Published online 3 October 2011 *Nucleic Acids Research, 2012, Vol. 40, No. 3* ***1267–1281***
doi:10.1093/nar/gkr792

Experimental identification and analysis of macronuclear non-coding RNAs from the ciliate *Tetrahymena thermophila*

Kasper L. Andersen and Henrik Nielsen*

Department of Cellular and Molecular Medicine and Center for Non-coding RNA in Technology and Health, The Panum Institute, University of Copenhagen, 3 Blegdamsvej, DK-2200N, Denmark

Received May 16, 2011; Revised September 2, 2011; Accepted September 11, 2011

ABSTRACT

The ciliate *Tetrahymena thermophila* is an important eukaryotic model organism that has been used in pioneering studies of general phenomena, such as ribozymes, telomeres, chromatin structure and genome reorganization. Recent work has shown that *Tetrahymena* has many classes of small RNA molecules expressed during vegetative growth or sexual reorganization. In order to get an overview of medium-sized (40–500 nt) RNAs expressed from the *Tetrahymena* genome, we created a size-fractionated cDNA library from macronuclear RNA and analyzed 80 RNAs, most of which were previously unknown. The most abundant class was small nucleolar RNAs (snoRNAs), many of which are formed by an unusual maturation pathway. The modifications guided by the snoRNAs were analyzed bioinformatically and experimentally and many *Tetrahymena*-specific modifications were found, including several in an essential, but not conserved domain of ribosomal RNA. Of particular interest, we detected two methylations in the 5′-end of U6 small nuclear RNA (snRNA) that has an unusual structure in *Tetrahymena*. Further, we found a candidate for the first U8 outside metazoans, and an unusual U14 candidate. In addition, a number of candidates for new non-coding RNAs were characterized by expression analysis at different growth conditions.

INTRODUCTION

Non-coding RNAs (ncRNA) represent a large proportion of the transcribed sequences in eukaryotes (1,2). Although functional studies are lagging behind the sequencing efforts, ncRNAs have been associated with aspects of many fundamental cellular processes, and hence also implicated in a wide range of diseases (3). NcRNA is a heterogeneous group of transcripts ranging in size from 20 to 24 nt micro RNAs (miRNAs) (4) to several thousands of nucleotides, such as the Xist RNA (5). Their overall sequence conservation, even within the same class of ncRNAs, can be very low which makes them hard to convincingly predict and map by *in silico* methods.

The most abundant classes of ncRNAs beside the ribosomal RNAs are small RNAs well below the size of mRNAs. The smallest, 20–25 nt ncRNAs, such as miRNA and siRNAs have been implicated in regulatory processes including transcriptional and post-transcriptional gene silencing (4). Another prominent large class of ncRNAs is small nucleolar RNAs (snoRNAs). Based on specific sequence motifs, snoRNAs are divided into two subclasses, box C/D and box H/ACA, each associated with a specific set of proteins and known to guide 2′-*O*-ribose methylations (N_ms) and pseudouridylations (Ψs), respectively (6,7). The majority of snoRNAs guide site-specific modification of nucleotides in the nascent pre-rRNA through base-pairing interactions. In addition, a few snoRNAs are important in the processing that lead to the excision of ribosomal spacers from pre-rRNA (8). In the recent years the complexity of the snoRNA biology has increased and snoRNAs targeting snRNAs, tRNAs and mRNAs have been described along with orphan snoRNAs with no apparent target among the main groups of RNA. Furthermore, some snoRNA genes have been shown to be differentially expressed depending on tissue type and to be among the relatively few genes that are parentally imprinted (9,10). Recently, miRNA and snoRNA biology converged when it was demonstrated in humans and in the protozoan *Giardia lamblia*, that miRNAs controlling mRNA translation could be produced from snoRNAs in a dicer-dependent manner (11,12). The genomic organization of snoRNAs is highly variable. SnoRNA genes can be found both as independent units transcribed from their own promoter or as intron encoded genes. The latter, have been shown to be processed to mature snoRNAs by either a splicing-dependent or independent

*To whom correspondence should be addressed. Tel: +45 35 32 77 63; Fax: +45 35 32 77 32; Email: hamra@sund.ku.dk

pathway. Furthermore, both intronic and independent snoRNA genes can be found as individual genes or as part of a cluster. In many cases, clustered snoRNAs have been shown to be transcribed as a single unit and subsequently processed by ribonucleases to the mature gene products. Typically, organisms have more than one type of genomic organization of their snoRNA genes. However, most show a strong preference toward a certain type(s) of organization (13,14).

