ends in 50  $\mu$ l of 50 mM Tris·Cl, pH 9.0 was incubated for 2 hours at 60° with 2  $\mu$ l calf alkaline phosphatase. Two additions of the phosphatase was made, 1  $\mu$ l was added at the start of the reaction and a second 1  $\mu$ l was added after 1 hour of incubation. After the two hours had expired, 200  $\mu$ l of 0.3 M NaOAc was added, the phosphatased DNA extracted with one volume of phenol:chloroform (1:1) twice, extracted with one volume of ether twice and precipitated with 2.5 volumes of ethanol for 15 minutes in a CO<sub>2</sub>/ethanol bath. DNA was spun in a microfuge for 15 minutes at 4°, drained, washed twice with 500  $\mu$ l ethanol and dried. The pellet was taken up in 11.5  $\mu$ l distilled water, 2.5  $\mu$ l 50 mM EGTA, 4.0  $\mu$ l 50 mM spermidine and incubated for 3 minutes at 90° in an oil heating block. After 3 minutes, the DNA was immediately placed on ice for 1 minute, and then 1.0  $\mu$ l 5 mg/ml BSA,



2.5  $\mu$ l 10X kinase buffer (500 mM Glycine-NaOH, pH 9.5, 100 mM  $MgCl_2$ , 50 mM DTT, 1 mM EDTA, 30% glycerol), 0.2-0.5 mCi [ $\gamma$ - $^{32}P$ ]ATP (3000 C/mmol; ICN) and 5-8 U of T4 polynucleotide kinase were added (25  $\mu$ l, total volume). The kinase reaction was incubated for 4 hours at 37° whereupon the reaction was terminated and precipitated as above.

An additional method was used to label double-stranded DNA with 5'-protruding ends, so that the DNA could be read 3'  $\rightarrow$  5' in Maxam and Gilbert sequencing. This procedure allowed one to read the complementary strand to a kinased 5'-protruding end-labeled fragment. Basically, 5-10 pmoles of restricted DNA with 5'-protruding ends was combined with 30  $\mu$ l. [ $\gamma$ - $^{32}P$ ]dNTP (NEN, 800 C/mmol, 13  $\mu$ M), 5  $\mu$ l 10X cDNA salts (0.5 M Tris-Cl, pH 8.3, 0.6 M NaCl, 0.06 M  $MgCl_2$ , RNase-free), 2.5  $\mu$ l 0.4 M DTT, 2  $\mu$ l AMV reverse transcriptase (14 U/ $\mu$ l) and 10.5  $\mu$ l distilled water. The reaction was mixed and incubated at 37° for 1 hour, at which time the reaction was processed identically to the kinase reaction.

Maxam and Gilbert DNA Sequencing of Plasmid pBH6b-2.6: DNA sequencing by chemical modification was performed as outlined by Maxam and Gilbert (30) using at some points the modifications of Smith and Calvo (31).

Maxam and Gilbert DNA sequencing requires a singly end-labeled substrate; therefore labeled DNA was cut with restriction endonuclease, run out on a gel, electroeluted and precipitated. DNA to be sequenced was resuspended in 50  $\mu$ l distilled water and 5  $\mu$ l aliquots were placed in each of four 1.5 ml silanized Eppendorf tubes. Tubes were designated C, CT, AG and G to reflect the nucleotide which would be susceptible to chemical modification. Each tube received





1  $\mu$ l of salmon sperm DNA (1  $\mu$ g/ $\mu$ l) to act as carrier and the contents were mixed. Tube C received 15  $\mu$ l 5 M NaCl, 30  $\mu$ l hydrazine, was incubated for 10 minutes on ice and the reaction terminated with 200  $\mu$ l of hydrazine stop buffer (0.3 M NaOAc, 0.1 mM EDTA, 25  $\mu$ g/ml tRNA). The CT reaction was identical to the C reaction except that 15  $\mu$ l of distilled water was substituted for the 15  $\mu$ l of 5 M NaCl. Tube AG received 15  $\mu$ l distilled water, 50  $\mu$ l of formic acid, was incubated at 23° for 2 minutes and the reaction stopped with 180  $\mu$ l of hydrazine stop buffer. Reaction G consisted of 200  $\mu$ l of DMS buffer (50 mM sodium cacodylate, pH 8.0, 1 mM EDTA), 0.5  $\mu$ l dimethyl sulfate and was incubated for 10 minutes on ice and terminated with 50  $\mu$ l DMS stop buffer (1.5 M NaOAc, pH 7.0, 1.0 M mercaptoethanol, 100  $\mu$ g/ml tRNA). From this point, all reactions were treated the same. Modified DNA was precipitated by the addition of 600  $\mu$ l ethanol and 5 minutes incubation in a CO<sub>2</sub>/ethanol bath. The DNA was pelleted by centrifugation for 5 minutes in a microfuge at 4°, drained and taken up in 200  $\mu$ l 0.15 M NaOAc followed with 500  $\mu$ l ethanol. DNA was precipitated a second time as above, spun 5 minutes, drained, washed with 500  $\mu$ l ice cold ethanol, briefly respun, drained and dried in a dessicator. Each pellet was resuspended in 50  $\mu$ l 1 M piperidine (fresh) and incubated for 30 minutes at 90° in an oil heating block. Tubes were briefly placed on ice, given a quick spin to sediment condensation on the sides of the tubes and the contents of each tube were placed in a fresh 1.5 ml Eppendorf tube. DNA was precipitated with 50  $\mu$ l 0.3 M NaOAc and 250  $\mu$ l ethanol as before, spun 5 minutes and drained. Each pellet was washed with 250  $\mu$ l 70% ethanol, spun 5 minutes, drained, dried in a dessicator and taken up in 3  $\mu$ l of FA



loading buffer (90% formamide, 0.25 mM EDTA, 0.02% bromophenol blue, 0.02% xylene cyanol).

Several other modifications of the Maxam and Gilbert DNA sequencing procedure were used. At one point, chemical modification of the C and CT reaction was proceeding too far (excess degradation). Both reactions were modified so that the incubation time was changed to 4 minutes at 23°, however the reactions were terminated and the DNA precipitated as before. When it came time to add 200  $\mu$ l 0.15 M NaOAc, an additional 20  $\mu$ l of acetylacetone was added to completely neutralize the reactivity of any lingering hydrazine (32). This mixture was let stand 5 minutes at room temperature, 500  $\mu$ l ethanol added and the reaction processed as before.

Ambiguity between C and CT reactions was resolved by the incorporation of a fifth chemical modification reaction to the procedure. This reaction involved photoinduced cleavage of DNA to determine thymidine residues and was designated T>G (33). Briefly, 5  $\mu$ l of singly end-labeled fragment was mixed with 5  $\mu$ l of 2 M cyclohexylamine and exposed to an ultraviolet light source for 1 1/2 minutes (time can vary due to intensity and distance of light source). An addition of 100  $\mu$ l 0.3 M NaOAc, pH 5.2, 1  $\mu$ l salmon sperm DNA (1 mg/ml) as carrier and 500  $\mu$ l ethanol was added to the DNA. DNA was precipitated as before, spun 10 minutes, drained, washed with ice cold ethanol and dried in a dessicator. Pelleted DNA was resuspended in 50  $\mu$ l 1 M piperidine (fresh) and processed as above.

Prior to loading the gel, DNA samples were heated at 90° for 1 1/2 minutes in an oil heating block, cooled on ice for 1 minute and spun briefly to concentrate the sample. Three aliquots of 1  $\mu$ l of

each sample were loaded onto either a 6%, 8%, or 12% 84 x 18 cm, 0.4 mm thick polyacrylamide sequencing gel (for an 8% gel, 8% acrylamide, 0.4% Bis, 7 M urea, 90 mM TBE, filtered and degassed) in the order G, AG, TC, C (T>G optional). After the initial 1  $\mu$ l aliquot of each sample was loaded, the gel was electrophoresed 12-14 hours at 30-40 constant watts until the xylene cyanol had just run into the lower buffer well. At this time, a second set of each of the samples was loaded onto the gel and electrophoresed approximately 8 hours or until the xylene cyanol had migrated four-fifths of the way down the gel. A third set of samples was then loaded on the gel, and electrophoresed approximately four hours or until the bromophenol blue had migrated two-thirds of the length of the gel. Buffer used for the sequencing set-up was 90 mM TBE. Gels were removed from between the glass plates, overlayed onto Whatman 3 MM paper and dried for 20 minutes at 80° on a BioRad Model 1125B gel drier. Dried gels were exposed to Kodak X-Omat AR5 film with or without intensifying screens depending on the relative amounts of radioactivity.

Preparation and Isolation of Chicken RNA. Chickens were made anemic by injections of phenylhydrazine (2.5% w/v) over the course of 6 consecutive days. Anemic red blood cells were prepared and fractionated into cytoplasm and nuclei according to Longacre and Rutter (34). Cytoplasmic RNA was prepared by extensive phenol/chloroform extraction of red cell cytoplasm and ethanol precipitation. Poly (A)<sup>+</sup> red cell RNA was also prepared according to Longacre and Rutter (34). All other RNAs were prepared by the method of Chirgwin et al. (35).



