5-1990

Conservation of intron position indicates separation of major and variant H2As is an early event in the evolution of eukaryotes

A van Daal
Washington University in St. Louis

E White

Sarah C.R. Elgin
Washington University in St. Louis, selgin@wustl.edu

M Gorovsky

J Mol Evol (1990) 30:449–455

Journal of Molecular Evolution

Conservation of Intron Position Indicates Separation of Major and Variant H2As Is an Early Event in the Evolution of Eukaryotes

A. van Daal,[1,*] E.M. White,[2] S.C.R. Elgin,[1] and M.A. Gorovsky[2]

[1] Department of Biology, Washington University, St. Louis, Missouri 63130, USA
[2] Department of Biology, University of Rochester, Rochester, New York 14627, USA

Summary. Genomic clones of *Drosophila* and *Tetrahymena* histone H2A variants were isolated using the corresponding cDNA clones (van Daal et al. 1988; White et al. 1988). The site corresponding to the initiation of transcription was defined by primer extension for both *Drosophila* and *Tetrahymena* genomic sequences. The sequences of the genomic clones revealed the presence of introns in each of the genes. The *Drosophila* gene has three introns: one immediately following the initiation codon, one between amino acids 26 and 27 (gln and phe), and one between amino acids 64 and 65 (glu and val). The *Tetrahymena* gene has two introns, the positions of which are identical to the first two introns of the *Drosophila* gene. The chicken H2A.F variant gene has been recently sequenced and it contains four introns (Dalton et al. 1989). The first three of these are in the same positions as the introns in the *Drosophila* gene. The fourth intron interrupts amino acid 108 (gly). In all cases the sizes and the sequences of the introns are divergent. However, the fact that they are in conserved positions suggests that at least two of the introns were present in the ancestral gene. A phylogenetic tree constructed from the sequences of the variant and major cell cycle-regulated histone H2A proteins from several species indicates that the H2A variant proteins are evolutionarily separate and distinct from the major cell cycle-regulated histone H2A proteins. The ancestral H2A gene must have duplicated and diverged before fungi and ciliates diverged from the rest of the eukaryote lineage. In addition, it appears that the variant histone H2A proteins analyzed here are more conserved than the major histone H2A proteins.

*Present address: Department of Biochemistry, G.P.O. Box 498, Adelaide, S.A. 5001, Australia
Offprint requests to: A. van Daal

Key words: Histone H2A — Histone variant — Intron position — *Drosophila* — *Tetrahymena*

Introduction

Histones are among the most highly conserved proteins found in eukaryotes. They are small, basic proteins that have the conserved function of packaging DNA within the nucleus. Two molecules of each of the four core histones, H2A, H2B, H3, and H4, interact to form a core particle around which DNA is wrapped. In addition to the major cell cycle-regulated histone proteins there are quantitatively minor nonallelic variants that differ significantly from their major histone counterparts. A number of recent studies have shown that although the H2A variants are only about 60% similar in amino acid sequence to the major histone H2As within a species, the variants themselves are highly conserved across species (85–100% similarity). This argues that they have an important function that differs in some aspects from that of the major H2A histone (Ernst et al. 1987; Hatch and Bonner 1988; van Daal et al. 1988; White et al. 1988). A number of observations have led to the suggestion that H2A variants may be preferentially associated with transcriptionally active chromatin (reviewed by Gorovsky 1985). The most compelling evidence comes from the fact that the H2A variant of *Tetrahymena* is found exclusively in the transcriptionally active macronucleus and is absent from the transcriptionally inert micronucleus of *Tetrahymena* (Allis et al. 1980).

In addition to the differences between the major and variant histone H2As seen at the level of the protein sequence, there are also differences in the

organization and expression of the genes (White and Gorovsky 1988; Dalton et al. 1989). Where analyzed, the major histone H2A genes are found to be transcribed primarily during S-phase, whereas the H2A variant genes are expressed throughout the cell cycle (reviewed in Schumperli 1988). In multicellular eukaryotes the H2A variant mRNAs are polyadenylated (Ernst et al. 1987; Hatch and Bonner 1988; van Daal et al. 1988), and the genes contain introns (Dalton et al. 1989; see below). This contrasts with the cell cycle-regulated histone genes, which, in most organisms, lack introns and encode mRNAs that are not polyadenylated.

In this paper we describe the sequencing of the genes for the H2A variants of *Drosophila* (H2AvD) and *Tetrahymena* (hv1). We show the presence of three introns in the *Drosophila* gene and two introns in the *Tetrahymena* gene. The two introns in hv1 are in identical positions to the two 5′-most introns found in H2AvD. The chicken H2A.F variant gene has recently been sequenced and shown to contain four introns (Dalton et al. 1989). Interestingly, the three introns seen in the *Drosophila* gene are in identical positions to the first three introns found in the chicken gene.