Considering the structural and functional diversity of ncRNA, it is of interest to explore these RNAs in a variety of organisms. As examples, recent reports have described small ncRNAs from *Arabidopsis* and the silkworm, *Bombyx mori* (15,16). Our emphasis is on the ciliate *Tetrahymena thermophila* that has been used as a model in RNA research for decades and has pioneered the ribozyme (17) and telomerase (18) fields. As most ciliates, *Tetrahymena* has two structurally and functionally distinct nuclei. The diploid micronucleus (MIC) is the germ line transmitting genetic information during sexual reproduction. However, the MIC is transcriptionally inactive during most of the lifespan of the cell and through vegetative cell divisions. The genome in the somatic macronucleus (MAC) is derived from the MIC genome during sexual reorganization in a process that involves DNA rearrangements. The genome is fragmented, as well as amplified from 5 MIC chromosomes to an estimated 250–300 MAC chromosomes and a ploidy of ~45 (19). Furthermore, 10–20% of the corresponding MIC genome is eliminated, preferentially by deletion of centromeres, foreign DNA, transposable elements and other repetitive elements (20,21). Since most classes of ncRNAs constitute a specific challenge due to their diversity and relatively poor sequence conservation, many genome projects are underrepresented in ncRNA annotation. In *Tetrahymena*, the smallest 20–25 nt long ncRNAs have been demonstrated to have a high complexity (22,23) and remarkably, it has been shown that a class of 27–30 nt small ncRNAs termed scnRNAs are involved in guiding the precise excision of DNA during MAC formation (24,25). In addition, the spliceosomal small nuclear RNAs (snRNAs) U1–U6 have been described (20,26,27), as well as a few snoRNAs (28,29). However, no systematic effort for describing ncRNAs other than the 20–40 nt and the most prominent rRNA, tRNA and snRNAs has been carried out.

Here, we generated and analyzed a cDNA library of 40–500 nt ncRNAs from the *Tetrahymena* MAC. We identified 80 ncRNAs, of which 64 (80%) were previously unknown and studied their genomic organization. The majority could be placed into one of the known classes of ncRNAs, predominantly snRNAs and box C/D and H/ACA snoRNAs. Among the snoRNAs, we identified the first U8 candidate outside metazoans and a *Tetrahymena* U14 candidate lacking one of two canonical U14 elements. The box C/D snoRNAs showed signs of an alternative maturation in that most did not include a small terminal stem known from box C/Ds in other organisms. Instead, most *Tetrahymena* box C/D snoRNAs had the potential to form extensive external and internal basepairing. Further, we determined possible targets for the identified snoRNAs and observed a *Tetrahymena*-specific pattern, which we tested experimentally. Among the predicted and experimentally verified modifications were two methylations in the 5′ stem-loop of U6 snRNA.

MATERIAL AND METHODS

Tetrahymena cell culture

SB210 and B1868VII isolates were maintained at 30°C in NEFF medium (0.25% proteose peptone, 0.25% yeast extract, 0.5% glucose, 30 μM $FeCl_3$). Heat shock and cold shock was induced by harvest of the cells in log phase growth ($1.5–3 \times 10^5$ cells/ml) followed by resuspension of the cell pellet in NEFF medium pre-warmed/cooled at the desired temperature. Subsequently, cells were kept shaking at the designated temperature for 2 h. Starvation was obtained by harvest of log phase cells, followed by two washes in 10 mM Tris–HCl (pH 7.5), resuspension of the washed cells in 10 mM Tris–HCl and continued growth at 30°C. Stationary phase cells were obtained by continuous growth in NEFF media to maximum density $\sim 1–2 \times 10^6$ cells/ml. Harvest of stationary phase cells was done after 20–24 h at maximum density.