S1 Nuclease Mapping. A singly end-labeled fragment was normally used for S1 mapping, this was generated as above and taken up in 100  $\mu$ l 0.3 M NaOAc, pH 7.0. The fragment was made RNase free by phenol:chloroform and ether extraction. Typically 50  $\mu$ l of singly end-labeled fragment was placed in a 1.5 ml Eppendorf tube along with, for example, 40  $\mu$ l total cytoplasmic red cell RNA (3 mg/ml), 10  $\mu$ l RNase free 5 M NaCl and 300  $\mu$ l ethanol. A control was prepared with 50  $\mu$ l singly end-labeled fragment, 10  $\mu$ l RNase free 5 M NaCl, 40  $\mu$ l RNase free distilled water and 300  $\mu$ l ethanol in a 1.5 ml Eppendorf tube and run under identical conditions to the RNA-containing reaction. In some cases 100  $\mu$ g of yeast tRNA was added to the control. Singly end-labeled DNA and cytoplasmic RNA were precipitated overnight at  $-20^{\circ}$ , spun down for 15 minutes at  $4^{\circ}$  in a microfuge and drained. The pellet was washed once with 500  $\mu$ l ice cold ethanol, spun 5 minutes in a microfuge at  $4^{\circ}$ , drained and dried. The pellet was resuspended in 20  $\mu$ l FAHB (80% formamide, 0.4 M NaCl, 0.04 M PIPES, pH 6.4, 1 mM EDTA, RNase free) incubated for 2 minutes at  $90^{\circ}$ , then at  $55^{\circ}$  for 2 hours followed by incubation at  $50^{\circ}$  for a final 3 hours. The S1 nuclease reaction was quenched with 300  $\mu$ l 1X S1 salts (25% glycerol, 0.15 M NaOAc, pH 4.5, 5 mM  $\text{ZnSO}_4$ , 250 mM NaCl, 50  $\mu$ g/ml denatured salmon sperm DNA) and the reaction volume divided equally into three tubes. Tube 1 typically received 50 units of S1 nuclease, tube 2 received 150 units of S1 nuclease and tube 3 received 400 units. These reactions were incubated at  $37^{\circ}$  for 15 minutes after which 10  $\mu$ l of 1 M NaCl, 0.25 M EDTA was added to each reaction. Each reaction was extracted once with one volume of phenol:chloroform (1:1), precipitated by the addition of 2.5 volumes of ethanol and

incubation for 15 minutes in a CO<sub>2</sub>/ethanol bath and spun 15 minutes in a microfuge at 4°. Pelleted DNA was drained, washed with 250 µl ice cold ethanol, spun 5 minutes, drained and dried. Each reaction was resuspended in 3 µl FA loading buffer (see above). As stated earlier, the no RNA control was run under exactly the same reaction conditions. Exact RNA and S1 levels are described for specific experiments in the figure legends.

S1 nuclease reactions were loaded on a 36 x 18 cm, 0.4 mm thick 8% sequencing gel and run for 2-3 hours (bromophenol blue migrates two-thirds length of gel) at 25 constant watts. Gels were removed from the glass plates, overlaid onto Whatman 3 MM paper and dried at 80° for 20 minutes on a BioRad gel drier. Dried gels were exposed to Kodak X-Omat AR5 film with intensifying screens for 2 days to 1 week depending on the intensity of the radioactive signal.

Primer Extension Analysis. The labeled DNA fragment used for primer extension was prepared and hybridized to RNA as described above for S1 mapping. The RNA:DNA hybrid preparation was precipitated (above) and taken up in a 100 µl reaction identical to the one described for 3' end labeling DNA fragments except all dNTPs were unlabeled and at 100 µM and AMV reverse transcriptase was used at 560 U/ml. The reaction was incubated at 42° for 2 hr, phenol:CHCl<sub>3</sub>-extracted, and 5 µl of 5 N NaOH was added to the aqueous phase followed by incubation at 90° for 5 min. The reaction was neutralized with 5 N HCl; 100 µl of 0.3 M NaOAc were added followed by 600 µl ethanol and the DNA was precipitated and prepared for gel electrophoresis as described for the S1 analyses.

## B. MATERIALS

Restriction endonucleases were purchased from New England Biolabs, Bethesda Research Labs, Amersham, Biotec and Collaborative Research, Inc. AMV reverse transcriptase was obtained from Life Sciences, Inc. through the Office of Program Resources and Logistics, Viral Cancer Program, National Institutes of Health. Ribonuclease A, lysozyme and ampicillin were purchased from Sigma. T4 polynucleotide kinase was purchased from Amersham and Worthington. T4 DNA ligase was purchased from Worthington. T4 DNA ligase was purchased from Worthington. Calf alkaline phosphatase was purchased from Boehringer Mannheim and further purified by column chromatography by D. Grandy (our lab). [ $\gamma$ - $^{32}\text{P}$ ]Deoxynucleotide triphosphates were purchased from both Amersham and New England Nuclear. Hydrazine, DMS and x-ray film were from Eastman Kodak. S1 nuclease was purchased from PL Biochemicals. Dimethylsulfoxide was purchased from Aldrich and formic acid, cyclohexylamine and piperidine were purchased from Fisher Scientific. Glass plates, comb and sequencing stands were all purchased from Bethesda Research Laboratories.



## RESULTS

Outline of Protocol. As stated earlier, the first example of a histone gene which contained intervening sequences was isolated and characterized by Engel, Sugarman and Dodgson (21). This gene was isolated from a set of  $\lambda$  Charon 4A recombinant clones known to contain chicken histone DNA sequences. Initially, Engel and Dodgson (16), had screened a chicken DNA-containing  $\lambda$  Charon 4A recombinant library with sea urchin histone H2A and H3 heterologous probes. Sugarman et al. (18) subsequently extensively mapped and characterized 15 of these  $\lambda$  recombinants and they went on to demonstrate the existence of H1, H2A, H2B, H3 and H4 chicken histone DNA sequences within the clones. These workers also attempted to see if any of the  $\lambda$  chicken histone clones contained chicken histone H5 DNA sequences. Since it was not known at the time if histone H5 was linked to other histone genes, the 50 original  $\lambda$  recombinants were screened to look for sequences typical of H5 sequences. At that time, histone H5 was the only vertebrate histone known to have its message polyadenylated (8) and it was also known to be expressed specifically in red blood cells. Thus, a cDNA probe complementary to red cell poly (A)+ mRNA was hybridized to the set of  $\lambda$  histone clones. It has since been shown that H5 is not linked to any of the other histone genes. However, several of the 50 histone clones did in fact hybridize to the cDNA made to red cell poly (A)+ mRNA. The first of these clones to be characterized ( $\lambda$ CH4b)

turned out not to contain the H5 histone gene, but instead it contained a histone H3 gene (H3.3-1) which surprisingly contained two intervening sequences which split the coding regions of the gene. This gene has been shown to code for a variant H3 histone protein designated H3.3. Although the H3.3-1 gene hybridized to mRNA in the poly(A)+ fraction, the more common histone H3 gene (H3.2) was shown not to produce polyadenylated mRNA. Originally Engel, Sugarman and Dodgson proposed that the H3.3-1 clone may have hybridized to the oligo-dT primed cDNA to poly(A)+ mRNA due to a d(A<sub>11</sub>GAg) sequence in the mRNA antisense strand 255 base pairs 3' to the termination codon (21). If transcribed, this sequence could account for the selection of the gene's RNA on oligo-dT cellulose. However, it now appears that the message encoded by this variant gene may indeed be post-transcriptionally polyadenylated (J.D. Engel, personal communication). This H3.3-1 gene was the first histone gene from any organism shown to contain intervening sequences.

Since the initial characterization of the H3.3-1 variant histone gene (1982), one other report has appeared in the literature pertaining to the isolation of histone genes with intervening sequences. Woudt et al. (24) have demonstrated the existence of one intron in a histone H3 gene and two introns in a histone H4 gene from Neurospora crassa. It also appears that a variant chicken H2A histone gene (Showman et al.; J.R.E. Wells, personal communication) and a human H3.3 histone gene (L. Kedes, personal communication) contain introns.

This report describes efforts to characterize a second clone isolated from the chicken histone  $\lambda$  recombinants ( $\lambda$ CH6b) (16) which

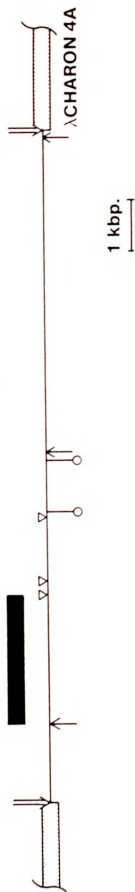
hybridized to both the heterologous sea urchin histone H3 probe and to the probe made from cDNA prepared against adult red cell poly(A)+ mRNA (21). To determine if this clone also contained a histone H3 gene possessing intervening sequences, it was necessary to determine its DNA sequence. This would tell us if the presence of introns was a common aspect of several chicken histone genes, and thus if there may be a whole family of such genes which may also show similarities in their pattern of expression (e.g. replacement histones).

The chicken histone H3-hybridizing region which was subcloned and characterized in this report was contained in the  $\lambda$  Charon 4A recombinant clone designated  $\lambda$  CH6b (Figure 2). This  $\lambda$  recombinant was initially restriction mapped by Sugarman *et al.* (18), and the 2.3 kilobase pair (kb). Bam HI/Hind III fragment was shown to hybridize to both the heterologous H3 sea urchin (and *Drosophila*) probe and to cDNA prepared against adult red cell poly(A)+ mRNA (J. Dodgson, unpublished results). The 2.3 kb H3-hybridizing fragment was excised with a Bam HI/Hind III double-digest, isolated and subcloned into the plasmid vector pBR322. The plasmid containing the H3-hybridizing region from  $\lambda$  CH6b ligated to Bam HI/Hind III cut pBR322 was designated plasmid pBH6b-2.6.

A fine structure restriction endonuclease map was generated for the 2.3 kb insert fragment of pBH6b-2.6 (Figure 3). This map was important for two reasons. Of major importance, was the fact that knowledge of restriction endonuclease sites determined my sequencing strategy. This strategy is illustrated in Figure 3 above the restriction map, each arrow indicating the site end-labelled and the direction and approximate distance sequenced. Of secondary importance



Figure 2. Organization of the  $\lambda$  Charon 4A recombinant clone  $\lambda$ CH6b. Shaded regions indicate  $\lambda$  Charon 4A sequences. Chicken DNA sequences are represented by a straight line. The 2.3 kb H3-hybridizing region is designated by a solid block above the chicken DNA. Restriction endonuclease sites within the recombinant are represented by symbols described below.