Materials and Methods

Isolation of Genomic Clones. *Drosophila* H2AvD genomic clones were obtained by screening a *Drosophila* Canton S library (Maniatis et al. 1978) with an H2AvD cDNA clone insert (van Daal et al. 1988), which was labeled with ^{32}P by nick translation. Filters containing plaques representing eight genome equivalents were prehybridized for 4 h at 65°C in 5× standard saline citrate (SSC: 0.15 M NaCl, 0.015 M sodium citrate), 2× Denhardt's (0.04% Ficoll, 0.04% polyvinylpyrrolidone 360, 0.04% BSA), 0.2% SDS, and 50 μg/ml salmon sperm DNA. Hybridization was overnight at 65°C in the same solution plus 6% polyethylene glycol (molecular weight 8000) and 2 × 10^6 cpm/ml radioactive probe. Filters were washed twice for 30 min each at room temperature and twice at 65°C in 0.1 × SSC, 0.2% SDS.

Tetrahymena hv1 genomic clones were isolated from a partial library constructed from HindIII-cut size-selected genomic DNA. Genomic Southern blots revealed that a 2.2-kb HindIII fragment contained the hv1 gene (data not shown). Genomic DNA was digested with HindIII and electrophoresed on a 1% low melting temperature agarose gel. HindIII fragments (2.2-kb) were extracted from the agarose using Elutip columns (Schleicher & Schuell Inc.) according to the manufacturer's suggestions, except that the entire procedure was carried out in a 37°C room. These fragments were ligated to HindIII-digested vector DNA (Bluescript-Stratagene Cloning Systems) and then used to transform *Escherichia coli* (DH5α). Colonies from the resultant library were screened with the hv1 cDNA clone insert (White et al. 1988), which was labeled with ^{32}P by the random primer method (Feinberg and Vogelstein 1983). Filters were prehybridized at 65°C in 5 × SSPE (SSPE: 0.18 M NaCl, 10 mM NaH_2PO_4, 1 mM EDTA pH 7.4), SPED (0.1% Ficoll, 0.1% polyvinylpyrrolidone 360, 0.1% BSA), and 1 mg/ml BSA for 6 h and hybridized overnight in the same buffer with 2.5 × 10^5 cpm/ml of radioactive probe. Filters were washed in 2 × SSPE, 0.1% SDS for 15 min four times at room temperature and twice at 65°C.

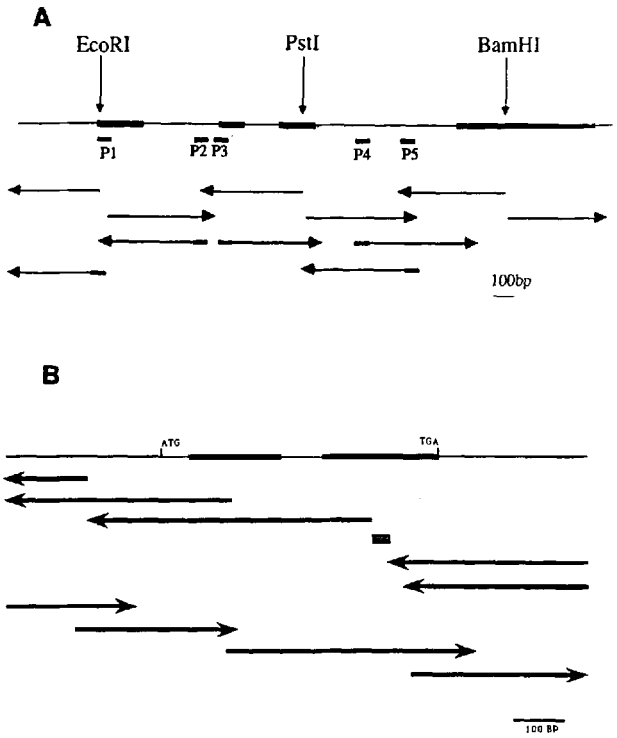

Fig. 1. **A** Sequencing strategy of H2AvD. Fragments of the H2AvD genomic clone were subcloned into M13 and the regions sequenced by the dideoxy chain termination method are indicated (Sanger et al. 1977). The positions of five oligonucleotide primers synthesized to complete the sequencing of both strands are labeled P1–P5, and the regions sequenced using these primers are indicated. The heavy line represents regions contained in the cDNA clone (van Daal et al. 1988). **B** Sequencing strategy of hv1. Arrows represent the regions of deletion clones sequenced. The hv1 genomic clone is represented at the top with the coding region in bold. The box shows a short region (~40 bp), which was only sequenced on one strand, but which has been confirmed in the cDNA clone (White et al. 1988).

Sequencing. Fragments of the H2AvD genomic clone indicated in Fig. 1A were subcloned into M13 and sequenced by the dideoxy chain termination method (Sanger et al. 1977). In addition, five oligonucleotide primers were constructed and used to complete the sequencing of larger inserts. The *Tetrahymena* hv1 genomic clone was sequenced using a set of unidirectional deletions (indicated in Fig. 1B) made from both orientations of the 2.2-kb HindIII fragment utilizing the Bluescript mung bean nuclease/exonuclease III system as described by the manufacturer (Stratagene Cloning Systems). Sequencing was done by the dideoxy chain termination method using the Sequenase [^{35}S]dATP sequencing kit (United States Biochemical Corp.).

Primer Extension. Oligonucleotide primer (0.2 pmol, 5′GAGCGTGCGAGTTGTTT) was end-labeled with [^{32}P]dATP and then annealed to 5 μg of denatured total *Drosophila* RNA in the presence of 20 units of RNAsin and 20 mM Tris-HCl, pH 8.3, 7.5 mM $MgCl_2$, 7.5 mM KCl, 2 mM dithiothreitol by incubation at 42°C for 1 h. Extension was carried out by the addition of dATP, dCTP, dGTP, and dTTP to 0.6 mM and 20 units of reverse transcriptase. Samples were incubated at 42°C for 1 h and analyzed by denaturing polyacrylamide gel electrophoresis.