Isolation of nuclei and RNA preparation

For extraction of nuclear RNA, cells (0.5–1 l of culture) in early log phase were harvested and resupended in 70 ml ice-cold Tris-Magnesium-Sucrose buffer (10 mM Tris pH 7.5, 10 mM $MgCl_2$, 3 mM $CaCl_2$, 250 mM sucrose). Lysis was obtained by adding 200 μl NP40 and shaking vigorously for 10 min at 0°C. A quantity of 63 g sucrose was added and shaking continued for 1 h. The lysate was spun at 6500 rpm in a HB4 swing rotor (Sorvall) and the supernatant removed. Nuclei were washed once in 5 ml TMS buffer before they were resuspended in 1 ml proteinase K buffer (10 mM Tris pH 7.3, 100 mM NaCl, 0.5% SDS) supplemented with 10 μl proteinase K (10 mg/ml) and incubated at 30°C for 30 min. The reaction was precipitated with 2.5 × volume 96% ethanol and DNase treated for 30 min at room temperature in DNase buffer (10 mM Tris pH 7.3, 3 mM $MgCl_{2,}$ 50 mM KCl) supplemented with 5 μl DNase I (Invitrogen). The RNA was phenol extracted and the concentration determined by spectrophotometry. Whole cell RNA was prepared by the TRIzol (Gibco BRL) method.

Library construction

The cDNA libraries were constructed from RNA by the RNomic approach essentially as described by Hüttenhofer *et al.* (30). Briefly, ~5 μg of isolated macronuclear RNA was separated on a 5% polyacrylamide gel (50% urea, 1 × TBE) and RNA in the size range between 40 and 500 nt excised, eluted from the gel, and tailed with a poly(C)-tail using *Escherichia coli* poly(A) polymerase (Invitrogen) in C-tailing buffer (50 mM Tris–HCl pH 8.0, 200 mM NaCl, 10 mM $MgCl_2$, 2 mM $MnCl_2$, 0.4 mM EDTA, 1 mM DTT, 2 mM CTP and trace amounts of [α-P^{32}]CTP). Then poly(C)-tailed RNA was reverse transcribed using a 3′ poly(G_{15})-primer including restriction sites. For second strand cDNA synthesis, followed by 5′-end double strand DNA adapter ligation, the SuperScript Choice System for cDNA Synthesis (Invitrogen) was applied. The cDNA was then amplified by 20 cycles of PCR, digested with XbaI and EcoRI and ligated into the pUC19 vector. The resulting plasmids were introduced into competent DH5α cells (Invitrogen) according to the manufacturer's instructions. DNA from positive colonies (determined by β-gal activity) was PCR amplified directly from the colony using plasmid targeted oligos and PCR products were analyzed by 2% agarose gel electrophoresis. PCR products with a length indicating the presence of an insert were sequenced by Sanger sequencing. Additional snoRNAs were found by scanning the surrounding genomic sequence of cloned ncRNAs for snoRNA sequence motifs. The expression of a few snoRNA candidates identified in this way was confirmed by primer extension analysis. Sequences and genomic localization of the clones were confirmed by BLAST/BLAT against the *Tetrahymena* genome at http://www.ciliate.org (20). All unique sequences were folded using the RNA mfold server at http://mfold.bioinfo.rpi.edu/ (31) and selected RNAs were compared to the Rfam 10.0 models (cutoff score 0) using the Infernal software 1.02. Some sequences were elongated manually by using genomic sequence to include canonical box C/D and H/ACA structures (32). A subset of the RNAs was analyzed by primer extension (as described below) giving information on the mature ncRNA 5′-ends and expression. Primers used for cloning, sequencing, northern blotting and primer extension are listed in Supplementary Table S1.

