Nucleic Acids Research, 1996, Vol. 24, No. 15 **3023–3030**

Cloning and characterization of the major histone H2A genes completes the cloning and sequencing of known histone genes of *Tetrahymena thermophila*

Xiuwen Liu[+] and Martin A. Gorovsky*

Department of Biology, University of Rochester, Rochester, NY 14627, USA

Received March 18, 1996; Revised and Accepted June 20, 1996

GenBank accession nos L18892 and L18893

ABSTRACT

A truncated cDNA clone encoding *Tetrahymena thermophila* histone H2A2 was isolated using synthetic degenerate oligonucleotide probes derived from H2A protein sequences of *Tetrahymena pyriformis*. The cDNA clone was used as a homologous probe to isolate a truncated genomic clone encoding H2A1. The remaining regions of the genes for H2A1 (*HTA1*) and H2A2 (*HTA2*) were then isolated using inverse PCR on circularized genomic DNA fragments. These partial clones were assembled into intact *HTA1* and *HTA2* clones. Nucleotide sequences of the two genes were highly homologous within the coding region but not in the noncoding regions. Comparison of the deduced amino acid sequences with protein sequences of *T.pyriformis* H2As showed only two and three differences respectively, in a total of 137 amino acids for H2A1, and 132 amino acids for H2A2, indicating the two genes arose before the divergence of these two species. The *HTA2* gene contains a TAA triplet within the coding region, encoding a glutamine residue. In contrast with the *T.thermophila HHO* and *HTA3* genes, no introns were identified within the two genes. The 5′- and 3′-ends of the histone H2A mRNAs were determined by RNase protection and by PCR mapping using RACE and RLM-RACE methods. Both genes encode polyadenylated mRNAs and are highly expressed in vegetatively growing cells but only weakly expressed in starved cultures. With the inclusion of these two genes, *T.thermophila* is the first organism whose entire complement of known core and linker histones, including replication-dependent and basal variants, has been cloned and sequenced.

INTRODUCTION

The DNA of eukaryotes is packaged into nucleosomes consisting of a core and a linker region. The core is composed of two molecules each of histones H2A, H2B, H3 and H4, around which is wound ~146 bp of DNA in $1^{3}/_{4}$ left-handed superhelical turns. The linker consists of variable amounts of DNA associated with one molecule of histone H1 (reviewed in 1). The histones are basic proteins which have been highly conserved in evolution, albeit to somewhat different degrees (H4=H3 > H2B=H2A > > H1; see 1,2).

In most eukaryotes, each histone class is encoded by a small multigene family. Expression of most major histone genes is coupled to DNA replication. In contrast with these replication variants, some histone genes encode replacement/basal variants that are transcribed throughout the cell cycle and in non-growing cells. The yeast *Saccharomyces cerevisiae* is the only organism in which all of the genes for all of the known histones have been cloned and sequenced and in which the function of histone genes modified *in vitro* can be studied *in vivo* using transformation and gene replacement. Unfortunately, yeast is unusual among eukaryotes in lacking histone H1 and distinct replacement variants of the core histones (3–5).

The histone complement of the ciliated protozoan *Tetrahymena thermophila* has been extensively studied and shown to contain both histone H1 and replacement variants (reviewed in 2,6,7). Recently, methods for mass transformation (8) and for gene replacement (9–11) have been developed for *Tetrahymena* that should enable detailed functional analyses of histone genes in that organism, including analyses of H1 and of minor replacement variants that are lacking in yeast. A prerequisite to such an analysis is the cloning of the histone gene complement of *Tetrahymena*. To date, genes encoding all of the known *Tetrahymena* histones except for two major H2As have been cloned and sequenced (2,12–20). In this report we describe the structure, sequence and expression of the *T.thermophila HTA1* and *HTA2* genes encoding histone H2A1 and H2A2 and summarize the organization of the first complete histone gene complement from an organism containing all of the histone subtypes common to most eukaryotes.

MATERIALS AND METHODS

Cell culture, nuclei isolation, DNA and RNA preparations

Tetrahymena thermophila (strain CU 428, mating type VII) were grown and harvested, nuclei isolated, and DNA and RNA prepared as previously described (16,21–23).

* To whom correspondence should be addressed

[+]Present address: Laboratory of Biochemistry and Metabolism, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD 20892, USA

Southern blotting and determination of oligonucleotide hybridization conditions

Restriction digests and Southern blots were carried out according to standard protocols (24).

Based on codon usage information available for *Tetrahymena* histone genes (15,25 and T. Thatcher and M. A. Gorovsky, unpublished observations) two degenerate oligonucleotides were synthesized corresponding to two regions (amino acid residues 73–80 and 90–96) that are highly conserved in all H2As including those of *Tetrahymena pyriformis* (26).

Probe 1

$$GC\frac{T}{C}GC\frac{T}{C}AA\frac{A}{G}GA\frac{T}{C}AA\frac{T}{G}AA\frac{A}{G}AA\frac{A}{G}ACN$$

$$A_{73}A_{74}K_{75}D_{76}N_{77}K_{78}K_{79}T_{80}$$

Probe 2

$$GC\frac{T}{C}AT\frac{T}{C}AGA\frac{T}{C}AA\frac{T}{C}GA\frac{T}{C}GA\frac{A}{G}GA\frac{A}{G}$$

$$A_{90}I_{91}R_{92}N_{93}D_{94}E_{95}E_{96}$$

Probes were labeled with [γ-^{32}P]ATP and polynucleotide kinase (24) to a specific activity of $>1 \times 10^8$ c.p.m./µg. The optimum conditions for both probes were hybridization at 30°C, followed by washing at 35–40°C. Hybridization solution was 5× SSPE, 1× SPED (8× SPED = 0.8% Ficoll, 0.8% PVP360, 0.8% BSA, 48 mM SDS, 16 mM pyrophosphate, 16 mM EDTA). Washing was done in 2× SSPE and 0.1% SDS.