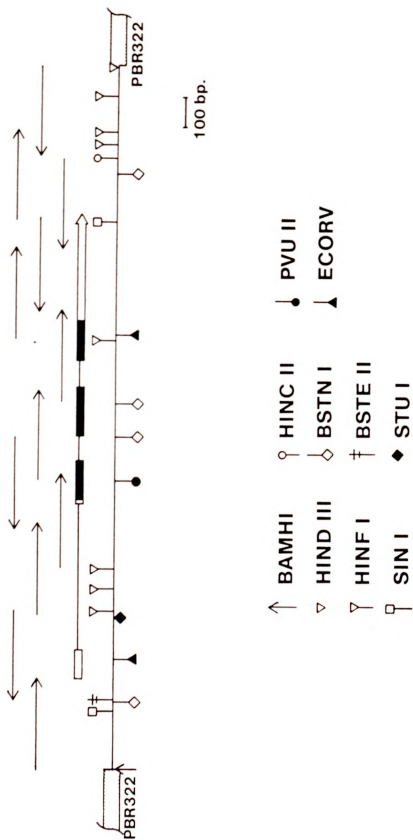
$\parallel$  EcoRI LINKER  
 $\uparrow$  BAM HI  
 $\nabla$  HIND III  
 $\downarrow$  SAC II







Figure 3. Restriction endonuclease map, Maxam and Gilbert sequencing strategy and relative position of H3.3 histone gene for subclone pBH6b-2.6. Shaded areas represent vector pBR322 sequences, while the subcloned chicken DNA sequences from  $\lambda$ CH6b are signified by a line in between. Restriction endonuclease sites are indicated with symbols below. Arrows above the restriction map illustrate the site labeled for Maxam and Gilbert DNA sequencing, the direction and the distance sequenced. The open and filled large arrow between the restriction map and sequencing strategy represents the relative position of the H3.3-2 gene within pBH6b-2.6. Dark areas represent coding portions, with light areas signifying nontranslated portions contained within H3.3-2 mRNA. The thin lines between these blocks give the locations of the three introns. (Note: The 3' end of the message is inferred from the DNA sequence, see text. Arrow points in the 5' to 3' direction).



was the localization of unique restriction endonuclease sites which were commonly placed in the coding portions of H3.3-1 and the 2.3 kb insert of pBH6b-2.6. For example, the Pvu II site present in the first exon of H3.3-1 is the only Pvu II site present in the entire gene. There is also only one such site in the 2.3 kb insert (see Figure 3). Therefore, this was chosen as a good region to look for sequence homology to the coding portion of the H3.3-1 gene.

The total DNA sequence of the 2.3 kb insert of pBH6b-2.6 has been sequenced by Maxam and Gilbert DNA sequencing (Figure 4). Initial DNA sequence comparison around the Pvu II site of the 2.3 kb insert fragment to the DNA sequence around the Pvu II site in the first exon of H3.3-1 revealed extensive homology. Twenty-seven nucleotides around the Pvu II site between the two were identical. Additionally, a unique EcoRV site is present in the third exon of H3.3-1. Therefore, I searched for additional homology in the DNA sequence around the two EcoRV sites present in the 2.3 kb insert of pBH6b-2.6 (see Figure 3). The existence of such homology would orient the gene and suggest an area upon which to concentrate further DNA sequencing efforts. DNA sequence homology with the third exon of H3.3-1 (24 of 27 nts) was demonstrated around the right-hand EcoRV site (see Figure 3, position 1415). This last finding turned out to be rather surprising. Initially, it was thought that if this clone represented a variant histone gene similar to H3.3-1, then the intron sizes would be similar. We based this assumption on the finding that the intervening sequences in, for example, globin genes diverge heavily in sequence between different species, but their position and approximate size is highly conserved. The fact that the distance between these





Figure 4. Complete nucleotide sequence of the 2.3 kb insert of subclone pBH6b-2.6. Upper case letters designate sequences present within the spliced transcript. Numbering of the sequence starts at the site where transcription is initiated, the cap site (+1). Sequences upstream are designated with negative numbers. All sites thought to be relevant to proper transcription are underlined. These include the "TATA" box (-31), the "CCAAT" box (-93) and the proposed polyadenylation site (+1493). The splice donor/acceptor sites occur at 102/568, 713/791, 947/1033. Also underlined are the start (AUG) and stop (UAA) codons necessary for translation.

-280 -260 -240 -220  
 g gatacaaaacc gtgccccgtt tegtccaaatc agcgcgcaga ggcgcgcgt cgtctattgg ccccttctcg tctcaaaata tccaaatcagc  
 -200  
 cgcgcgacttt aacaacccaat cagcggcggg gtaggccccca atcagcgcga argaacggga gacaggccct tgtgtcttat taactcaagg cgcgcgcggg  
 -180 -160 -140 -120  
 -100  
 ggtccccggca atcgaagcggc gcgctgcttc caggcaattgt kaggtcaacg cgcggtgtgc cgcgcgcgcgt ataaaaagtg cgtccccgcg gcgcgcgcggc  
 -80 -60 -40 -20  
 cap 20 40 60 80 100  
 GTAGTGTGGG GTAGTGTGGG TTGGGAGTTG GTTGTGTGTG GAGCAAGGCA GGCTTGGGTG GCGCGCGATA TCGCGGCTGG TCGTCTCCTT TTCTCTGGGA  
 120 140 160 180 200  
 ggtaaagtggg gctcaaccggg cgcgtttctgc ggtgggctgc ggtgctgcgc ggccttccga gactttgatt attcttttt tccccccgag ctgggcggag  
 220 240 260 280 300  
 agttcaggcc ttgcctacgt gccctgagtc acggcccggtt cccatccccc cccgcctcct ccgaggtagc ccgggtggc ccgggagggg gcgtgactca  
 320 340 360 380 400  
 gtccgccttc cccggggagg ggagcggtcg ggccacaacg cggagggggc ggtttctgccc ccgtgactcc tctctcctct cccggggggg gccgctggcc  
 420 440 460 480 500  
 gcacgccccg ctcaacggcg ggggggagggg ggccggccccc gacatctgcc gttctgggg ctggagggga ttgaaaggcg ggggagcgcc gcatcgcggg  
 520 540 560 580 600  
 gccccggcggg tcgcgcgcgc gcgaaactcg aaagaagcgg gccctgcctc acgggtgctg gtctctgtag GTAAGTGAGA AAAAATGGCC CGTACAAAGC  
 620 640 660 680 700  
 AGACCGCCCG CAAGTCCACC GGGGGGAAGG CTCCGGCCAA GCAGCTGGCC ACCAAGGCGG CCGGAAAG CGCTCCCTCT ACCGGCGGCG TCAAGAAGCC  
 720 740 760 780 800  
 TCACCGCTAC AGgtaaggct gccggcgtgc tgtgctgcgg cccggctgc gctgctgcac cgtctccctcc tgttccaccc agCGCGGGGA  
 820 840 860 880 900  
 CCGTCGGGCT CCGTGAGATC CGTCGTACC AGAAGTCCAC GGAGCTGCTG ATCCGCAAGC TGCCCTTCCA CCGGCTGGTC AGGAAATCG CCCAGGATT





920	940	960	980	1000
CAAAACGCAC	TTGAGCTTC	CATCGCTCG	CTGCAGctg	tgctgtgtg cgtgtgtt tgctgaagg t*ctaaaggc
1020	1040	1060	1080	1100
ggttctaacg	tttcttcttg	tcctcttggc	acagcagcgc	agcgaacct atctgtctgg tctgtttcaa gacacaaacc tctggccat ccattgccaac
1120	1140	1160	1180	1200
AGACTCACCA	TCATGCCCAA	AGATATCCAG	TTGGCTGCCA	GGATACGGCG AGACAGACT TAAGTGAAGC CTGTTTTAT GGTGTTTTGT ACTAAATTCT
1220	1240	1260	1280	1300
GTAAATACT	TTGTGTTTAA	TTTGACTT	TTTTCTAAGA	AATTCTTTAT AATATCTTGC ATTTCATTC
1320	1340	1360	1380	1400
GCTAAAAAGT	ACTGTTGACA	TAAACCTCAG	TGATGTGAGC	CTTGTGCTC AGGAGTGACA AGTTGCTAAT ATGCAGAAGG GATGGGTGAT CTTTCTTGCT
1420	1440	1460	1480	1500
TCTCATGCAT	GTTTCTGTAT	GTTAATGACT	TGTTGGGTAG	CTAAAACTTGT AAGTCTACTAG AATTGATATA AATGTGTACA GGTGCTCTTT GCATATAAAC
1520	1540	1560	1580	1600
TGTTTATGAC	TTGATccaa	tggttaacaa	tggggctgt	tagtctgacc atacatcact ggatecaat gteggacttt tcagagggtg aaactacaac
1620	1640	1660	1680	1700
tcttaaccac	agtgtaaact	acagtttct	aaaaagctaa accctggcgc	tatagaatac actatgtgca ttataatag ctattttata tattgtagyg
1720	1740	1760	1780	1800
tcaacaytt	taaattaaat	gttttacc	cacatgagag	gagctctttg cutttggtt cctatggctt ggagaagctg attctcgctt ccagtaigt
1820	1840	1860	1880	1900
agcggtagat	gcttgaata	ctggcggctg	ctggggcagc	ttagctttct gccactgaag cactgctca ccttgctcc ccttgatgt ttttaggagac
1920	1940	1960	1980	
ctgagtcagc	tcttgcctag	ccaggaggag	cagctctgaca	ggcagaggcg aggaagcgc ttcttccct ctctgtggtg ctctctgtt taagctt

two conserved sites in H3.3-1 including intervening sequences was almost 4 kb compared to the 480 base pairs between these two sites in pBH6b-2.6 suggested two possible explanations. Either introns were not present in the H3 gene of pBH6b-2.6, or introns were present but they must be smaller in size (or number). Without further analysis of the sequence, it appeared that the latter must be the case since the distance between the Pvu II site and the EcoRV site in an uninterrupted gene would be 312 base pairs versus the observed 480 base pairs. Further DNA sequence analysis revealed homology to the second exon of H3.3-1 at a point equidistant from the Pvu II site and right-hand EcoRV site in pBH6b-2.6. By inferring the amino acid sequence from the DNA sequence in these regions of homology, three coding portions or exons became apparent.