Northern blot

For northern blot analysis, 5–10 µg whole cell RNA was resolved on a 5% denaturing (50 % urea) polyacrylamide gel. RNA was visualized by SYBR Gold (Invitrogen) staining and transferred to a Hybond-N$^+$ membrane (GE). End-labeled oligos were hybridized to the membrane-bound RNA at 42–50°C in 6× SSC, 0.1% SDS, 4× Denhardts solution [0.08% (w/v) BSA, 0.08% (w/v) polyvinylpyrrolidone and 0.08% (w/v) ficoll] followed by washes in 3× SSC/2× SSC, 0.1% SDS. For detection, membranes were exposed to a phosphor imager screen and scanned with a Typhoon 8600 scanner.

Prediction and mapping of modified nucleotides and RNA 5′-ends by primer extension

Targets of the identified box C/D snoRNAs were predicted using the SnoScan web service (http://lowelab.ucsc.edu/snoscan/) (33) against a local database of *Tetrahymena* rRNA, tRNA, snRNA, telomerase and, SRP RNA species. Targets for box H/ACA snoRNAs were predicted by local implementation of snoGPS (34) and searching against the above mentioned local database. For further specification on the search parameters, output filtering and database use, see Supplementary Tables S2 and S3. The 2′-OH-ribose methylations were detected with primer extension analysis by limiting the dNTP concentration from 1 to 0.04 and 0.004 mM in the reaction mixture (35). The Ψ's were modified by *n*-cyclohexyl-*N*′-β-(4-methylmorpholinium) ethylcarbodiimeide *p*-tosylate (CMC) and subsequently detected by primer extension (36). Additionally, primer extension reactions were used to confirm expression of cloned ncRNAs and determine their 5′-ends. Primer extension was carried out on 1–5 µg whole cell or nuclear RNA at 42°C in 10 µl reactions with 1 pmol of the appropriate 5′–end-labeled primer and M-MuLV H$^-$ reverse transcriptase (Fermentas). The resulting DNA was analyzed on a (6, 8 or 10%) denaturing (50% urea) polyacrylamide gel next to the appropriate direct RNA sequencing reaction primed by 1 pmol of the oligo used for detection of the modification. For 5′-end determination, the primer extension reaction was electrophoresed next to a known but unrelated sequencing reaction.

RESULTS

Library construction and classification of ncRNAs

We generated a cDNA library of small RNAs prepared from the transcriptionally active macronucleus of exponentially growing *Tetrahymena* cells. The smallest ncRNAs have already been addressed in *Tetrahymena* (22,23) so we concentrated on medium-sized ncRNAs in the range of 40–500 nt which excludes miRNA-sized RNAs, most mRNAs and the larger rRNA species. Further, since the macronucleus can be efficiently separated from the remainder of the *Tetrahymena* cell, we simultaneously enriched against cytoplasmic mature rRNA, tRNA and mRNAs species. Thus, we did not need, as seen in related studies, any experimental elimination step or pre-sequencing screen to avoid massive representation of these highly abundant RNAs in the cDNA library.

We searched the 600 obtained cDNA sequences against the *Tetrahymena* genome database (http://www.ciliate.org/) to verify the sequence, determine the genomic localization and detect overlaps with known ncRNAs or open reading frames. Of a total of 92 compiled non-overlapping putative ncRNAs, 82 mapped to the assembled macronuclear genome and 10 to the unassembled 'trace sequences' (Figure 1). By searching against GenBank, all but one of the trace sequences in the cDNA library could be identified as rRNA or EST/cDNA derivatives. None of the trace sequence matches were included in the following analysis. Similarly, six sequences mapped to already known snRNAs and were not analyzed further.