cDNA cloning

A library of *T.thermophila* log phase cDNAs in λgt11 constructed by D. Shapiro (16) was probed with the two oligonucleotides under the conditions described above. Plaques that hybridized positively to both probes were selected and re-screened. Phage DNA was isolated essentially as described by Maniatis *et al.* (27) with the following modifications. Phage lysate was spun twice to remove contaminating chromosomal DNA. Fifty µl of 1 mg/ml DNase I and 25 µl of 10 mg/ml RNase A were then added to 50 ml of supernatant and incubated for 1 h at 37°C. After addition of 2.92 g of NaCl, 5 g of PEG and incubation on ice for 30 min, phage were pelleted at 12 000 *g* for 10 min at 4°C, and solubilized in 2 ml of a solution containing 10 mM Tris–HCl, pH 8.0, 1 mM EDTA, 0.2% Sarkosyl. After addition of 5 µl of 20 mg/ml Proteinase K and incubation at 55°C for 20 min, phage DNA was isolated by phenol–chloroform extraction and ethanol precipitation.

Although the library was constructed by cloning into the *Eco*RI site of λgt11, the cDNA inserts could not be released with *Eco*RI (subsequently, sequence information indicated that the junctions were disrupted). Therefore, *Sac*I/*Kpn*I-digested insert fragments, which also include 2.1 kb of λgt11 sequences, were subcloned into *Sac*I/*Kpn*I sites on the plasmid vector Bluescript KS(+) (Stratagene) and sequenced using a pair of λgt11 sequencing primers (forward primer = GACTCCTGGAGCCCG; reverse primer = GGTAGCGACCGGCGC) according to standard protocols for double-stranded DNA sequencing (24). A truncated *HTA2* cDNA clone (pXL40) was identified.

Construction of size-selected mini genomic libraries

The 145 bp *Acc*I–*Bal*I fragment from pXL40 was blunt-end ligated into the *Sma*I site of pBluescript to obtain pXL41, from which the insert could be released with *Bam*HI and *Pst*I to obtain a homologous probe to facilitate cloning the remainder of the *Tetrahymena* H2A genes. Genomic DNA was digested with *Bam*HI, *Pst*I, *Eco*RI, *Hin*dIII and *Kpn*I and probed with the 145 bp *HTA2* cDNA sequence (data not shown). *Hin*dIII-digested fragments of 4.1 and 8.0 kb in size were identified and used for construction of a size-selected genomic library. Fifty µg of genomic DNA was digested and run on a 0.5% agarose gel. DNA fragments between 3.5–4.5 and 7–9 kb were eluted as described (28), then ligated into Bluescript KS(+) vector, transformed into DH5α cells and screened (24). This method yielded a 4.1 kb truncated *HTA1* clone (pXL50).

Inverse PCR amplification

Since both *HTA1* and *HTA2* clones obtained were truncated, we cloned the missing parts of each gene by an inverse PCR method (29). When *Bgl*II-digested *Tetrahymena* genomic DNA was probed with the *HTA2* cDNA clone (pXL41), two bands, 2.2 and 2.0 kb in size, were observed. *Bgl*II-digested genomic DNA was ligated at a concentration of 2 µg/ml to achieve 95% monomeric circles according to the formula of Collins and Weissman (30). These ligated circles were then amplified with a single pair of primers,

Oligo 1408	CTTCAAGAATCTGG**G**AATTCTACC
(H2A1 & H2A2	CTTCAAGAATCTGG**A**AATTCTACC)

and

Oligo 2407	GTTCT**C**GAATT**C**GCTGGTAACGC
(H2A1	GTTCT**T**GAATT**G**GCTGGTAACGC)
(H2A2	GTTCT**C**GAATT**G**GCTGGTAACGC),

derived from portions of the coding region of *HTA1* and *HTA2* known to be similar from sequencing the previously obtained partial clones. Mutations (underlined) producing *Eco*RI sites were introduced into the two primer sequences to facilitate subsequent cloning. PCR products were digested with *Eco*RI and cloned into the *Eco*RI site on the pKS(+) vector. Three independent colonies were obtained from clones derived from each of two duplicate reactions, sequenced and compared to detect any mutations occurring during PCR amplification. One completely correct clone for *HTA1* (pXL51) and two correct clones for *HTA2* (pXL42a and pXL42b) were identified.

Assembly of complete *HTA* genes

For construction of pXL53 containing the reconstituted *HTA1* gene encoding H2A1, the 1.9 kb *Bal*I–*Bgl*II fragment containing part of the coding sequences plus the entire 3′ noncoding sequences from pXL51 was ligated into the *Bal*I–*Bam*HI sites of pXL50 with the removal of the small (159 bp) *Bal*I–*Bam*HI fragment from pXL50.

For assembly of pXL46 containing the intact *HTA2* gene encoding H2A2, pXL42b was cut with *Bgl*II and *Xba*I and resealed by blunt end ligation to remove the 3′-half of the gene as well as the *Xba*I site on the vector. The *Xba*I site within the region encoding H2A2 was not cut because this site is dam^s. This gives pXL43. pXL40 and pXL43 were then grown in a dam$^-$

Escherichia coli strain to isolate plasmids in which the *Xba*I sites could be cleaved. Then the 1.36 kb *Xba*I–*Sac*I (blunt) fragment of pXL40 was ligated into the *Xba*I–*Eco*RV sites on pXL43 to obtain pXL44. pXL42b was also digested with *Bgl*II and *Sal*I and blunt end ligated to remove the 5′-half of the *HTA2* gene, resulting in plasmid pXL45. The 1.66 kb *Bal*I–*Xho*I fragment from pXL45 was then ligated into the *Bal*I–*Xho*I sites of pXL44 with the corresponding fragment removed from pXL44. A schematic presentation of the major steps of these procedures is given in Figure 1.