The amino acid sequence, predicted from the complete sequence of the pBH6b-2.6 insert is identical to the amino acid sequence of the histone H3.3-1 variant (Figure 5). Therefore the histone H3-hybridizing region contained in subclone pBH6b-2.6 represents a second H3.3 variant histone gene. This gene has been designated H3.3-2. If we compare the amino acid sequence of the H3.3 proteins to the amino acid sequence of the major chicken H3 polypeptide (called H3.2) we see four amino acid differences (Figure 5). These include changes from Ser to Ala (31), Ala to Ser (87), Ile to Val (89) and Gly to Met (90). Despite the identical nature of the predicted amino acid sequence for H3.3-1 and H3.3-2, we can see that these two H3.3 genes are, however, quite different in nucleotide sequence (Figure 5). The primary sequence of H3.3-2 varies from H3.2 by 19%. The primary sequence of H3.3-2 varies from H3.3-1 by 18%. From this we can see



Figure 5. Comparison of the protein and nucleotide sequences of histone H3.3-2 to histones H3.3-1 and H3.2. Line A is the nucleotide sequence of histone H3.3-2. Above it is the amino acid sequence of the H3.3 variant polypeptide coded for by H3.3-2. Lines B and C represent differences in DNA sequence between H3.3-2 and H3.3-1 (B) or H3.2 (C). Four amino acid differences are apparent between H3.3-2 and H3.2. These are underlined within the H3.3-2 amino acid sequence and the appropriate amino acid substitution is given below line C. Intron sequences have been removed from the H3.3 DNA sequences for purposes of this comparison.



that H3.3-2 is almost as divergent from H3.3-1 as from H3.2 even though the two H3.3 genes code for identical protein sequences. The implications of this finding will be considered further in the Discussion. The location of the three exons of H3.3-2 relative to the PvuII and EcoRV sites can be seen in Figure 3.

The total DNA sequence analysis of H3.3-2 clearly shows that the gene contains two small intervening sequences that interrupt the coding portions of the gene (see Figure 4). These two intervening sequences occur at exactly the same sites within the coding sequence of H3.3-2 as do the introns in H3.3-1 (5' and 3' relative to transcription). However, these intervening sequences are 80 base pairs and 88 base pairs, respectively, versus 766 and about 2,800 base pairs for the introns of H3.3-1 (see Figure 8). As I stated earlier, we might have anticipated that the introns locations and sizes would be conserved. Only one of these assumptions turned out to be true.

Several other important pieces of information could be deduced from the DNA sequence of pBH6b-2.6 (see Figure 4). Examination of the intron-exon junctions in H3.3-2 reveal the proper consensus donor and acceptor sites (36). No termination codons are present within the protein coding portions of H3.3-2. Additionally, the proper start codon (ATG) and stop codon (UAA) are present flanking the H3.3-2 gene. In fact, it appears that this gene contains the necessary information required to code for a full-length, functional H3.3 variant histone polypeptide. This suggested, but did not prove, that the H3.3-2 gene was actually expressed. However, analysis of the DNA sequence some 200 bp. upstream of the ATG start site for the H3.3-2 protein did not reveal any consensus sequences for the initiation of transcription by

RNA polymerase II (Figure 4). Neither a TATA sequence nor a CCAAT box was identifiable, and this suggested one of three possibilities must exist for H3.3-2. It was conceivable that this gene might not be transcribed despite the encouraging data previously mentioned. It was also possible that this gene did not contain "standard" consensus promoter sequences typical of most other RNA polymerase II-transcribed genes. Since consensus promoter sequences were also not seen directly upstream of the first coding exon of H3.3-1, maybe novel sequences involved in the initiation of transcription existed for both genes. A third alternative was that H3.3-2 might contain another intervening sequence (or several) in the 5'-nontranslated portion of the gene. The existence of intervening sequences in the 5'-nontranslated portion of a gene has been shown for both ovalbumin and insulin genes (23). To determine whether a novel putative promoter sequence existed or if another intervening sequence in the 5'-nontranslated portion had gone undetected, S1 nuclease mapping was undertaken.

S1 mapping utilizes a specific single-stranded singly end-labelled DNA probe to determine where a message begins or where an intron-exon junction occurs. The requirements for the probe in this case were that the DNA be singly end-labeled inside the first exon and that the probe extend 5' to the ATG protein start site. Initially, the singly end-labelled DNA probe is denatured and hybridized at the proper temperature to promote the formation of RNA:DNA hybrids versus reannealing of the two DNA strands of the probe (see Materials and Methods). Once the hybrids are formed, they are treated with S1

nuclease which will attack single-stranded DNA but not RNA:DNA hybrids. The RNA:DNA hybrid formed between the first coding exon and the message will protect labelled DNA of a length equal to the distance between the label and the beginning of the exon. The shortened DNA fragment is denatured and run out on a polyacrylamide gel with size markers. The size of the fragment tells where transcription is initiated or where the nearest intron:exon junction is, since the label and exon length fragment are protected but the intron does not hybridize with the mRNA and is degraded (intron sequences are not present in mature cellular mRNA). This method would therefore shed some light on which of the possible explanations outlined above was correct: no H3.3-2 transcription occurred, the sequences surrounding the transcription initiation site were novel, or an undetected intron existed in the 5'-nontranslated region of H3.3-2.

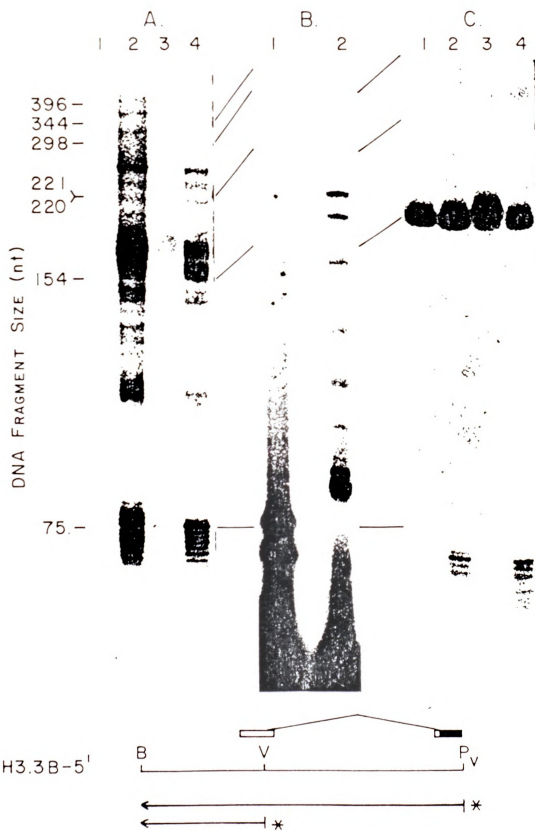
The results of such an S1 mapping experiment are shown in Figure 6(A). The source of RNA was total cytoplasmic RNA from chicken anemic reticulocytes. Since this gene was isolated by hybridization to a probe made of cDNA prepared against adult anemic red cell poly(A)+ mRNA, it was known that the H3.3-2 gene was represented in the above RNA. The probe used in this case consisted of a 932 base pair PvuII/BamHI fragment isolated from pBH6b-2.6 (see Figure 3). This fragment was singly end-labelled at the PvuII site (644), placing the label 60 base pairs from the A of the initiator ATG. The probe contained 875 base pairs upstream this ATG sequence. As we can see in Figure 6A (lanes 2 and 4), a strong signal is indicated at approximately 72-75 base pairs. This would mean that only about 15



and hybridized either in the absence of added RNA (Lanes 1,3) or to 120  $\mu$ g of total anemic red cell cytoplasmic RNA. Digestion reactions were carried out as in A, with S1 nuclease levels at 500 U/ML (Lanes 1,2) and 1500 U/ML (Lanes 3,4). All reactions contained in this figure were run out on 8%, 0.4 mm sequencing gels (Maxam and Gilbert, 30; Materials and Methods), dried and exposed to one intensifying screen for 3-7 days. Arrows indicate the position of labeled marker DNA fragments (Hinf I digest of plasmid pBR322) run in separate lanes. The figure below the autoradiogram represents the location of the intervening sequence in the 5' nontranslated leader sequence and the location of the two singly end-labeled fragments used in the above S1 nuclease reactions.

Figure 6. Identification of the leader exon of the H3.3-2 gene.