In order to further classify the macronuclear matching sequences, they were scanned for known snoRNA sequence motifs [C (RUGAUGA), D (CUGA), H (ANANNA) and ACA] and their secondary structure was determined by use of the mfold webserver (31). Some sequences containing box D, D′, C′, H or ACA sequence motifs were suspected to be 5′ truncated and were extended manually by using genomic sequence to include canonical box C/D and H/ACA structures (32) when applicable. A subset of the RNAs was analyzed by primer extension to give information on mature ncRNA 5′-ends and verify the added 5′ sequence extensions and/or by northern blotting (see below) to obtain information on the full-length size of the RNAs. Based on the above, we concluded that the library include 80 ncRNAs, of which 64 are newly described RNAs

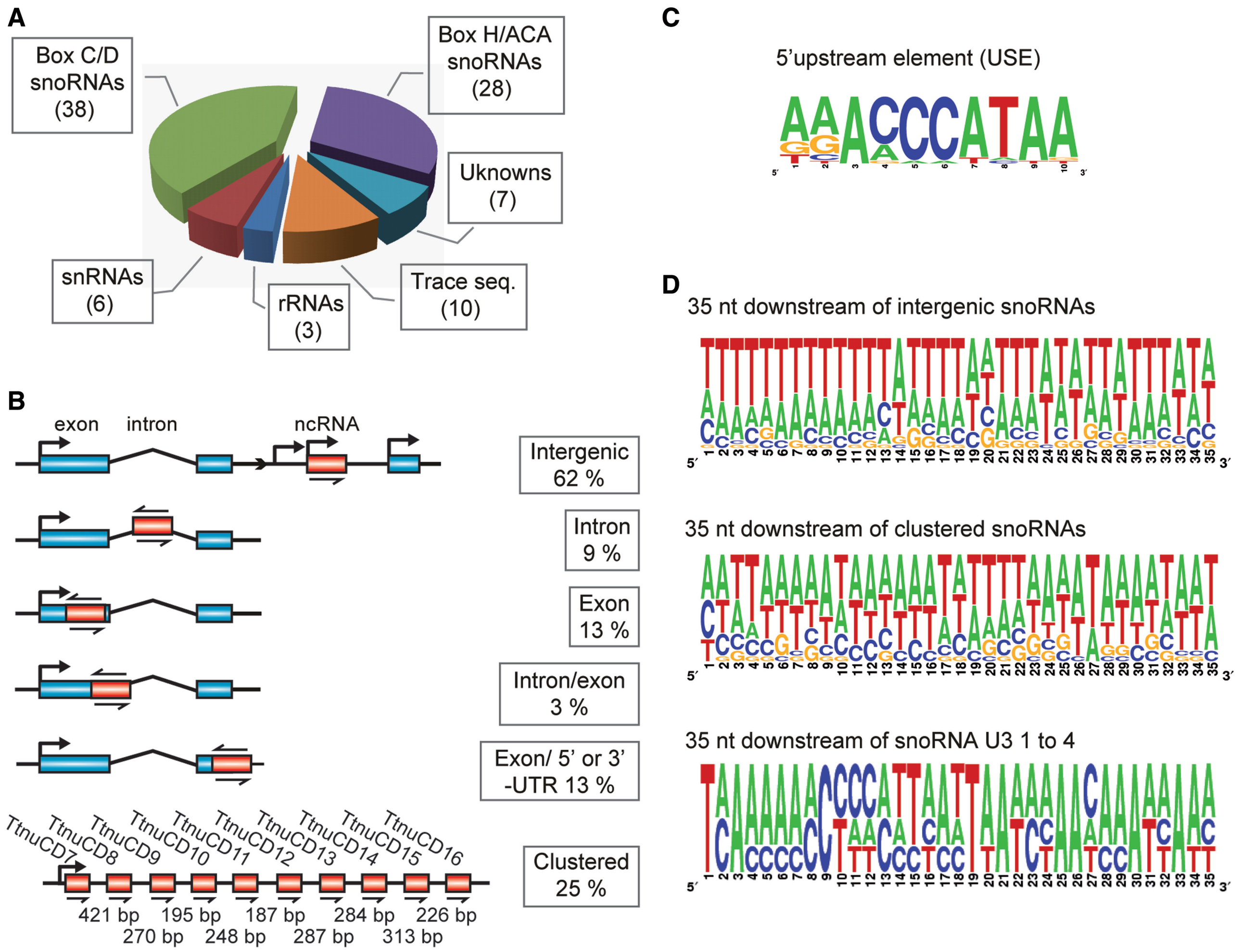

Figure 1. New RNAs and their genomic organization. (**A**) Distribution of cDNA library sequences into RNA classes. The number of RNAs in each class is given in parentheses. (**B**) Schematic outline of genomic localization and organization of ncRNAs in *Tetrahymena*. NcRNAs are shown as red boxes, protein-coding gene exons as blue boxes connected by broken lines representing introns. Orientation of ncRNA and protein-coding genes, as well as transcription starts are indicated with arrows. The largest cluster of snoRNAs is shown at the bottom. Names of snoRNAs and distances between genes are given. (**C**) Frequency logo of the upstream sequence element (USE) found 54–179 bp upstream of intergenic snoRNAs. (**D**) Frequency logo of 35 nt downstream of intergenic monocistronic snoRNAs, clustered snoRNAs and snoRNA U3 genes. The frequency logos were created using WebLogo (http://weblogo.berkeley.edu/) (64).

and 16 (rRNA, snRNAs and some snoRNAs) were previously known. The majority of the RNAs could be assigned to one of the two classes of snoRNAs namely 38 box C/D and 28 box H/ACA (TtnuCDs and TtnuHACAs, respectively) (Figure 1 and Supplementary Tables S2 and S3).