RNase protection assay

For 5′-end analysis, PCR clones pXL51 (for *HTA1*) and pXL42a (for *HTA2*) were cut with *Bgl*II and transcribed with T7 RNA polymerase in the presence of [α-^{32}P]UTP to generate probes with 105 nt of coding sequences plus 153 and 144 nt of 5′-flanking sequences for *HTA1* and *HTA2* respectively.

Two new constructs were made to map the 3′-ends of H2A messages. pXL52 (for *HTA1*) was obtained by ligating the 1.75 kb *Hin*dIII–*Bgl*II fragment of pXL51 into the *Hin*dIII/*Bam*-HI sites of pBluescript KS(+) so that the antisense probe for the 3′ region of *HTA1* messages can be generated by transcribing with T7 RNA polymerase after cleavage with *Hin*dIII. pXL47 (for *HTA2*) was constructed by subcloning the 1.65 kb *Bal*I–*Bgl*II fragment of pXL42a into *Eco*RV/*Bam*HI-digested pBluescript KS(+). The antisense transcript for the 3′ region of *HTA2* messages was synthesized by digesting with *Hin*dIII followed by T7 transcription.

RNase protection was done as described (15). Optimum RNase digestion conditions were found to be at room temperature for 15 min with RNase A at 8 µg/ml and RNase T1 at 180 U/ml.

RACE and RLM-RACE

The 3′-ends of *HTA1* and *HTA2* messages were also mapped by the method of rapid amplification of cDNA ends (RACE; see 31–33). cDNA was made using an oligo(dT) adapter–primer (5′-GACTC-GAGTCGACATCGATTTTTTTTTTTTTTT-3′) and then amplified using the adapter (the above oligo without the 17 Ts) and Oligo 2407 (described above). PCR products were cut with *Eco*RI (in Oligo 2704) and *Sal*I (in the adapter oligo), cloned into *Eco*RI/*Sal*I sites on pBluescript KS(+), and sequenced.

RNA ligase mediated RACE (RLM-RACE) was used to map both 5′- and 3′-ends of H2A1 and H2A2 messages as described (23).

Northern blot analysis

Northern blotting of 1.2% formaldehyde–agarose gels was done according to standard protocols (24). The probe for *HTA1* and *HTA2* messages was derived from the 2 kb *HTA2* insert from pXL46, which recognizes both messages. The probe (phv1) for *HTA3* was derived from the genomic clone of *HTA3* (17) by the removal of the second exon (X. Liu and M. A. Gorovsky, unpublished observations). Quantitation was done using a Molecular Dynamics PhosphorImager™ (Molecular Dynamics, Sunnyvale, CA).

RESULTS

Cloning of *Tetrahymena HTA1* and *HTA2* genes

When a λgt11 cDNA library (16) was screened with the two oligonucleotides based on amino acid residues 73–80 and 90–96 of histone H2A of *T.pyriformis* (see Materials and Methods), five positive clones were obtained which hybridized to both probes. Inserts from these clones were purified, subcloned into the plasmid vector Bluescript KS(+) and sequenced on both strands. When the deduced protein sequences were compared with published *T.pyriformis* H2A protein sequences (26), they all were found to be truncated clones encoding H2A2 but missing the 3′ half of the gene. The longest cDNA (pXL40) contained 26 bp of 5′-untranslated sequence and 354 bp of coding sequence, encoding the first 117 of the 132 amino acid residues. We then constructed size-selected genomic libraries. When a *Hin*dIII-digested genomic Southern blot was probed with the *HTA2* cDNA clone, two bands, 4.1 and 8.0 kb, were observed (data not shown). We succeeded in cloning only the lower band which contained a truncated *HTA1* clone (pXL50), containing ~3.7 kb 5′ upstream sequence and 403 bp of coding sequence, encoding all except the last four amino acid residues. Cloning of the upper band was unsuccessful, probably because the AT-richness of *Tetrahymena* noncoding sequences makes cloning of large fragments extremely difficult (34).

Finally, inverse PCR (29) was used to clone the missing parts of the *HTA1* and *HTA2* genes. Two *Bgl*II fragments, 2.2 and 2.0 kb, respectively, were isolated corresponding to the entire *HTA1* (pXL51) and *HTA2* (pXL42a, and pXL42b in the opposite orientation) clones except for the region between the pair of primers used in the PCR reaction. This complex cloning process is schematically presented in Figure 1.

Assembly of the truncated clones into intact genes

Since all of the clones contained incomplete genes, they were then assembled into intact genes *in vitro* (Fig. 1; for details see Materials and Methods). Briefly, the 3′-half of *HTA1* from the PCR clone (pXL51) was connected to the 5′-half of the genomic clone (pXL50) at the *Bal*I site in the coding region. *HTA2* assembly was somewhat more complicated. The final construct contained the 5′- and 3′-halves from the PCR clone (pXL42b) with the middle part (*Xba*I–*Bal*I) from the cDNA clone (pXL40). The assembled constructs (pXL53 and pXL46 for *HTA1* and *HTA2*, respectively) now contain the full length genes. The coding regions from these two clones were re-sequenced to check for any errors during the assembling procedures. None were found. Also, these reconstructed genes have been shown to be able to function as the only H2A genes in the yeast, *S.cerevisiae* (5). It is worth noting that, while this somewhat baroque scheme for cloning and reconstructing the *HTA1* and *HTA2* genes is the most complicated one we have ever utilized for cloning any *Tetrahymena* gene, all of the more straightforward approaches we tried failed. It is also worth noting that difficulties in cloning particular *Tetrahymena* genomic DNA fragments, especially larger ones, are not uncommon.