(A.) S1 nuclease analysis of the splice acceptor site 5' to the ATG initiation codon. A DNA fragment singly end-labeled at the PvuII site (+644, Figure 4) and extending 931 nucleotides to the Bam HI site of pBH6b-2.6 (see Figure 4) was prepared as outlined in Materials and Methods. Approximately 0.5  $\mu$ g of the singly end-labeled DNA fragment was hybridized either in the absence of added RNA (Lanes 1 and 3) or to 400  $\mu$ g of total anemic red cell cytoplasmic RNA (Lanes 2 and 4). After hybridization, reactions were quenched into S1 reaction buffer, split into three aliquots, digested and processed as outlined in the Materials and Methods. S1 nuclease levels were 1500 U/ML (Lanes 1,2) and 4000 U/ML (Lanes 3,4). (Equivalent results obtained with 500 U/ML, not shown). (B) Primer extension analysis of H3.3-2. Approximately 1  $\mu$ g of the singly end-labeled DNA fragment described in A was digested with the restriction endonuclease Hae III and the resulting 56 bp singly end-labeled PvuII/Hae III fragment isolated. This fragment was hybridized to either 500  $\mu$ g of yeast tRNA (Lane 1) or 300  $\mu$ g of total red cell cytoplasmic RNA (Lane 2). Hybridization and primer extension with AMV reverse transcriptase are described in Materials and Methods. (C) Localization of the 5' end of the H3.3-2 mRNA by S1 nuclease analysis. A DNA fragment singly end-labeled at the EcoRV site (+69, Figure 4) and extending 357 nucleotides to the Bam HI site of pBH6b-2.6 (see Figure 4) was prepared



bases upstream of codon one were protected. Examination of the DNA sequence data in this area (Figure 4) reveals no consensus promoter sequences. However, there is a consensus intron splice acceptor site (3' end of an intervening sequence, Figure 9) in this region around +570 in Figure 4. This data therefore suggested that at least one further intron existed in the 5'-nontranslated region of H3.3-2. Note also that this S1 protection experiment shows that the H3.3-2 gene (or one of identical sequence) is specifically expressed in red cell RNA. The H3.3-1 and the H3.2 genes, for example, diverge completely from H3.3-2 upstream of the ATG initiation codon. Thus, at best, transcripts from these genes could protect only 60 bases of the probe (PvuII site to ATG distance). In fact weak bands at about 60 bases are seen in the gel, possibly due to cross-hybridization to transcripts from these other genes. However the bands at 72-75 bases are almost certainly due specifically to transcription of the H3.3-2 gene. In fact since probe DNA is in considerable sequence excess to the H3.3-2 transcript in the RNA, the intensity of the bands at 72-75 are a specific measure of the level of H3.3-2 transcription in a given RNA sample (as discussed later). This was confirmed by using varying levels of RNA in the S1 experiment (results not shown).

Primer extension was next used to attempt to verify the existence of the intron in the 5'-nontranslated portion of H3.3-2 and to estimate the amount of exon sequence 5' to the first coding exon. \*In this experiment, a singly end-labeled primer is hybridized to the message and the primer is extended to the end of the mRNA using AMV reverse transcriptase. In this case, the probe used in the previous S1 mapping experiment (Pvu II\*/Bam HI) was cut with Hae III and a

56 bp. primer fragment contained entirely in the first exon of H3.3-2 was generated (see Figure 4). This was hybridized to the total red cell cytoplasmic RNA and the primer extended as indicated (Materials and Methods). The labeled DNA was denatured and run out on a polyacrylamide gel (Figure 6B). Because primer extension can be inefficient and terminate prematurely, several bands are seen above the unextended primer band in Figure 6B. However, the largest significant band probably results from complete extension of the primer to the end (5') of the mRNA. In this case, that band is about 180 bases in length. This distance should equal the length of the primer (56 bases) plus the distance between the primer and the splice acceptor site (another 18 bases) plus the length of all further 5' exons (one or more). This latter distance is therefore approximately 106 bases ( $180 - 56 - 18$ ). From this, it became clear that, as expected from the analysis of the sequence data, the site identified by the initial S1 mapping was due to an intron-exon junction and not a transcriptional start site. If the latter were the case, the primer should not have been extended past this point (74 bases). However, in this case the primer was extended approximately 100 additional bases, so there exists another 100 base pair exon 5' to the coding portion of the gene (or more than one exon whose total length is about 100 base pairs).

Since an intron in the 5'-nontranslated region of H3.3-2 was indicated, the next step was to find out exactly where transcription was initiated. For this, a second probe was constructed. By examining the DNA sequence upstream of the consensus acceptor site, we were able to locate a region which resembled a consensus intron donor site



(the 5' end of an intron, Figure 9). This region is apparent at position 101 in Figure 4. (Furthermore, about 130 base pairs upstream of this donor are sequences which resemble consensus promoter sequences, see below). In constructing a DNA probe for the second round of S1 mapping, it was important that the label would be upstream of this region. The label must be in a sequence that is present in the mRNA if it is to be protected from S1 nuclease digestion. The probe also had to be long enough to extend past the promoter or past another splice site if additional introns were implicated. The fragment chosen in this case was a 357 base pair EcoRV\*/BamHI fragment isolated from pBH6b-2.6. The EcoRV site was end-labeled with [ $\gamma$ - $^{32}$ P]ATP (see Materials and Methods), the fragment hybridized to RNA and subjected to S1 digestion (Figure 6C) as before. A strong signal was apparent at around 69 bases indicating the location of the putative transcriptional start site. A single exon extending from this site (cap site, +1 in Figure 4) to the splice donor sequence discussed above (+102 in Figure 4) would account for the length of mRNA sequence upstream which was copied in the primer extension experiment. As mentioned above, consensus promoter sequences are visible in the appropriate positions upstream of the predicted transcription initiation or cap site (see Figure 4). In sum, these data are consistent with the exon organization of the 5' end of the H3.3-2 gene shown at the bottom of Figure 6.

Results of experiments described later suggested that the H3.3-2 mRNA like that of H5 and possibly H3.3-1 is polyadenylated. The appropriate signal sequence for polyadenylation AATAAA (in DNA) was seen at position 1492 (see Figure 4). From this, we predicted the





actual poly(A) addition site by comparison to the 3' ends of other polyadenylated messages as shown in Figure 4. Due to the low level of mature mRNA made from the H3.3-2 gene, it has not been possible to date to definitively map the actual polyadenylation site. Further attempts at this mapping are in progress.

At this point it is interesting to compare the H3.3-2 gene in detail to other analogous genes. Figure 7 is a comparison of the promoter region sequence between the variant H3.3-2 gene and the major chicken erythrocyte H3 histone, H3.2. Two major observations arise from examination of this Figure. First is that the spacing between these sequences seems to be fairly well conserved. Whether these distances are important with respect to orientation of the promoter during transcription initiation is unknown, but it appears these distances are conserved in a wide variety of genes transcribed by RNA polymerase II (although not all). The second thing involves the conservation of the "CCAAT" box, "TATA" box and "cap" box (or RNA initiation site) relative to the consensus sequence. The "TATA" box appears to be the most highly conserved of the promoter sequences, with the "CCAAT" box next, followed by the "cap" box. In addition the "cap" box of the H3.3-2 gene looks very different from those of most genes since it is not nearly as pyrimidine-rich. The H3.3 variants may contain weak promoters corresponding to their low level of expression, and this unusual "cap" box may be related to the H3.3-2 promoter strength.

The next features of the H3.3-2 gene to be compared are the introns. Figure 8 shows a bar graph representation of the size and location of the introns contained in H3.3-2 compared to those in



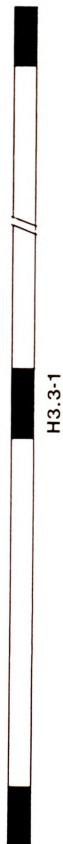
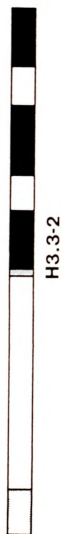
Figure 7. Comparison of major chicken H3 histone (H3.2) versus variant chicken H3 histone (H3.3-2) consensus sequences thought to be essential for the proper initiation of transcription.

	"CCAAT" Box		"TATA" Box		"CAP" Box		
H3.3-2	CGGCAATCG	-42-	GTGCC	-7-	CTATAAAAAG	-20-	GCGTAGT
H3.2	TTTCAATCA	-28-	GATGG	-7-	CTATAAAAGC	-18-	TCTATCT
CONSENSUS	G <sub>H</sub> CCCAAT-A <sub>G</sub>	-(25-50)-	GATCC	-(6-10)-	GTATAAATAG	-(16-24)-	T <sub>H</sub> CATTG <sub>H</sub>





Figure 8. Bar graph representation of the variation in intron size among the histones known to contain intervening sequences. Open regions represent introns, dark regions coding portions and shaded regions 5'-nontranslated sequences. The 3' intron H3.3-1 is actually about 2.8 kb. The H3.3-1 gene probably also contains one or two introns in its 5' nontranslated region which have not yet been definitively mapped (J.D. Engel, personal communication).



100 bp







Figure 9. Comparison of the intron/exon junctions within the histone genes known to contain intervening sequences.

# DONOR SEQUENCE

## ACCEPTOR SEQUENCE

CONSENSUS

H3, 3-2 (1ST INTRON)

CAG/GT<sup>A</sup><sub>G</sub>AGT

(T)<sub>11</sub>N<sup>C</sup>AG/G

H3, 3-2 (2ND INTRON)

GAG/GTAAGT

(T)<sub>8/11</sub>GTAG/G

H3, 3-2 (3RD INTRON)

CAG/GTAAGG

(T)<sub>16/19</sub>CCAAG/G

H3, 3-1 (1ST INTRON)

CAG/GTAGTG

(T)<sub>16/20</sub>ACAG/G

H3, 3-1 (2ND INTRON)

CCG/GTACGC

(T)<sub>22/26</sub>AAAG/G

NL CRASSA

H3 (1ST INTRON)

CTC/GTAAGT

(T)<sub>9/16</sub>CCAAG/G

H4 (1ST INTRON)

GAC/GTAAGT

(T)<sub>13/18</sub>TCAG/G

H4 (2ND INTRON)

CCA/GTACGT

(T)<sub>11/19</sub>CAG/T



H3.3-1 as well as to the intervening sequences of the H3 and H4 histone genes of N. crassa. It becomes immediately apparent from this diagram that the introns in H3.3-2 are quite a bit smaller than those in H3.3-1. This is despite the fact that the intervening sequences are located at identical sites relative to the histone gene coding sequences. Note that it is not yet clear whether or not there is another intron (or more than one) in the 5' untranslated region of the H3.3-1 gene. If we compare H3.3-2 to the H3 and H4 histone genes from N. crassa, we can see that their intron sizes more closely resemble that of the H3.3-2 gene rather than those of the H3.3-1 gene. However, the H3 histone gene from N. crassa only contains one intron in comparison to the two or more contained in the H3.3-1 gene and the three in the H3.3-2 gene. Note also that the location of the intron in the N. crassa H3 histone gene relative to the coding sequences differs from any of the three in H3.3-2. The H4 histone gene of N. crassa contains two introns, but this is the first known example of an H4 histone gene to contain introns, so there is no analogous data to which to compare it. Whether or not intron number, locations and/or sizes in histone genes are important to the functioning of these genes is a topic for further investigation.