The remaining seven sequences could not be assigned to any known class of RNAs. They were all searched against several ncRNA databases, Rfam (http://rfam.sanger.ac.uk/), Non-coding RNA database (http://biobases.ibch.poznan.pl/ncRNA/) and Functional RNA database (http://www.ncrna.org/frnadb/) but returned no significant matches. These were all classified as unknowns (TtnuUkn's) and could potentially be novel ncRNAs. However, three of these mapped to the *Tetrahymena* genome in introns or exons of protein-coding genes in the mRNA sense direction and could be degradation products of mRNAs rather than functional transcripts (Figure 1, Supplementary Table S4 and Supplementary Figure S1).

Structure of *Tetrahymena* ncRNAs

When folded by mfold, most box H/ACA snoRNA candidates with typical lengths ranging from 124 to 136 nt adopted the characteristic two stem structures as known from, e.g. vertebrate box H/ACA snoRNAs. The two stems contain each an internal loop, the pseudouridylation pocket and are separated by a single stranded region containing the box H (ANANNA) motif (29,32). However, some of the TtnuHACAs deviated from the canonical box H/ACA snoRNA in having AUA (6 in total), UAA (1), UCA (1) and AAA (1) as alternative sequence in place of ACA sequence box towards the 3′-end (Supplementary Table S3).

The *Tetrahymena* box C/D snoRNA candidates also correspond to the canonical vertebrate box C/D structure containing the hallmark 3′ box D (CUGA) and a 5′ box C (RUGAUGA), albeit, without the preference for a purine at the 5′ position. In nearly all TtnuCDs, additional internal boxes C′ and D′ could be identified with a slight deviation from the box C and D sequence motif as typically seen (32). However, the canonical 4–6 bp terminal stem that is joining the 5′- and the 3′-end of box C/D snoRNAs, was often very short or not present at all in TtnuCDs. For several box C/D snoRNAs a potential for an alternative, longer stem of 5–25 bp was present just adjacent to or a few nucleotides away from the 5′- and 3′-ends of the mature transcript (Supplementary Figure S2). In addition, some of the TtnuCDs had a potential for forming extensive internal base-pairing (data not shown).