The complete sequences of the *HTA1* and *HTA2* genes

The DNA sequences for both genes are presented in Figure 2. The deduced amino acid sequences matched those of the histone H2A proteins from *T.pyriformis* (26) except for two conservative amino acid replacements ($I_{48}\rightarrow V_{48}$ and $I_{138}\rightarrow L_{138}$; note that

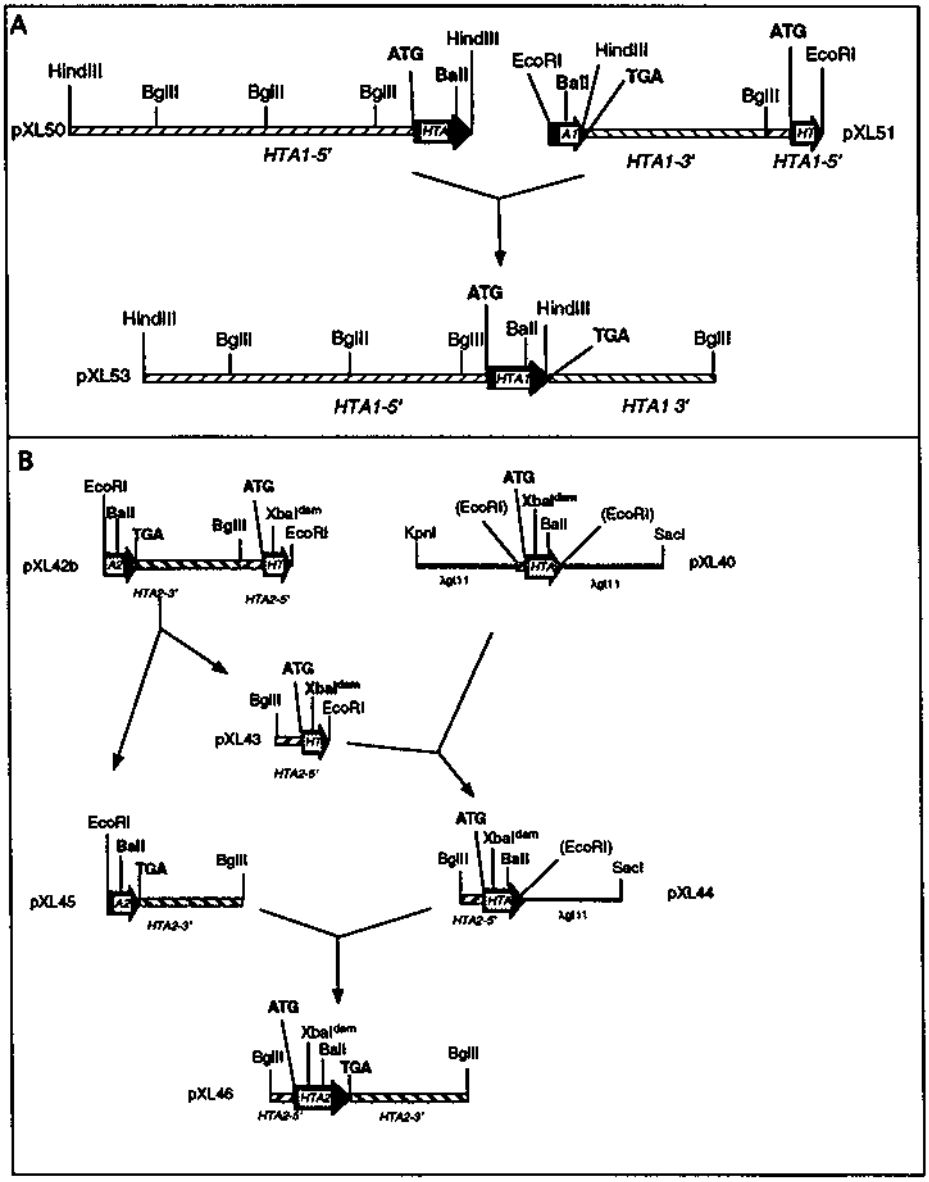

Figure 1. Schematic illustration of the procedures involved in cloning and assembly of *Tetrahymena HTA1* (**A**) and *HTA2* (**B**) genes. See Materials and Methods for details.

locations are codon positions shown in Figure 2, with the initiator methionine codon as number 1, and not amino acid positions in the processed protein from which the initiator methionine is removed) in H2A1, one amino acid replacement ($A_{131}\rightarrow P_{131}$) and one inversion ($T_{128}S_{129}\rightarrow S_{128}T_{129}$) in H2A2. In both *T.pyriformis* and *T.thermophila*, the two H2As are highly similar except for the extreme C-termini, where the sequences are highly diverged and the H2A1 proteins are five amino acids longer than the H2A2 proteins. The differences in H2A1 and H2A2 within each species and the virtual identity of H2A1 and H2A2 between the two *Tetrahymena* species argue that the gene encoding *Tetrahymena* H2A duplicated before the two species diverged and that the two genes do not perform identical functions.

When the nucleotide sequences of the coding regions (except for the 3′-region coding for the divergent carboxyl ends) of the two genes were compared, 27 codons were found to be different, 24 of which were synonymous, encoding identical amino acids either by codon redundancy (5/24) or mostly by codon degeneracy differing only at the third position of the codons (19/24). Three substitutions changed the underlying amino acids.

The histone *HTA2* gene cloned here contains a TAA triplet as the 28th residue. In many ciliates TAA (and by wobble TAG) is used as a codon for glutamine instead of as a stop codon (13,25,35). A glutamine tRNA containing a UUA anticodon has been isolated and the corresponding gene has been cloned from *T.thermophila* (36,37).