While examining the intron sizes, it is also helpful to look at the intron:exon junctions present within these four genes (Figure 9). From this, we can see that the donor and acceptor sites agree very closely between the consensus sequence and those of H3.3-2. Histone variant H3.3-1 shows slightly more variation, with genes of N. crassa demonstrating the most variability. It is presently unclear whether



splicing out of introns in N. crassa resembles more closely the mechanism used in yeast or the somewhat different mechanism in higher eukaryotes. Even though some variability does exist, we can see that on the whole the donor:acceptor sites in histone variant H3.3-2 closely resemble consensus sequences shown to be related to the proper excision of introns.

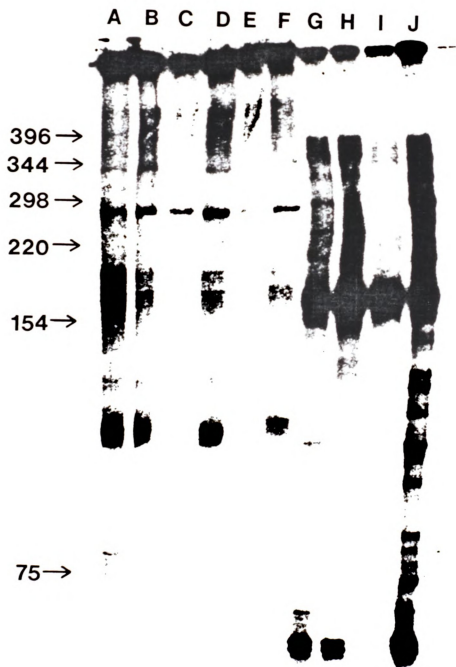
At this point, it appears that all the necessary recognition sequences for eukaryotic transcription, processing and translation are present in the H3.3-2 gene. The next step then, was to look at the expression of the H3.3-2 gene in a variety of tissues. We already knew the H3.3-2 message must be expressed since the S1 mapping was successful. However, according to Wu and Bonner (26), mammalian H3.3 variant histone is expressed at low levels throughout the cell cycle in all tissues. Figure 10 represents the initial attempts to examine the expression of the H3.3-2 histone gene in chickens. Two probes were used; these included the PvuII\*/BamHI fragment from pBH6b-2.6 used in Figure 6A and an analogous PvuII\*/EcoRI fragment of the H3.2 gene from the subclone of H3.2 called pSplink 2A. The PvuII site is in a conserved region of the H3 histone sequence, so it is present at an identical site in both the H3.3-2 and H3.2 genes. The H3.3-2 fragment was hybridized to total anemic red cell cytoplasmic RNA (lane A), anemic red cell poly(A)<sup>-</sup> RNA (lane B), poly(A)<sup>+</sup> anemic red cell RNA (lane C), total adult liver RNA (lane D), total adult breast muscle RNA (lane E) and total chicken embryo fibroblast (CEF) RNA (lane F). In each case 60 µg of RNA was used except for lane C where 0.6 µg was used. By examining the gel in Figure 9, several things become apparent with regard to the expression of this gene. In lane





Figure 10. Expression of variant H3.3-2 and total H3.2 histone genes in different RNA samples. The singly end-labeled PvuII/Bam HI fragment of the H3.3-2 gene was prepared as described in Figure 6. An equivalent DNA fragment was isolated for the H3.2 gene and prepared in a similar fashion. This consisted of A 0.4 kb fragment, singly end-labeled at the analogous PvuII site and extending to the EcoRI site of the H3.2-containing plasmid pSplink 2A. Approximately 0.5  $\mu$ g of the H3.3-2 fragment was hybridized to total anemic red cell cytoplasmic RNA, Lane A; anemic red cell poly(A)<sup>-</sup> RNA, Lane B; anemic red cell poly(A)<sup>+</sup> RNA, Lane C; total adult liver RNA, Lane D; total adult breast muscle RNA, Lane E; and total chicken embryo fibroblast (CEF) RNA, Lane F. In each case 60  $\mu$ g of RNA was used except for the poly(A)<sup>+</sup> anemic red cell RNA where 0.6  $\mu$ g of RNA was hybridized. Approximately 0.3  $\mu$ g of the H3.2 fragment was hybridized to equivalent amounts of total anemic red cell RNA, Lane G; total adult liver RNA, Lane H; total adult breast muscle RNA, Lane I; and total CEF RNA, Lane J. Hybridization and S1 digestion conditions are given in Materials and Methods. Samples were digested with 4000 U/ML of S1 nuclease and analyzed as described in Figure 6.







A, we can see as before in Figure 6A that the probe hybridizes with the total cytoplasmic red cell RNA and gives rise specifically to a ladder of bands at about 75 bases in length. If we look at lanes B and C, the signal in lane B is very weak compared to lane A and less than that in lane C. This is enlightening, since 60  $\mu$ g of RNA was used in lane B and only 0.6  $\mu$ g was used in lane C. In this RNA prep about 1% of the red cell RNA was recovered in the poly(A)+ fraction so these relative RNA levels correspond to about an equal number of cells. The fact that the hybridization signal was equal or stronger in lane C when related to the differences in RNA concentration suggests that over half of the H3.3-2 mRNA is recovered in the 1% of total RNA that bound to oligo(dT)-cellulose (two passages). This strongly suggests that most, if not all, of the H3.3-2 mRNA is normally polyadenylated. Note that this must be post-transcriptional polyadenylation since the H3.3-2 gene does not have the downstream A-rich sequence that may be transcribed as part of the H3.3-1 gene. If we now examine lanes D, E and F we can see relatively equal low level hybridization signals. This bears out the hypothesis that this H3.3 variant, H3.3-2 is expressed at low levels in several tissues, both dividing (CEF) and nondividing (adult breast muscle and liver). The signal is weakest in the breast muscle RNA lane. Since this is a rather transcriptionally inactive tissue, it may be that the level of poly(A)+ RNA is especially low in this sample or that this RNA sample was more degraded on purification than the others. It is interesting that H3.3-2 is clearly expressed in the dividing CEF cells. This suggests that the gene is expressed at a low constitutive level in most, if not all, tissues. Results of similar experiments with the



H3.3-1 gene are in agreement with this interpretation (J.D. Engel, personal communication).

The additional lanes in Figure 10: G, H, I and J contain RNA samples hybridized to the H3.2 probe. These include total anemic red cell RNA (lane G), liver RNA (lane H), breast muscle RNA (lane I) and CEF RNA (lane J). These results are included for comparison of expression between H3.2, which represents the major chicken H3 histone polypeptide and H3.3-2 which represents a histone variant. In these lanes two major S1 digestion products are of interest. The darker bands at about 60 bases are due to protection from the PvuII site to the ATG initiation codon. These result from the expression of all other H3.2 genes which are closely homologous to the H3.2 gene used as probe in this experiment. (It is not known how many such genes exist in the chicken genome; probably about 5-10). RNA specifically made from the H3.2 gene used as probe will be homologous to all the probe from the PvuII site to the cap site about 42 base pairs upstream of the ATG, thus giving rise to a series of bands about 100 bases in length (see lane J). Thus the 60 base ladder is a rough estimate of total H3.2 gene expression, while the 100 base band represents specifically expression from the H3.2 gene used as probe. Note that, as mentioned briefly above, the 60 base band in lanes A-F is rather weak. This either means that the H3.3-2 gene used as probe in these lanes is the major H3.3 gene transcribed or, more likely, that the H3.3 gene family is rather divergent in DNA sequence between the ATG and the PvuII site. This would result in S1 digestion of the partially homologous RNA:DNA hybrids between the H3.3-2 probe and the other H3.3 transcripts which therefore would give rise to little or





no distinct signal on the gel autoradiogram.

As we can see, some differences in expression are evident from this gel between the two H3 genes under study. For both the specific 100 base bands and total 60 base bands the H3.2 probe hybridizes most strongly in lane J, moderately in lane G, and least in lanes H and I. Lane J represents dividing tissues (CEF) versus lanes H and I which contain mature nondividing tissue. (As mentioned above, the breast muscle RNA used for lane I may be low in intact poly(A)<sup>+</sup> mRNA.) From this we can see that the major chicken H3 histone, H3.2 represents a member of a class of genes whose expression is regulated more closely according to the replication state of the tissue examined than is that of the H3.3 variant, H3.3-2. However, this difference in relative expression is not as striking as was previously proposed for the H3.2 genes relative to the H3.3-1 gene (Engel, Sugarman, and Dodgson). This is probably because our results and those on H3.3-1 (J.D. Engel, personal communication) suggest that the H3.3-2 gene is somewhat (3-10 fold) more active than the H3.3-1 gene, although the relative activity of the two genes in different tissues probably does not vary extensively. The regulation of H3.3 histone expression is considered further in the Discussion.



## DISCUSSION

This thesis details the primary sequence and expression of a chicken H3 histone gene variant which contains intervening sequences. This variant histone gene, designated H3.3-2 and the similar variant histone gene H3.3-1 isolated by Engel, Sugarman and Dodgson (21) are members of the replacement class of histone genes. The replacement histones are not to date as completely characterized as are the replication histones. However, the major difference which, by definition, sets these two histone classes apart relates to the regulation of their expression. The replication histones are used to meet the needs of rapidly dividing cells which require extensive histone synthesis to organize their newly synthesized DNA into chromatin. Expression of the replication histone genes is tightly coupled to DNA synthesis and thus occurs only in S-phase of the cell cycle. On the other hand, the replacement histones are expressed in relatively low levels throughout the cell cycle, in both dividing and nondividing tissues. Although at present we do not understand why different histone variants are required with these two types of expression pattern, our initial goal has been to ascertain structural differences between the two gene types, which could be related to their differences in expression. The data presented here, along with that presented by Engel, Sugarman and Dodgson (21) allow us to



identify several common structural differences which exist between the replacement and replication histone genes that have been studied to date.