Primer extension with *Tetrahymena* RNA isolated either from growing or starved cells was performed with an oligonucleotide

```
1     TAACCCAGTA GGACTACTGT AAAAACGACG TATTTCCAGG CTATTGCCGA GTCATTGAAA
61    CATAAAATAA AAAAGGATGA TTATGATTTC AAGGATATAT TACGATAGAC AGCTCTGCAA
121   AACGGCGCCA CTTGGCAATG TGCGGCCACA AAAGGAAGCT ACCATCTAGG GGTGAGGGCA
181   TATCGATTGG CCAAACATCG ATATATTCCG TTCCACCCCT GGCGTTCCGC TGACGGCGTC
241   TTTTCCCCGA AAAAATTTCA AGTAGTCGAA ACCGAATTCC GTAGAAACAA CTCGCACGCT
301   CCGGTTTCGT GTTGCAACAA AATAGGCATT CCCATCGCGG CAGTTAGAAT CACCGAGTGC
361   CCAGAGTCAC GTTCGTAAGC AGGCGCAGTT TACAGGCAGC AGAAAAATCG ATTGAAGAGA
421   AATGGTAATA TTGCGGTGAA TTTTTGAGCG GCAGCGCATC TCGCTTTTCC CACTAGCGCT
       Met
481   GCCCCCGCGA ATTAAATATT GAAGTGCGGG GATTACGGAG ACGCGGACAA CAAATGCACG
541   CAAATGGACA CCACGTAGGC CGCAACAACA AGGAGCAGCG AAATCGGCGG GCATTGTTAT
601   TGTGCTGCTA GCGGGCAATC GGCCGTAACC TCACTTTGGG CAGCGGTGCC TCCATTTTGT
661   GTTTTCTTCT GCGACTCGTG CCACATTCGC TTTAATTCGT TATTTTAAGA AATGCATTGT
721   GCTGTGTCTA TTCGCAGGCT GGCGGTAAAG CAGGCAAGGA TTCGGGCAAG GCCAAGGCGA
                     Ala GlyGlyLysA laGlyLysAs pSerGlyLys AlaLysAlaL
781   AGGCGGTATC GCGTTCCGCG CGCGCGGGTC TTCAGGTGAG TTTTACAATC TGTACCTCGG
      ysAlaValSe rArgSerAla ArgAlaGlyL euGln
841   TTTTGGCGCC TTTTCGGTAC CCACTCTGCC AACGAGCATA TTCCACACAT AACAGGGTGG
901   CCCGCCTTCG TTGGTTGGTT GGGGTTGACA AAAACATAAC AAGCCAGCCG GTCAAGCCAA
961   TTAACCAGTT ACCTGCCTCT TATAGTCCCG GACAGTTCCC CGTGGGTCGC ATCCATCGTC
                                          PhePr oValGlyArg IleHisArgH
1021  ATCTCAAGAG CCGCACTACG TCACATGGAC GCGTCGGAGC CACTGCAGCC GTGTACTCCG
      isLeuLysSe rArgThrThr SerHisGlyA rgValGlyAl aThrAlaAla ValTyrSerA
1081  CTGCCATATT GGAATACCTG ACCGCCGAGG TAAGTGTGCT TCCGCCGAGT TTTCCCGCTT
      laAlaIleLe uGluTyrLeu ThrAlaGlu
1141  TTCTCTCGTG GTACTTTTTC TCGCTTCAAT GTAAACAGCA TTTACCAGAA ATATTCAAGG
1201  GAAACATTTT TTTTTCCACA GTTGCATTCT CTCCACTTTT CTTATTGTGT ACAGTTATTT
1261  CGAATTTCCC CGTTCCAATT CAAAAACACA CTTTCTTATT GACATTTCGT GTGGAATGAG
1321  AATGATGAAA ATGTACATCT AATTCATGAA ATGCTCGAGG CGCACATTAT CAGTTGTTTA
1381  AATCTAAATA ACAGTAGTTT CGATTTAATG TATAACGCTT TAATTTTCTA GATATACAGT
1441  TTATAGAGTT TGCGCCATTA GCAGTACACT TTCCGCTTTT CCGAATATAA CGTTACGTGA
1501  GTTTTGTTTT AATGTGTGAT GTTATTAAAG AAATTTATTA AAAGATTCAA AGTTTGTTTT
1561  GGTAATGAGC GATACAAGTA AATAAATTCG CTTTAGCTTT TCGTTATCCT CAATTGAAAT
1621  CTTTATTAAT TTCTTTGTTT CCCAATGCAT TTAGGTCCTG GAGTTGGCAG GCAACGCATC
                                        ValLeu GluLeuAlaG lyAsnAlaSe
1681  GAAGGACTTG AAAGTGAAAC GTATCACTCC TCGCCACTTA CAGCTCGCCA TTCGCGGAGA
      rLysAspLeu LysValLysA rgIleThrPr oArgHisLeu GlnLeuAlaI leArgGlyAs
1741  CGAGGAGCTG GACAGCCTGA TCAAGGCAAC CATCGCTGGT GGCGGTGTCA TTCCGCACAT
      pGluGluLeu AspSerLeuI leLysAlaTh rIleAlaGly GlyGlyValI leProHisIl
1801  ACACAAGTCG CTGATCGGCA AGAAGGAGGA AACGGTGCAG GATCCACAGC GGAAGGGCAA
      eHisLysSer LeuIleGlyL ysLysGluGl uThrValGln AspProGlnA rgLysGlyAs
1861  CGTCATTCTG TCGCAGGCCT ACTAAGCCAG TCGGCAATCG GACGCCTTCG AAACATGCAA
      nValIleLeu SerGlnAlaT yrEnd
1921  CACTAATGTT TAATTCAGAT TTCAGCAGAG ACAAGCTAAA CACGACGAGT TGTAATCATT
1981  TCTGTGCGCC AGATATATTT CTTATATACA ACGTAATACA TAATTATGTA ATTCTAGCAT
2041  CTCCCCAACA CTCACATACA TACAAACAAA AAATACAAAC ACACAAAACG TATTTACCCG
2101  CACGCATCCT TGGCGAGGTT GAGTATGAAA CAAAAACAAA ACTTAATTTA GAGCAAAGTA
2161  ATTACACGAA TAAATTTAAT AAAAAAAACT ATAATAAAAA GCAATCATGT TATTTCAAAA
2221  AAAAAAAACT AGACGAGATG TTGCGCTGTG TCGTA
```