Mapping the 5′- and 3′-ends of *HTA1* and *HTA2* transcripts

Using RNase protection, the 5′-ends of *HTA1* and *HTA2* messages were mapped to –61, –58 and –52, –51 respectively; 3′-ends of the two messages were mapped to 890 and 538–539 respectively

Figure 2. Sequence alignment for *HTA1* and *HTA2* (GenBank accession nos L18892 and L18893 respectively) genes and their deduced proteins. DNA sequence differences between the two genes are indicated by bold letters, as are the differences between the two protein sequences. The positions of all the oligonucleotides used in this report within the two genes are indicated by underlines. Oligo 2407 and Probe 1 overlap by 2 nt (GC). Transcription start sites are marked by underlined bold letters (mapped by RNase protection) or italic bold letters (mapped by RLM-RACE), as are the sites for polyadenylation. The*Bgl*II box region of homology is also indicated in bold at the 5′-end of the sequence.

(Fig. 3, and summarized in Fig. 2). We also mapped these ends using the method of RLM-RACE (23). The 5′-ends of *HTA1* were mapped to –60, –56 and –44; 5′-ends of *HTA2* were mapped to –49, –48 and –43. The 3′-ends of *HTA1* were at 878–886, and for *HTA2* they were at 532–534 (23). It seems to be a general feature for most *Tetrahymena* messages to have multiple start and stop sites (23). The small discrepancy between the results obtained by RNase protection and those by RLM-RACE probably reflects the inaccuracies in measuring the real length of protected RNA fragments on sequencing gels when DNA sequencing ladders are used as markers (24). The ends mapped by sequencing the RLM-RACE products are likely to be more accurate.

In contrast with the high DNA sequence homology (91.0%, excluding the divergent C-termini) within the coding regions between the two genes, both 5′ and 3′ flanking region were highly divergent. The 5′ sequences of the *HTA1* and *HTA2* genes are likely to be involved in the control of H2A transcription. However, when these regions are examined for common sequences, little conservation was found (Fig. 2). A canonical TATA box was absent even though these regions are extremely AT-rich. There is one significant exception to this lack of homology. A conserved box (referred to as the *Bgl*II Box), AGATCTG(A/C)AATCTGATTGG(C/A)TAA, is identical in 21 of 23 residues between the two genes and its sequence contains

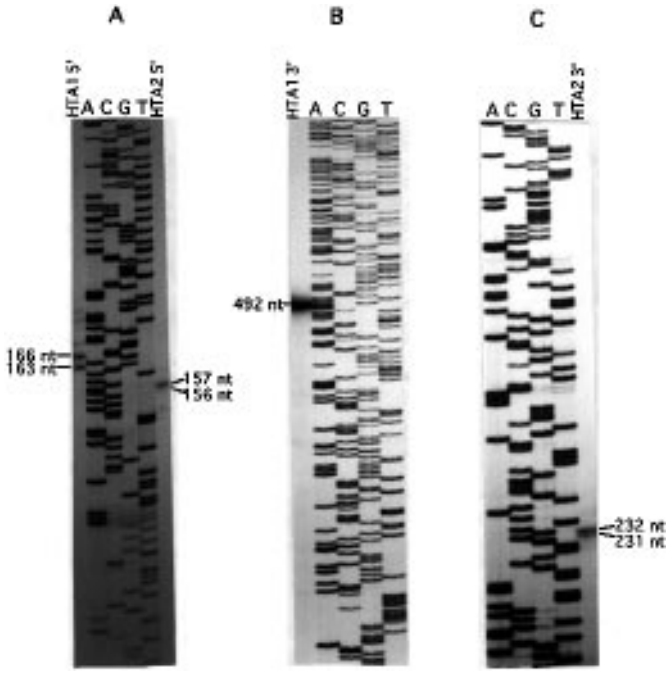

Figure 3. Mapping 5′- and 3′-ends of *HTA1* and *HTA2* messages by RNase protection. Protected fragments were run against a sequencing reaction of the '–40 primer' primed M13mp18 DNA. Sizes of protected fragments are indicated. Position of transcription start/stop sites can then be calculated. (**A**) The 5′-ends of *HTA1* messages are at –61 and –58 (subtracting 105 from 166 and 163, because the probe used should protect 105 nt of coding sequence); the 5′-ends of *HTA2* are at –52 and –51 (subtracting 105 from 157 and 156). (**B**) The 3′-end of *HTA1* is at 890 (492 plus 398, the probe starts 398 nt downstream from ATG). (**C**) The 3′-ends of *HTA2* are at 538 and 539 (231 plus 307 and 232 plus 307, the probe starts 307 nt downstream from ATG).

a previously identified element containing a 'CCAAT' box (TCTGATTGGATA) that is a common feature of the 5′ flanking regions of many *Tetrahymena* genes (23,38). Most interestingly, this box consists of 38.9% GC, in contrast with the low GC content(<20%) typical of the flanking regions of the *Tetrahymena* genome. There is a second incomplete copy of this motif in *HTA1* (GATTGGATA) downstream from the first. The strong conservation of this sequence motif indicates it is an important promoter element whose function merits further investigation *in vivo*, using newly developed methods to transform *Tetrahymena* with cloned genes (8). The only conserved element in the 3′ untranslated region is the discontinuous motif (TGTGTN$_{1–8}$TAAN$_{0–11}$AAGTATT) noted previously in these H2A genes and in the *T.thermophila* H4 genes (23).