One of the first clues that the replacement histone genes might be structurally different arose as part of the isolation procedure of the H3.3 clones, although it was not realized at the time. As stated earlier, the H3.3-2 gene described herein came from a  $\lambda$  recombinant clone selected from a pool of histone clones because of its unique hybridization to oligo(dT)-primed cDNA made against anemic red cell poly(A)+ mRNA. This probe was used in an attempt to identify the histone H5 gene since H5 histone is uniquely expressed in red cells and since H5 mRNA is polyadenylated while that of most other histones is not. As it turned out several  $\lambda$  recombinant histone gene clones hybridized to this probe, and it has since been shown that none of these contain H5 sequences. Of the two such recombinants characterized to date ( $\lambda$ CH4b,  $\lambda$ CH6b) both have been shown to contain H3.3 variant histones of the replacement class. Whether or not the remaining as yet uncharacterized histone clones contain H3.3 variants or other variants is unknown. It is conceivable that a subfamily of replacement histone H3.3 variants may exist. Two such genes have been characterized in detail and it seems quite possible that other H3.3 genes exist. In any case, the hybridization to oligo(dT)-primed cDNA suggested, and we have later confirmed, that both replacement genes code for polyadenylated mRNA while all replication genes we have studied code primarily for non-polyadenylated mRNA.

The work of Engel, Sugarman and Dodgson (21) had shown that the initial H3.3 gene studied differed in several ways from its



replication variant analogue, the H3.2 gene. In order to learn whether the differences were common to all histone gene variants or were idiosyncratic, we decided to analyze the structure of another likely variant gene, the gene we now refer to as the H3.3-2 gene. The initial and most important part of this work involved determination of the complete nucleotide sequence of this gene.

Careful examination of the sequence of H3.3-2 reveals a wealth of information. The first thing we can do is infer the amino acid sequence coded for by H3.3-2 and compare this sequence to other H3 histones (see Figure 5). Even though  $\lambda$ CH6b hybridized to the H3 histone specific probe (18), positive proof that H3.3-2 codes for an H3 histone peptide arose only through this comparison. Note that H3.3-2 codes for the identical polypeptide that H3.3-1 does (thus the use of the term H3.3-2) and exhibits the same four codon substitutions from H3.2 (see Figure 5). However, despite the fact that they code for identical amino acid sequences, analysis of the primary sequence of H3.3-1 and H3.3-2 points out that these two genes are really quite different. By examining Figure 5, we can see that H3.3-2 differs from H3.2 (major chicken replication H3 polypeptide) by 78 nucleotides, roughly 19% divergence in primary sequence. Some degree of sequence divergence between H3.3-2 and H3.2 might be expected since the polypeptide sequence varies by 4 codons out of 135 and since the genes differ in expression pattern. However, if we now compare the nucleotide sequence of H3.3-2 to that of H3.3-1 we find a difference of 71 nucleotides or an 18% divergence. This is striking in lieu of the identical amino acid sequence H3.3-2 and H3.3-1 exhibit. This means that the histone variant H3.3-2 has diverged almost as much





from H3.3-1 as it has from H3.2. Since the differences in nucleotide sequence between H3.3-1 and H3.3-2 are completely silent with respect to their amino acid sequence, there obviously has been strong selective pressure to maintain the protein sequence during the considerable evolutionary time in which the primary sequence of H3.3-2 and H3.3-1 diverged. This strong selective pressure immediately suggests two things. First, the two H3.3 genes are probably expressed, and, furthermore, they probably play an essential role in the organism's life cycle. Unfortunately, it is not possible to calculate just when these variant genes diverged evolutionarily from their nucleotide sequence data, since the coding nucleotides have obviously been under strong selective pressure while the high degree of third base differences (83%) suggests that these base pairs have reached an equilibrium level of divergence.

The non-coding portions of the H3.3-2 gene also provided interesting information. The 5' and 3' flanking sequences were analyzed to identify putative consensus sequences known to be essential for transcription of other eukaryotic genes by RNA polymerase II. Upon examination of 5' sequences upstream of the coding region (see Figure 4), several consensus sequences were apparent. At position -31, H3.3-2 contains the sequence TATAAA known as the "TATA" box. At position -93 there is a partially homologous "CCAAT" box, GCAAT. In addition, the sequence CCAATCAG is repeated four times at positions -163, -185, -209 and -266. Even though duplications of this sort have been documented in histone genes before (37), the relative significance of such repetitive CCAAT sequences, if any, for transcription is not known yet. Besides these consensus



sequences necessary for transcription, one also finds a cap site and a start codon (AUG). The cap site or RNA start site was identified by S1 mapping experiments (see Figure 6). Therefore, all the known necessary elements for H3.3-2 to be expressed are present 5' to the coding region.

An examination of the 3' flanking sequences also is interesting. There is a stop codon (UAA) and the sequence AAUAAA which has been linked to the polyadenylation of eukaryotic messenger RNAs (38). As described above, it appears that H3.3 mRNA exists within the poly(A)+ RNA class unlike that of the replication variants. This is confirmed for H3.3-2 in Figure 4. The presence of the sequence AAUAAA 3' to H3.3-2 coding sequences reinforces this idea. Proudfoot has shown that when the sequence AAUAAA is altered, mRNAs are not polyadenylated and improper termination occurs. Recently Gil and Proudfoot (39) have demonstrated that the sequence AAUAAA is not sufficient alone to insure proper excision of messenger RNAs. Sequences some 35 bp downstream appear to be important. With this in mind I have examined the nucleotide sequence of H3.3-2 downstream of the sequence AAUAAA. It is interesting to note that H3.3-2 contains an inverted repeat (ACTTGATCCAAGT) separated by the three nucleotides ATC some 10 base pairs after the AAUAAA sequence. From what we know of other RNAs, secondary structures can play an important role in termination. This inverted repeat could form a hairpin loop in the transcript, and provide a secondary signal for proper excision. Whether or not this is the case, can't be determined at this time. However, it does appear that all the known necessary consensus sequences are present

in the flanking regions for H3.3-2 to be expressed into a post-transcriptionally polyadenylated mRNA.

Since the H3.3-1 gene was the first histone gene shown to contain intervening sequences, we were especially interested in examining the corresponding introns of the H3.3-2 gene. As described above, the H3.3-2 gene contains three introns, two of which interrupt the coding sequence (at locations identical to the two coding introns in H3.3-1). Consensus sequences have been shown for both donor (5' end of intron) and acceptor (3' end of intron) sites of introns (36). If we look at Figure 9 we can see that while some sequence polymorphisms exist, the H3.3-2 introns contain the appropriate consensus sequences and should be spliced correctly. Changes in the GT/AG ends of an intron usually result in the loss of splicing activity. Weiringa et al. (40) have shown that these two sites are not the only prerequisites for proper splicing though. After constructing a mini-intron (30 nt.) containing the 5' and 3' consensus sequences of the large intron of rabbit  $\beta$ -globin they found that it could not be spliced. Keller and Noon (41) suggest that there may be a conserved internal signal sequence some -60 to -10 nt. from the 3' splice site which may be required in addition to the 5' and 3' consensus sequences. Keller (42) even proposes how such a sequence could be essential for correct splicing of mRNAs in the RNA Lariat model of splicing. Suffice it to say that this work is still in the speculative stage, but it is interesting to note the proposed conserved internal consensus signal is  $\text{CTAA}^{\text{G}}\text{C}^{\text{C}}\text{T}$  (41). In H3.3-2 the sequence CTCAC is contained within the intron present in the 5' nontranslated region and also in the first intron (80 nt.) within the coding region. The sequence CTAAC is contained in the second intron (88 nt.) within the



coding region. Both the CTCAC and CTAAC sequences within these introns are positioned -60 to -10 nt. from the 3' end of the respective introns (see Figure 4).

Besides considering intron splicing, a comment should also be made in regards to the sizes of the introns themselves. If we compare the two introns within the coding region of H3.3-2 to the analogous introns of H3.3-1, we can see that the introns are present at exactly the same location relative to the nucleotide sequence. However, if we now compare the sizes of analogous introns to each other we see that those of H3.3-1 (764 nt., 2.9 kb) are more than an order of magnitude larger than those of H3.3-2 (80 nt., 88 nt.). This is surprising since intron size has been shown to be fairly constant among  $\alpha$ - and  $\beta$ -globin genes. The fact that H3.3-2 and H3.3-1 genes exhibit large differences in intron size, is further proof that these two genes probably diverged very long ago.

At this point I think it would be advantageous to compare the data presented herein regarding H3.3-2 to what is known about H3.3-1. It was stated previously that such a comparison might reveal several distinguishing characteristics exclusive to the replacement histone genes as a whole. H3.3-2 and H3.3-1 have several things in common. For instance, both genes contain introns and these introns occur in both the coding region and the 5' nontranslated region. Both genes contain long 5' and 3' nontranslated regions with a 5' nontranslated leader sequence requiring splicing to the coding region of the mRNA. Both genes code for an identical H3.3 variant histone protein and both have transcripts which are polyadenylated. As far as differences are concerned, I have already dealt with the variation in nucleotide sequence and intron size. However, I did not mention the notable





difference in primary transcript size (1.5 kb vs. >5 kb) or that the exact number of introns within the 5' nontranslated region of H3.3-1 has not been determined yet (J.D. Engel, personal communication). It appears more than one intron may be present in this region of the H3.3-1 gene.