Fig. 2. DNA and derived amino acid sequence of H2AvD. Three introns are found as indicated. The first intron is 313 bp, located immediately after the initiator ATG at position 425; the second is 180 bp at position 816; and the third is 545 bp at position 1110. The transcription start site is indicated by the bold A at nucleotide position 264.

complimentary to the region coding for amino acids 20–29 of hv1 as described (Horowitz et al. 1987) except that the hybridization temperature was 45°C or 55°C. RNAse protection assays were performed using an antisense transcript derived from the hv1 genomic clone as described previously (Horowitz et al. 1987).

Evolutionary Analysis. A phylogenetic tree was constructed from the known or deduced variant and major histone H2A amino acid sequences of *Tetrahymena, Drosophila,* sea urchin, chicken, and human. In addition, the major H2A amino acid sequences of *Saccharomyces cerevisiae, Schizosaccharomyces pombe, Aspergillus,* trout, and *Xenopus* were included (Wells 1986). The tree was constructed by maximum parsimony using the PAUP (phylogenetic analysis using parsimony) program (Swofford 1984).

Results and Discussion

Genomic Sequence of Drosophila H2AvD Gene

Seven overlapping genomic clones were isolated after screening eight genome equivalents of a *Drosophila* genomic library. M13 subclones from two of the clones were used to sequence 2255 bp from both strands. This sequence is shown in Fig. 2 and includes 421 bp of 5′ noncoding and 370 bp of 3′ noncoding sequences. The transcription initiation site is also indicated by bold type.

Start Site of H2AvD Transcripts

A primer complementary to the transcribed sequences 12–29 bp from the EcoRI site, which defined the 5′ end of the *Drosophila* cDNA sequence, was synthesized. This primer is indicated in Fig. 1A and was used to determine the start site of transcription by primer extension analysis. The result shown in Fig. 3 reveals that transcription initiates 10 bp upstream of the EcoRI site. The sequence surrounding the transcription start (AGTAGTC) shows good similarity to the insect consensus sequence for transcription start sites ($ATCA^{G/}{}_{T}T^{C/}{}_{T}$), where the bold A represents the transcription start (Hultmark et al. 1986). This site was confirmed by screening a *Drosophila* 0–4-h cDNA library (Brown and Kafatos 1988) in which full-length cDNAs were stringently selected for prior to cloning. Four clones were isolated, each of which began at the nucleotide defined as the transcription start site by primer ex-

tension (data not shown). It should be noted that there is no TATA sequence in the region 30 bp upstream of the transcription initiation site. The chicken H2A.F transcript has also been analyzed by primer extension and the cap site determined (Dalton et al. 1989). The H2A.F gene also showed a lack of TATA homology in the expected region. Dalton and his colleagues suggest the sequence TTCAAAT, 20 bp upstream of the transcription start site, as a possible site of polymerase interaction. Interestingly, there is a 5-bp sequence (GGCGT) positioned 27 bp and 26 bp upstream from the transcription start sites of the *Drosophila* and chicken genes, respectively, which may serve as a signal for a specific transcription factor. It should be pointed out that another *Drosophila* gene, glyceraldehyde 3-phosphate dehydrogenase, which also lacks a TATA box, does not contain GGCGT in the region 20–30 bp upstream of transcription initiation. This would suggest either that this sequence is specific to the H2A variant genes or to a subclass of genes without TATA sequences, or that its existence in both the *Drosophila* and chicken genes is coincidental.