The *HTA1* and *HTA2* genes are uninterrupted

HTA3, the gene encoding the H2A.F/Z variant hv1 of *Tetrahymena*, contains two introns (17), and the H2A genes of some fungi (39) and some plants (40) have been reported to contain introns. The sequences of the genomic clones of *HTA1* genes indicate that there are no interruptions within the protein coding regions. The RLM-RACE experiments show that there are no introns in the 5′- and 3′-UTRs, because the sequences of the mapped cDNA clones are identical with those of the genomic clones. For the *HTA2* gene, the cDNA clones (pXL40 plus 5′- and 3′-RLM-RACE clones) match perfectly with the genomic clones where both sequences

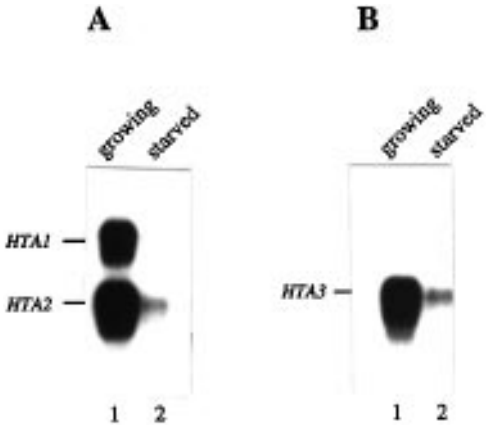

Figure 4. Northern blot showing the expression of the *HTA1*, *HTA2* and *HTA3* genes in both growing and starved cells. Ten µg of total cellular RNA from growing (lane 1) or starved (lane 2) cells were loaded on each lane and hybridized with the probes recognizing either *HTA1* and *HTA2* (**A**) or *HTA3* (**B**).

are available. Within the region between the two primers used in the inverse PCR reaction, the genomic sequences were not cloned. However, the size of the inverse PCR product (1.9 kb) is consistent with that of the genomic fragment on the Southern blot (2.0 kb) taking into account the 78 bp between the two primers. This argues that no intron exists in the *HTA2* gene either.

Expression of *HTA1* and *HTA2* genes

Northern blots of growing (dividing) and starved (non-dividing) *Tetrahymena* RNAs were hybridized with a probe recognizing both *HTA1* and *HTA2*, or with a probe recognizing *HTA3* (Fig. 4). Both the *HTA1* and *HTA2* genes are highly expressed in growing (dividing) cells and low levels of expression are also detectable in starved (non-dividing) cells. PhosphorImager quantitation (Molecular Dynamics, Sunnyvale, CA) of the Northern blots indicates that the *HTA1* gene is expressed at a 55-fold higher level in growing than in starved cells, and the *HTA2* messages are 34 times more abundant in growing versus starved cells. It is not clear whether these large differences in expression between growing and starved cells represent extremely low levels of expression in non-growing cells or indicate that a small fraction of the cells is still growing, perhaps slowly. The *HTA3* messages are clearly present in both growing and starved cells, although a 9-fold higher level of expression is detected in growing cells (see also 41). Thus, the *HTA1* and *HTA2* genes are more likely to be cell cycle regulated while the *HTA3* gene encoding hv1 is a partially replication independent variant, expressed throughout the cell cycle (41) and in non-growing cells.

DISCUSSION

The DNA sequences in Figure 1 establish the amino acid sequences of the H2A proteins in *T.thermophila*. The *HTA1* and *HTA2* genes encode different proteins. H2A1 is five amino acids longer than H2A2 and the two proteins diverge considerably at the C-terminal ends. There also are three internal amino acid differences between the two proteins. Both genes are expressed, as revealed by Northern blot analysis, RNase mapping experiments and by the presence of electrophoretically distinct proteins

on SDS–PAGE (42). Whether the structural differences between the two proteins reflect any functional differences can now be tested using the cloned genes and newly developed methods for creating gene knockouts in *Tetrahymena* (8–11).

When viewed in their entirety, the *Tetrahymena* histone genes illustrate many of the common properties of the histone gene superfamily of eukaryotes. In a few organisms, histone genes are present in large numbers of tandem repeats containing one copy each of the five types. However, in most organisms, each histone class is represented by a small number of genes dispersed at multiple sites in the genome but with some tendency to cluster (1). *Tetrahymena* shows this latter arrangement, but with minimal clustering. Thus, the only histone genes that appear to be physically linked are *HHF2* and *HHT2* encoding an H4 and an H3 which are divergently transcribed and are separated by ~400 bp in *T.thermophila* (12,15) and in other *Tetrahymena* species (38,43,44).

There are three reasons why genes are duplicated and retain function (as opposed to becoming pseudogenes): (i) to provide additional gene copies (dosage repetition); (ii) because the duplicate genes evolve to encode proteins with slightly different functions (variant or isotype repetition); or (iii) because they evolve different patterns of gene expression that are selectively advantageous (regulatory repetition). The *T.thermophila* histone gene superfamily appears to exhibit all of these phenomena. Thus, the two H4 genes, *HHF1* and *HHF2*, encode identical proteins (12,15) and are both expressed during macronuclear S-phase, probably for dosage reasons, but only *HHF2* is expressed during micronuclear S-phase (regulatory repetition; 45). *HHT1* and *HHT2* also encode identical proteins, while *HHT3* encodes a slightly different H3 (7,13). *HHT1*, *HHT2*, *HTA1* and *HTA2* are likely replication variants since they are expressed much more highly in growing than in starved cells (7,46 and this report), while *HHT3* is a replacement/basal or partially replication-dependent variant gene expressed in both growing and starved cells (7,46). The patterns of expression of these H2A and H3 genes in the cell cycle have not yet been analyzed. *HTA3*, the gene encoding hv1 is a replacement/basal variant or partially replication-dependent gene (41 and this report). *HTA3* probably also represents a case of variant repetition. It encodes an H2A.F/Z-type variant and phylogenetic analyses (2,16,17) have indicated that these variants diverged from the major H2A genes early in eukaryotic evolution and have been under even greater evolutionary constraint than the major H2A genes. It is also worth noting that the *HHF1* gene encoding one of the *Tetrahymena* H4s may have properties of both replication-dependent and of replacement/basal variants in as much as its expression is greater in growing than in starved cells (47). Because both the *HTA1* and the *HTA2* genes and the *HHF2* gene (encoding another H4 protein) are present at extremely low levels in starved cells compared with growing cells, it is not clear whether they show a low level of basal expression or just reflect a population of cells that is small and/or cycling slowly.