From such a comparison, what can we now conclude There are at least two histone genes (maybe more) which constitute the H3.3 replacement variant histone genes. Whether or not this family is larger is unknown at this time. Furthermore, the two H3.3 genes we have studied are the only true replacement histone genes characterized to date. The red cell-specific H5 histone has occasionally been called a replacement histone, but it more probably represents a different histone class - the tissue-specific histones. As other examples of replacement genes surface, as I feel they will, our understanding of these variant histone genes will grow. Already the two H3.3 genes exhibit similarities to the tissue-specific H5 histone gene. H5 mRNA is polyadenylated and, like the H3.3 genes, is not linked to other histone genes, however it does not contain intervening sequences. It is possible that polyadenylated messenger RNA and intervening sequences may be two characteristics of all true replacement histone genes. Whether this is true or not is not possible to say at this time, but this data and that of Engel, Sugarman and Dodgson do suggest areas in which to concentrate research efforts aimed at further characterizing replacement histone genes and their expression.

The final topic to be considered in this report is the expression of H3.3-2. Throughout the discussion of the structure of the H3.3-2



gene, I have noted that the sequence data suggested that H3.3-2 was expressed. Indeed, the success of S1 nuclease analysis used to pinpoint splice sites and the start of transcription (see Figures 6,10) confirm that H3.3-2 must be expressed. It is possible for me to make this statement based on information presented earlier in the Results (see above) section. Previously I detailed how a singly end-labeled DNA fragment was hybridized to RNA and then subjected to S1 nuclease digestion. Only regions where RNA:DNA hybrids form are protected, since S1 nuclease specifically degrades single-stranded DNA. (Conditions of the reaction are optimized to prevent as much reannealing of the probe DNA (DNA:DNA hybrids) as possible to decrease spurious results). The source of RNA was total cytoplasmic red cell RNA and the only way the single-stranded singly end-labeled DNA probe would not be completely degraded occurred when a complementary messenger RNA existed in the mRNA pool that would hybridize to the DNA probe (RNA:DNA hybrid). The fact that a 75 bp. signal showed up in the Results in Figure 6 was evidence that a H3.3-2 transcript must exist in the total cytoplasmic red cell RNA pool. Additionally, the argument was made in the Results based on the intensities of the H3.3-2 and H3.2 probe signals that the 75 bp. protected fragment was specific for the H3.3-2 gene itself and no other (see below). Thus, it is possible from the S1 data to conclude that H3.3-2 is expressed.

In addition, the S1 analysis revealed several other characteristics of H3.3-2 expression. If we look at the S1 analysis (see Figure 10) we see further evidence of the sequence divergence that must exist among the H3.3 replacement variants. A strong signal is apparent at 75 bp with very little signal at 60 bp for H3.3-2.



However, just the opposite is true for H3.2: a strong signal is present at 60 bp for H3.2 and a weaker signal at 100 bp. As explained in the Results, a strong 60 bp signal is representative of expression of all H3.2 genes whereas the weak 100 bp band indicates expression of one particular H3.2 gene. Likewise, the strong 75 bp band indicates the specific expression of H3.3-2 and the weak band at 60 bp presumably results from weak cross-hybridization with other members of the family of H3.3 genes. H3.3-2 thus shows very little hybridization to other members of its family which might be expected if all the putative H3.3 genes are as different in primary sequence as are H3.3-1 and H3.3-2. Thus we can see that the sequence data of H3.3-2 and H3.3-1 along with the S1 data point out that the H3.3 genes probably exist as a much more divergent population than that of the H3.2 subfamily. This is confirmed by sequence comparisons of two H3.2 genes analogous to our comparison of H3.3-1 and H3.3-2 (Doug Engel, personal communication).

An offshoot of this is that the signals in lanes A through F (see Figure 10) represent very specific indications of the level of H3.3-2 expression in various tissues. Note that H3.3-2 appears to be expressed in both dividing and nondividing tissues at a relatively low (basal) level. This trait is characteristic of replacement variant histones as described by Wu and Bonner (26). In contrast, H3.2 is expressed to a greater extent in dividing tissue, which is characteristic of replication variant histones. From the S1 analysis and other data not presented we can ascertain that H3.3-2 is expressed at approximately 5% of the level of total H3.2 expression in nondividing or slowly-dividing adult tissues such as reticulocytes and



liver. Despite this, H3.3 has been shown to account for up to 50% of histone H3 levels in adult (nondividing) tissue (43). From the S1 data it is possible to see that H3.3-2 shows no large increase in expression between dividing and nondividing tissues. This poses a puzzle then of how the increase in H3.3 histone content can be accounted for. It is doubtful the increase can be attributed solely to an increase in the level of transcription since the S1 data presented here does not bear this out. More than likely other factors are also involved such as differences in turnover rate and translational efficiency. It might even be possible that as H3.2 levels decrease, H3.3 may be able to compete more effectively for binding to DNA and thus increase its relative protein stability. Whatever the mechanism, additional work must be done before we completely understand the functional differences between replication and replacement variant histones and the regulation of their expression.





## REFERENCES



## REFERENCES

1. McGhee, J.D., and Felsenfeld, G. (1980) *Ann. Rev. Biochem.* 49:1115-1153.
2. Isenberg, L. (1979) *Ann. Rev. Biochem.* 48:159-191.
3. Weisbrod, S., Groudine, M., and Weintraub, H. (1980) *Cell* 19:289-301.
4. Lewin, B. (1980) *Gene Expression* 2, Wiley Interscience, New York, pp. 915-918.
5. Darnell, J.E., Jr. (1982) *Nature* 297:365-371.
6. Nurse, P. (1983) *Nature* 302:378.
7. Kedes, L. (1979) *Ann. Rev. Biochem.* 48:837-870.
8. Zwiedler, A. (180) in *Gene Families of Collagen and Other Proteins* (Prockop and Champe, eds.), Elsevier North Holland, Inc., p. 47-56.
9. Lewin, B. (1983) *Genes*, John Wiley and Sons, Inc., New York, pp. 456-499.
10. Hentschel, C., and Birnsteil, M. (1981) *Cell* 25:301-313.
11. Zwiedler, A. (1977) *Methods Cell. Biol.* 17:223-233.
12. Stein, G., Stein, J., and Marzluff, W. (1984) *Histone Genes*, John Wiley and Sons, New York.
13. Showman, R.M., Wells, D.E., Anstrom, J., Hursh, D.A., and Raff, R.A. (1982) *Proc. Natl. Acad. Sci. USA* 79:5944-5947.
14. Karp, R. (1979) Ph.D. Thesis, Stanford University, Palo Alto, California.
15. Crawford, R.J., Krieg, P., Harvey, R.P., Hewlish, D.A., and Wells, J.R.E. (1979) *Nature* 279:132-136.
16. Engel, J.D., and Dodgson, J.B. (1981) *Proc. Natl. Acad. Sci. USA* 78:2856-2860.
17. Harvey, R.P., Krieg, P.A., Robins, A.J., Coles, L.S., and Wells, J.R.E. (1981) *Nature* 294:49-53.



18. Sugarman, B.J., Dodgson, J.D., and Engel, J.D. (1983) J. Biol. Chem. 258:9005-9016.
19. Krieg, P.A., Robins, A.J., Gait, M.J., Titmas, R.C., and Wells, J.R.E. (1982) Nucleic Acids Res. 10:1495-1502.
20. Ruiz-Vasquez, R., and Ruiz-Carillo, A. (1982) Nucleic Acids Res. 10:2093-2108.
21. Engel, J.D., Sugarman, B.J., and Dodgson, J.B. (1982) Nature 297:434-436.
22. Urban, M.K., Franklin, S.G., and Zwiedler, A. (1979) Biochemistry 18:3952-3960.
23. Breathnach, R., and Chambon, P. (1981) Ann. Rev. Biochem. 50:349-383.
24. Woudt, L.P., Pastink, A., Kempers-Veenstra, A.E., Jansen, A.E.M., Magen, W.H., and Planta, R.J. (1983) Nucleic Acids. Res. 11:5347-5360.
25. Childs, G., Nocente-McGrath, C., Lieber, T., Holt, C., and Knowles, J.A. (1982) Cell 31:383-393.
26. Wu, R.S., Tsai, S., and Bonner, W.M. (1982) Cell 31:367-374.
27. Maniatis, T., Fritsch, E.F., and Sambrook, J. (1982) Molecular Cloning: A laboratory manual. Cold Spring Harbor Press, Cold Spring Harbor, New York.
28. Hanahan, D. (1983) J. Mol. Biol. 166:557-580.
29. Girvitz, S.C., Bacchetti, S., Rainbow, A.J., and Graham, F.L. (1980) Anal. Biochem. 106:492-496.
30. Maxam, A., and Gilbert, W. (1980) Meth. In Enz. 65:499-560.
31. Smith, D.R., and Calvo, J.M. (1980) Nucleic Acids. Res. 8:2255-2274.
32. Jay, E., Seth, A.K., Romens, J., Sood, A., and Jay, G. (1982) Nucleic Acids Res. 10:6319-6329.
33. Simonesitis, A., and Torok, I. (1982) Nucleic Acids. Res. 10:7959-7964.
34. Longacre, S., and Rutter, W.J. (1977) J. Biol. Chem. 252:2742-2752.
35. Chirgwin, J., Przybyla, A., MacDonald, R.J., and Rutter, W.J. (1979) Biochemistry 18:5294-5299.
36. Mount, S. (1982) Nucleic Acids Res. 10:459-471.



37. Grandy, D.K., Engel, J.D., and Dodgson, J.B. (1983) in Gene Expression, Alan R. Liss, Inc., New York, pp. 445-455.
38. Proudfoot, N.J. and Brownlee, G. (1976) Nature 263:211-216.
39. Gill, A., and Proudfoot, N.J. (1984) Nature 312:473-474.
40. Wieringa, B., Hofer, E., and Weissman, C. (1984) Cell 37:915-925.
41. Keller, E.B., and Noon, W.A. (1984) Proc. Natl. Acad. Sci. USA 81:7417-7420.
42. Noon, W.A. (1984) Cell 39:423-425.
43. Urban, M., and Zwiedler, A. (1983) Dev. Biol. 95:421-428.







MICHIGAN STATE UNIVERSITY LIBRARIES



3 1293 03082 1783