Genomic Sequence of Tetrahymena hv1 Gene

The *Tetrahymena* size-selected genomic library yielded ~20,000 recombinant colonies from which 12 hv1-positive clones were identified. The genomic sequence of hv1 is shown in Fig. 4. This is the first documentation of the 5′ DNA sequence of hv1, because the previously analyzed cDNA clone was not full length, but terminated seven amino acid residues from the initiation codon, as inferred from the amino acid sequence of the hv1 protein (Allis et al. 1986). It should be noted that the sequence beginning 116 bp downstream of the TGA termination codon is different from the sequence previously described for the cDNA clone. The difference has been shown to be due to a cloning artifact in the cDNA clone (data not shown).

Start Site of hv1 Transcripts

Primer extension using RNA from log-phase *Tetrahymena* hybridized to primer at 45°C yielded multiple bands with the most prominent corresponding to transcription initiation sites 21, 61, and 68 nucleotides upstream of the A of the initiation codon (Fig. 5). When the hybridization was performed at 55°C, only the band 21 nucleotides upstream was observed. When RNA from starved cells was used, the major band observed at both hybridization temperatures corresponded to the initiation site 21 bp upstream of the translation start site. Based on these studies, it seems likely that transcription of the hv1 gene initiates 21 bp upstream of the ATG and that

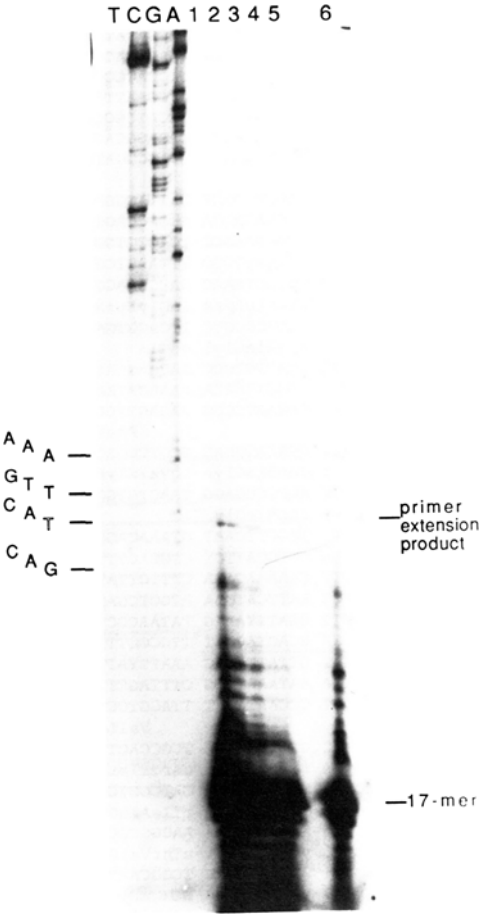

Fig. 3. Mapping of the 5′ terminus of the H2AvD transcript. Primer (P1, 0.2 pmol) was end-labeled and annealed to 10, 5, 1, 0.1, or 0.01 μg of total *Drosophila* RNA (lanes 1–5, respectively). Extension was carried out by the addition of deoxynucleotides and reverse transcriptase. Lane 6 represents labeled primer without the addition of RNA. The first four lanes are sequencing reactions of an M13 clone containing a BglII–PstI fragment (the PstI site is at position 1063 in Fig. 2), which contains the 5′ end of the cDNA clone. The P1 primer was annealed to this clone and sequencing was carried out according to the procedure of Sanger et al. (1977).

the higher molecular weight bands represent extension products derived from nonspecific hybridization to mRNAs that are abundant only in growing cells. It is interesting to note that, despite the extremely (80%) A/T-rich nature of the 5′ noncoding sequences, there is no good TATA sequence 20–30 bp upstream of the transcription start site. There is also no GGCGT sequence in that region.