All of the *Tetrahymena* histone genes analyzed to date are transcribed into polyadenylated messenger RNAs, a property shared with histone genes in fungi (48) and higher plants (see 40, for a recent example). In multicellular animals, only basal/replacement histone genes produce polyadenylated mRNAs. This evolutionary distribution argues that polyadenylation of histones messages is the primitive state and that absence of polyadenylation in the messages of replication-dependent variants in multicellular animal cells probably reflects an evolutionarily recent loss. Also, the stem–loop structure associated with transcription termination of non-polyadenylated histone messages of multicellular animals (see 49 for a recent review) is absent in polyadenylated histone messages. Interestingly, the messages encoding the mammalian H2A.X variant exist in both polyadenylated and non-polyadenylated forms (50,51).

The distribution of introns in *Tetrahymena* histone genes also is illustrative. The *HHO* gene, encoding *Tetrahymena* macronuclear H1, was the first replication variant shown to contain an intron (14). While intron containing genes encoding replication variants are not common, they have since been found in plants (40) and in fungi (39). In multicellular animals, introns have only been found in basal/replacement genes. However, introns are not a universal feature of basally expressed histone genes. In *Tetrahymena*, the basal/replacement gene (*HTA3*) encoding hv1 contains introns (17), while that (*HHT3*) encoding an H3.3 (formerly hv2) does not (7). As in the case of polyadenylation, this phylogenetic distribution makes it likely that introns were present in histone genes before the divergence of replication and basal/replacement variants and before the divergence of animals, plants, protists and fungi and were lost in the evolution of replication-dependent variants in multicellular animals.

Table 1. Summary of histone subtypes and gene number of *Tetrahymena* macro and micronuclear histones

Histone protein	Gene	Nuclear location		No. of genes	References
		Mac	Mic		
Linker histones					
H1	*HHO*	+	–	1	(14)
Mic LH	*MLH*				
α		–	+		
β		–	+	1[a]	(52)
γ		–	+		
δ		–	+		
H2A					
H2A.1	*HTA1*	+	+	1	this report
H2A.2	*HTA2*	+	+	1	this report
hv1	*HTA3*	+	–/+[b]	1	(16,17,53)
H2B					
H2B.1	*HTB1*	+	+	1	(18)
H2B.2	*HTB2*	+	+	1	(18)
H3					
H3.1	*HHT1, HHT2*	+	+	2	(7,13)
hv2	*HHT3*	+	–	1	(7)
H4	*HHF1, HHF2*	+	+	2	(12,15)

[a]The single Mic LH gene encodes a polyprotein precursor that is processed to yield four micronuclear linker histones.
[b]The hv1 protein is absent in micronuclei of vegetative cells but is present in the micronuclei during specific stages of conjugation (53).

In summary, with the studies described here, *T.thermophila* joins *S.cerevisiae* as the only organisms whose entire (known) histone gene complements have been cloned and sequenced (see Table 1). *Tetrahymena thermophila* is unique in having the only completely known histone gene complement containing genes for H1 and minor replacement/basal H2A and H3 variants which are present in most eukaryotes but appear to be lacking in *S.cerevisiae* (4). With the recent development of methods for mass transformation (8) and for gene replacement (9–11) in *Tetrahymena*, it should now be possible to analyze H1, and H2A

and H3 variants *in vivo* using the types of molecular genetic analyses that have been used to study the functions of major core histones in yeast.

ACKNOWLEDGEMENTS

The authors wish to thank Josephine Bowen for critical reading of the manuscript. This work was supported by Public Health Service Grant GM-21793 from the National Institutes of Health.

REFERENCES

1 Van Holde,K.E. (1989) *Chromatin.* Springer-Verlag, New York.
2 Thatcher,T.H. and Gorovsky,M.A. (1994) *Nucleic Acids Res.*, **22**, 174–179.
3 Grunstein,M. (1990) *Annu. Rev. Cell Biol.*, **6**, 643–678.
4 Grunstein,M. (1990) *Trends Genet.*, **6**, 395–400.
5 Liu,X., Bowen,J. and Gorovsky,M.A. (1996) *Mol. Cell. Biol.*, **16**, 2878–2887.
6 Gorovsky,M.A. (1980) *Annu. Rev. Genet.*, **14**, 203–239.
7 Thatcher,T.H., MacGaffey,J., Bowen,J., Horowitz,S., Shapiro,D.L. and Gorovsky,M.A. (1994) *Nucleic Acids Res.*, **22**, 180–186.
8 Gaertig,J. and Gorovsky,M.A. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 9196–9200.
9 Kahn,R.W., Andersen,B.H. and Brunk,C.F. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 9295–9299.
10 Gaertig,J., Thatcher,T.H., Gu,L. and Gorovsky,M.A. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 4549–4553.
11 Gaertig,J., Gu,L., Hai,B. and Gorovsky,M.A. (1994) *Nucleic Acids Res.*, **22**, 5391–5398.
12 Bannon,G.A., Bowen,J.K., Yao,M.-C. and Gorovsky,M.A. (1984) *Nucleic Acids Res.*, **12**, 1961–1975.
13 Horowitz,S. and Gorovsky,M.A. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 2452–2455.
14 Wu,M., Allis,C.D., Richman,R., Cook,R.G. and Gorovsky,M.A. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 8674–8678.
15 Horowitz,S., Bowen,J.K., Bannon,G.A. and Gorovsky,M.A. (1987) *Nucleic Acids Res.*, **15**, 141–160.
16 White,E.M., Shapiro,D.L., Allis,C.D. and Gorovsky,M.A. (1988) *Nucleic Acids Res.*, **16**, 179–198.
17 Van Daal,A., White,E.M., Elgin,S.C.R. and Gorovsky,M.A. (1990) *J. Mol. Evol.*, **30**, 449–455.
18 Nomoto,M., Imai,N., Saiga,H., Matsui,T. and Mita,T. (1987) *Nucleic Acids Res.*, **15**, 5681–5697.
19 Nomoto,M., Matsui,T., Saiga,H. and Mita,T. (1988) *Oxford Surveys on Eukaryotic Genes*, **5**, 251–278.
20 Gorovsky,M.A. (1986) In Gall,J.G. (ed.), *The Molecular Biology of Ciliated Protozoa.* Academic Press, Inc., Orlando, pp. 227–261.
21 Gorovsky,M.A., Yao,M.-C., Keevert,J.B. and Pleger,G.L. (1975) *Methods Cell Biol.*, **IX**, 311–327.
22 Pederson,D.S., Shupe,K. and Gorovsky,M.A. (1984) *Nucleic Acids Res.*, **12**, 8489–8507.
23 Liu,X. and Gorovsky,M.A. (1993) *Nucleic Acids Res.*, **21**, 4954–4960.
24 Ausubel,F.M., Brent,R., Kingston,R.E., Moore,D.D., Seidman,J.G., Smith,J.A. and Struhl,K. (1988) *Current Protocols in Molecular Biology.* Wiley Interscience, New York.
25 Martindale,D.W. (1989) *J. Protozool.*, **36**, 29–34.
26 Fusauchi,Y. and Iwai,K. (1983) *J. Biochem.*, **93**, 1487–1497.
27 Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
28 Girvitz,S.C., Bacchetti,S., Rainbow,A.J. and Graham,F.L. (1980) *Anal. Biochem.*, **106**, 492–496.
29 Ochman,H., Medhora,M.M., Garza,D. and Hartl,D.L. (1990) In Innis,M.A., Gelfand,D.H., Sninsky,J.J. and White,T.J. (eds), *PCR Protocols: A Guide to Methods and Applications.* Academic Press, San Diego, CA, pp. 219–227.
30 Collins,F.S. and Weissman,S.M. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 6812–6816.
31 Frohman,M.A., Dush,M.K. and Martin,G.R. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 8998–9002.
32 Frohman,M.A. (1990) In Innis,M.A., Gelfand,D.H., Sninsky,J.J. and White,T.J. (eds), *PCR Protocols: A Guide to Methods and Applications.* Academic Press, San Diego, CA, pp. 28–38.
33 Frohman,M.A. (1993) *Methods Enzymol.*, **218**, 340–356.
34 Anderson,B. and McDonald,G. (1993) *Anal. Biochem.*, **211**, 325–327.
35 Andreasen,P., Dreisig,H.D. and Kristiansen,K. (1987) *Biochem. J.*, **244**, 331–335.
36 Hanyu,N., Kuchino,Y. and Nishimura,S. (1986) *EMBO J.*, **5**, 1307–1311.
37 Kuchino,Y., Hanyu,N., Tashiro,G. and Nishimura,S. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 4758–4762.
38 Brunk,C.F. and Sadler,L.A. (1990) *Nucleic Acids Res.*, **18**, 323–329.
39 May,G.S. and Morris,N.R. (1987) *Gene*, **58**, 59–66.
40 Sundås,A., Tandre,K., Kvarnheden,A. and Engström,P. (1993) *Plant Mol. Biol.*, **21**, 595–605.
41 White,E.M. and Gorovsky,M.A. (1988) *Mol. Cell. Biol.*, **8**, 4780–4786.
42 Allis,C.D., Glover,C.V.D., Bowen,J.K. and Gorovsky,M.A. (1980) *Cell*, **20**, 609–617.
43 Brunk,C.F., Kahn,R.W. and Sadler,L.A. (1990) *J. Mol. Evol.*, **30**, 290–297.
44 Sadler,L.A. and Brunk,C.F. (1992) *Mol. Biol. Evol.*, **9**, 70–84.
45 Yu,S.-M., Horowitz,S. and Gorovsky,M.A. (1987) *Genes Dev.*, **1**, 683–692.
46 Bannon,G.A., Calzone,F.J., Bowen,J.K., Allis,C.D. and Gorovsky,M.A. (1983) *Nucleic Acids Res.*, **11**, 3903–3917.
47 Yu,S.-M. and Gorovsky,M.A. (1986) *Nucleic Acids Res.*, **14**, 7597–7615.
48 Osley,M.A. (1991) *Annu. Rev. Biochem.*, **60**, 827–861.
49 Marzluff,W.F. (1992) *Gene Expression*, **2**, 93–97.
50 Mannironi,C., Bonner,W.M. and Hatch,C.L. (1989) *Nucleic Acids Res.*, **17**, 9113–9126.
51 Nagata,T., Kato,T., Morita,T., Nozaki,M., Kubota,H., Yagi,H. and Matsushiro,A. (1991) *Nucleic Acids Res.*, **19**, 2441–2447.
52 Wu,M., Allis,C.D., Sweet,M.T., Cook,R.G., Thatcher,T.H. and Gorovsky,M.A. (1994) *Mol. Cell. Biol.*, **14**, 10–20.
53 Stargell,L.A., Bowen,J., Dadd,C.A., Dedon,P.C., Davis,M., Cook,R.G., Allis,C.D. and Gorovsky,M.A. (1993) *Genes Dev.*, **7**, 2641–2651.