Conserved Intron Positions

Figure 6 illustrates the alignment of the *Drosophila, Tetrahymena,* and chicken histone H2A variant genomic DNA sequences. Introns I, II, and III from *Drosophila* and intron II from *Tetrahymena* were positioned by comparison of the genomic sequences to the cDNA sequences. Intron I from *Tetrahymena*

```
1     ATACTTTTCT ATTGCAGTAT GAATGAAATA ATATACAATA TGAATGATGA TGATGATAAT
61    AATAAAGGAA AATCAAAGCA AGATTCCAAT TGATCAATCA ATAAATAATC ATAATAATAC
121   CTTCTTTCTT TTATCTTTGA TTTATCATCC TTTTGTTTTG TTTTGATTTG ATTTGATTTT
181   ATTTATTGTG TGATCTTTAA TTATATTGGA TTCTCTTCTC AAGAAGAAGA AAAAGAAAGC
241   AAATAAACAA AAGAAAAAAG TAAAAAGCAA AAAAAGAAAA ATTAAAAGAA AAAGAAATAA
301   AAACAAGTAA AAAAGAAAAT GGTAATTTAA ATTTTTTTTA GAGTGGCTTA ATCTCATTTT
                        Me t
361   TAATCAATTT TCAAATAAAA TAGGCTGGCG GAAAAGGCGG TAAAGGTGGT AAAGGTGGCA
                             AlaGlyG lyLysGlyGl yLysGlyGly LysGlyGlyL
421   AAGGTGGTAA AGTCGGAGGC GCCAAGAATA AGAAGACTCC TCAATCACGT TCTTATAAGG
      ysGlyGlyLy sValGlyGly AlaLysAsnL ysLysThrPr oGlnSerArg SerTyrLysA
481   CTGGTTTATA AGTAATATTG GCATCAGGAT GTCCATCATT CTGAGCATTT GACTCAGTTG
      laGlyLeuGl n
541   TGATTTTATT AACGCTAAAA TATTATAATA ATAAAGTTCC CAGTCGGTAG AATCCACAGA
                                          PheP roValGlyAr gIleHisArg
601   TTTTTGAAGG GTAGAGTTAG TGCTAAGAAC AGAGTTGGTG CTACTGCTGC TGTTTATGCT
      PheLeuLysG lyArgValSe rAlaLysAsn ArgValGlyA laThrAlaAl aValTyrAla
661   GCTGCTATTT TGGAATATTT AACAGCAGAA GTTTTGGAAT TGGCTGGTAA TGCTTCTAAG
      AlaAlaIleL euGluTyrLe uThrAlaGlu ValLeuGluL euAlaGlyAs nAlaSerLys
721   GATTTCAAAG TCAGAAGAAT CACTCCTCGT CACTTGCTCT TGGCTATTAG AGGTGATGAA
      AspPheLysV alArgArgIl eThrProArg HisLeuLeuL euAlaIleAr gGlyAspGlu
781   GAATTAGATA TTTTGATCAA GGCTACCATT GCTGGTGGTG GTGTCATTCC TCACATCCAT
      GluLeuAspI leLeuIleLy sAlaThrIle AlaGlyGlyG lyValIlePr oHisIleHis
841   AAAGCTCTCT TGGGTAAGCA CTCTACTAAA AACAGATCTA GTGCTAAGAC TGCTGAACCT
      LysAlaLeuL euGlyLysHi sSerThrLys AsnArgSerS erAlaLysTh rAlaGluPro
901   CGTTGAGTAG TAATGTACAT GATTTAAAAA AAAATTACAA AACAACTCAA TAAAATTCAA
      ArgEnd
961   TATTATAATA ATTCAACCTA TATATATATT ATTACTATGC TGACTGGTTC ACGGATGGAG
1021  GAAGGAAAAG CAGTACTGCG CTCATAGCAT AAATATTTTT GGAAATTCTC TCAAATACTT
1081  TTATCAACCT CAACATGAAA TAATAATTAA CACTGTAAAA ATACAAAAAA TCTATAAATT
1141  CTCCTACTTC TAAAAGATTG TTCTCAACAC CTAAAGCTCA TTTAGGATTA CTATCTTCAG
1201  ATTCTGCAAG CTCTAAGGAC A
```

Fig. 4. DNA and derived amino acid sequence of hv1. Two introns are found as indicated. The first is 62 bp located immediately after the initiation codon at position 322, and the second is 85 bp located at position 492.

could not be localized in this way, as the cDNA clone is incomplete. However, it is likely that there is also an intron immediately following the initiation codon in *Tetrahymena* for several reasons. There is no ATG codon immediately preceding the first amino acid of the protein. It is unlikely that the hv1 gene uses an alternative initiator codon, and so an intron is inferred. The ATG indicated is the most likely initiator as it is preceded by AGAAA, a sequence very similar to the conserved *Tetrahymena* translation initiation sequence [(A/G)(C/T)A_{3-4}; Horowitz et al. 1987]. Note that immediately following this ATG is a 5′ splice consensus sequence, and RNAse mapping (data not shown) gave the expected 108-nucleotide second exon, arguing strongly that the 3′ end of the intron is at this position. This is the first ATG encountered. The next ATG is found in a cluster of repeated ATGs over 300 nucleotides upstream from the 3′ splice junction. None of these ATGs have a sequence resembling the *Tetrahymena* translation consensus sequence on the 5′ side or a 5′ splice junction on the 3′ side.

The gene for the chicken histone H2A variant (H2A.F) has also been sequenced (Dalton et al. 1989) and has been shown to contain four introns. The positions of the introns are shown in Fig. 6. All three species contain introns I and II in identical positions. However, the sizes of the introns vary and there is no apparent homology except that of the splice sites. Intron I is 62 bp, 313 bp, and 1536 bp long and intron II is 85 bp, 180 bp, and 246 bp long in *Tetrahymena, Drosophila,* and chicken, respectively. Both *Drosophila* and chicken contain a third intron at amino acid position 64, but again the size

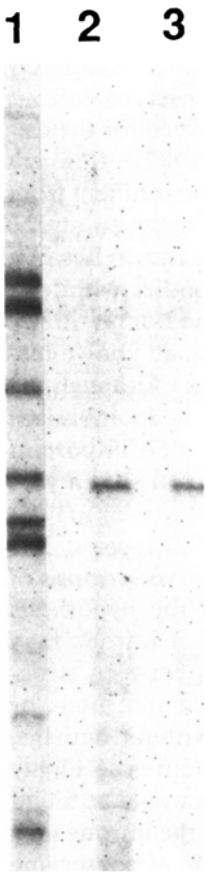

Fig. 5. Mapping of the 5′ terminus of the hv1 transcript. Primer, [0.004 pmol (lane 1) or 0.04 pmol (lanes 2 and 3)], was end-labeled and annealed to 40 μg of *Tetrahymena* total RNA from log cells (lanes 1 and 2) or starved cells (lane 3). Hybridization temperature was 45°C for lanes 1 and 3, and 55°C for lane 2 with extension by reverse transcriptase at 42°C as described previously (Horowitz et al. 1987).

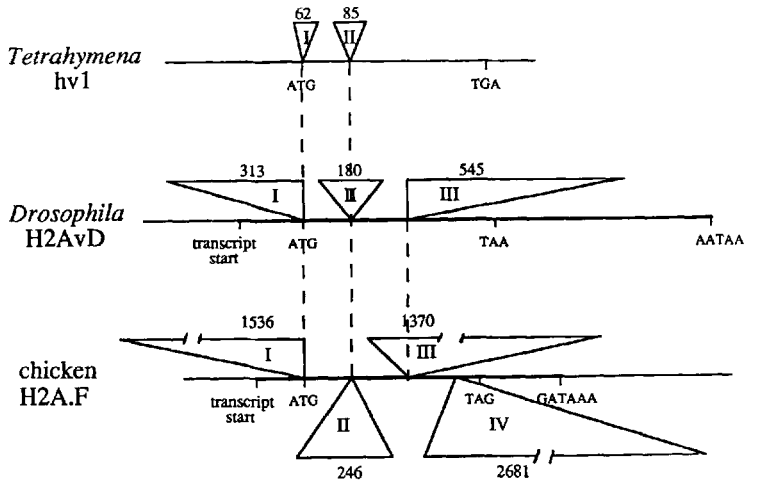

Fig. 6. Comparison of introns in hv1, H2AvD, and chicken H2A variant genes. The positions of the introns are indicated by the triangles. The sizes of the introns are shown in base pairs.

and sequence of the intron is different (545 bp in *Drosophila* and 1370 bp in chicken). Intron IV is unique to the chicken gene and is 2681 bp in length. The data suggest that introns I and II existed prior to the divergence of *Tetrahymena* from other eukaryotes.

The introns in these H2A variant genes all lie in extremely well-conserved regions of the protein. Intron I is present at an unusual location, immediately following the initiation codon. Conserved introns at this position have also been reported for the calmodulin gene family (Smith et al. 1987). Although the amino terminus of the hv1 protein differs from that of H2AvD and H2A.F, the first seven amino acids are well conserved and so the intron lies in a conserved region. The second intron lies within an area known as the H2A box (West and Bonner 1980), which defines a peptide conserved in all known histone H2A and H2A variant proteins. Although the function of this region is not known, its invariant nature suggests that it lies in an extremely important part of the protein. The third intron is also in a very well-conserved region of the protein, similar in both H2A and H2A variant proteins of all species. The location of the introns in well-conserved regions of the protein is not consistent with the hypothesis proposed by Craik et al. (1983) that introns may interrupt protein-coding sequences between structural or functional domains so that insertions or deletions can be accommodated without damage. However, because most of the protein is so highly conserved, it is likely that the introns would be found in such regions, and the location of the introns may not have great significance in terms of conserving the surrounding splice junctions.

Evolutionary Analysis

The availability of the protein sequences for both the variant and major histone H2A proteins of a number of species allowed the construction of the phylogenetic tree shown in Fig. 7. The protein sequence of *S. pombe* histone H2A-A was used as the hypothetical outgroup to construct the most parsimonious tree. The robustness of the tree was illustrated by the fact that exactly the same topology was produced no matter what species was used as the hypothetical ancestor (data not shown). This clearly indicates that the maximum parsimony analysis of the data is appropriate. The consistency index obtained for the tree illustrated in Fig. 7 is 0.839, indicating that greater than 80% of all amino acid changes need only have occurred once. It is not possible to assume an evolutionary clock, as there is too much heterogeneity in the branch lengths of the tree. It is clear, however, that the H2A variants are clustered separately from the major cell cycle-regulated H2As, which indicates that the ancestral H2A gene must have duplicated and diverged into the major and variant type of H2A genes before fungi and ciliates diverged from the rest of the eukaryotic lineage. This is strongly supported by the fact that an H2A variant protein deduced from the gene sequence from *S. pombe* (J. Hindley and P. Nurse, unpublished) falls into the cluster of H2A variants and is clearly distinct from the group of major cell cycle-regulated histone H2A proteins. This, in fact, suggests that the tree is probably rooted between the two nodes marked with circles in Fig. 7.

In addition, it can be seen that the two H2A protein lineages have been under different selective pressures, as the variants have evolved more slowly. We believe that this high degree of conservation among the histone H2A variants (even higher than the major histones) suggests that they have an important function distinct from that of the major histone H2As. The conserved genome organization of the H2A variant genes, different from that of the major cell cycle-regulated H2As, suggests that they

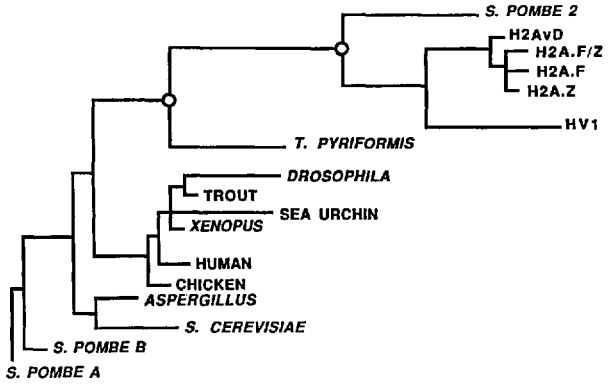

Fig. 7. Phylogenetic tree of H2A sequences. This tree was constructed by the PAUP computer program. The branch lengths are drawn to scale. *S. POMBE 2* (Hindley and Nurse, personal communication), H2AvD, H2A.F/Z, H2A.F, H2A.Z, and HV1 represent the *S. pombe, Drosophila,* sea urchin, chicken, mammalian, and *Tetrahymena* histone H2A variants, respectively. *T. PYRIFORMIS, DROSOPHILA,* CHICKEN, SEA URCHIN, TROUT, *XENOPUS,* HUMAN, *ASPERGILLUS, S. CEREVISIAE, S. POMBE* A, and *S. POMBE* B represent the *Tetrahymena pyriformis, Drosophila,* chicken, sea urchin, trout, *Xenopus,* human, *Aspergillus, S. cerevisiae,* and *S. pombe* A and B major cell cycle-regulated histone H2As, respectively.

have a different form of regulated transcription and RNA processing, which may be related to the distinct function of the H2A variants.

Acknowledgments. This work was supported by grants from the NIH to S.C.R.E. (GM31532) and M.A.G. (GM21793). We are grateful to Josephine Bowen La Rose for technical assistance and to Paul Nurse and John Hindley for permission to use their unpublished data.

References

Allis CD, Glover CVC, Bowen JK, Gorovsky MA (1980) Histone variants specific to the transcriptionally active, amitotically dividing macronucleus of the unicellular eucaryote, *Tetrahymena thermophila.* Cell 20:609–617

Allis CD, Richman R, Gorovsky MA, Ziegler YS, Touchstone B, Bradley WA, Cook RG (1986) hv1 is an evolutionarily conserved H2A variant that is preferentially associated with active genes. J Biol Chem 261:1941–1948

Brown NH, Kafatos FC (1988) Functional cDNA libraries from *Drosophila* embryos. J Mol Biol 203:425–437

Craik CS, Rutter WJ, Fletterick R (1983) Splice junctions: association with variation in protein structure. Science 220: 1125–1129

Dalton S, Robins AJ, Harvey RP, Wells JRE (1989) Transcription from the intron-containing chicken histone H2A.F gene is not S-phase regulated. Nucleic Acids Res 17:1745–1756

Ernst SG, Miller H, Brenner CA, Nocenta-McGrath C, Francis S, McIsaac R (1987) Characterization of a cDNA clone coding for a sea urchin histone H2A variant related to the H2A.F/Z histone protein in vertebrates. Nucleic Acids Res 15:4629–4644

Feinberg AP, Vogelstein B (1983) A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal Biochem 132:6–13

Gorovsky MA (1985) Ciliate chromatin and histones. In: Gall J (ed) The molecular biology of ciliated Protozoa. Academic Press, New York

Hatch CL, Bonner WL (1988) Sequence of cDNAs for mammalian H2A.Z, an evolutionarily diverged but highly conserved basal histone H2A isoprotein species. Nucleic Acids Res 16:1113–1124

Horowitz S, Bowen JK, Bannon GA, Gorovsky MA (1987) Unusual features of transcribed and translated regions of the histone H4 gene family of *Tetrahymena thermophila.* Nucleic Acids Res 15:141–160

Hultmark D, Klemenz R, Gehring WJ (1986) Translational and transcriptional control elements of the untranslated leader of the heat-shock gene hsp22. Cell 44:429–438

Maniatis T, Hardison RC, Lacy E, Lower J, O'Conner C, Quon D, Sim GK, Efstradiadis A (1978) The isolation of structural genes from libraries of eucaryote DNA. Cell 15:687–701

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain terminating inhibitors. Proc Natl Acad Sci USA 74: 5463–5467

Schumperli D (1988) Multi-level regulation of replication dependent histone genes. Trends Genet 4:187–191

Smith VL, Doyle KE, Maune JF, Munjaal RP, Beckingham K (1987) Structure and sequence of *Drosophila melanogaster* calmodulin gene. J Mol Biol 196:471–485

Swofford DL (1984) PAUP—phylogenetic analysis using parsimony. Illinois Natural History Survey, Urbana–Champaign, Illinois

van Daal A, White EM, Gorovsky MA, Elgin SCR (1988) *Drosophila* has a single copy of the gene encoding a highly conserved histone H2A variant of the H2A.F/Z type. Nucleic Acids Res 16:7487–7497

Wells DE (1986) Compilation analysis of histones and histone genes. Nucleic Acids Res 14:r119–r149

West MHP, Bonner WM (1980) Histone H2A, a heteromorphic family of eight proteins. Biochemistry 19:3238–3245

White EM, Gorovsky MA (1988) Localization and expression of mRNA for a macronuclear-specific histone H2A variant (hv1) during the cell cycle and conjunction of Tetrahymena thermophila. Molec Cell Biol 8:4780–4786

White E, Shapiro DL, Allis CD, Gorovsky MA (1988) Sequence and properties of the message encoding *Tetrahymena* hv1, a highly conserved histone H2A variant that is associated with active genes. Nucleic Acids Res 16:179–198

Received July 6, 1989/Revised October 25, 1989