Genomic organization of ncRNA in *Tetrahymena*

The alignment of the identified ncRNAs to the assembled *Tetrahymena* genome allowed for an analysis of the genomic organization of ncRNAs in *Tetrahymena*. NcRNAs were found in different genomic contexts, mostly in intergenic regions, but also within protein-coding genes overlapping exons, introns or borders between these (Figure 1B). It should be noted that some of the annotated protein-coding genes hosting ncRNA genes are unrealistically small and in some cases occupy only slightly more sequence than the ncRNA. We suspect that the higher GC-content of ncRNAs relative to surrounding sequence has given rise to several false protein-coding gene predictions. Nevertheless, ncRNAs was found in introns of *bona fide* protein-coding genes, as well as in hypothetical protein-coding genes with attributes indicating a reliable gene prediction. The two examples of *bona fide* protein-coding genes hosting the intronic snoRNAs TtnuCD30 and TtnuHACA17 were not proteins involved in ribosomal functions as typical for vertebrate genes hosting snoRNAs. Instead, they were alternative housekeeping genes, namely a dynein chain encoding gene and a preprotein translocase encoding gene, respectively (Supplementary Table S4).

SnoRNAs mapping to intergenic regions were found both as independent units and as part of clusters. We identified five regions where snoRNA genes were clustered and separated by 169–421 nt. In Figure 1B, the largest cluster is schematically outlined with snoRNA genes and gene space distances noted. Curiously, all but one of the snoRNAs found in clusters could be assigned to the box C/D subgroup.

Upstream and downstream sequence elements

We next compared 200 bp of upstream and downstream sequences in clustered and isolated, intergenic snoRNAs. The genome of *Tetrahymena* is very A/T rich (78%) so a stretch of three consecutive C's upstream of most intergenic snoRNAs was readily identified. When these were aligned, a putative promoter upstream sequence element (USE) with the consensus sequence 5′-AAACCC ATAA was noted. A frequency plot of the conserved USE is depicted in Figure 1C. The USE was identified for 22 of 27 intergenic snoRNAs and was situated at a distance of 53–125 bp upstream from the 5′-end of the mature box C/D snoRNAs and 53–179 bp upstream from box H/ACA snoRNAs. The rather large difference in the position of the putative promoter element could indicate that some snoRNAs are initially synthesized as precursors. The USE was found less frequently upstream of clustered snoRNAs and also showed a tendency towards deviating from the consensus sequence (e.g. 5′-TAAGCC ATAA in TtnuCD35 and 5′-TTTCCCAATA in TtnuCD12). Nevertheless, for 11 out of 19 clustered snoRNAs we could identify a putative USE and these included snoRNAs that were not first in a cluster. The USE was not found in the *TtnuUkn* genes except for TtnuUkn4 that had a somewhat deviating USE upstream of both the mature 5′-ends deduced from expression analysis.

Alignment of 35 bp downstream of intergenic monocistronic snoRNAs revealed stretches of 3–13 consecutive T's (Figure 1D). In contrast, the 3′ flanking sequence in clustered snoRNAs had no sequence feature and reflected the general high AT-richness of the genome. A distinct downstream element was previously found in *Tetrahymena* snoRNA *U3* genes (26) and is included in Figure 1D for comparison.

Candidates for the first U8 outside metazoans and an unusual U14

The typical TtnuCD snoRNA is 55–77 nt long. However, five TtnuCDs are considerably longer: TtnuCD7 (108 nt), TtnuCD10 (117 nt), TtnuCD25 (89 nt), TtnuCD26 (98) and TtnuCD32 (98 nt). By comparing them with the Rfam database using Infernal, we suggest that four of the long TtnuCDs belong to the snoZ7/snoR77 (TtnuCD7), U15 (TtnuCD10), U14 (TtnuCD25) and U8 (TtnuCD26) families, respectively. Using this analysis TtnuCD32 could not be assigned to a specific snoRNA family. The affiliations of TtnuCD7 and TtnuCD10 to the snoZ7 and U15 families were further supported by conservation of their respective predicted targets at 17S rRNA U572 (plant snoR77Y targeting U580) and 26S rRNA A2274 (human U15 targeting A3764) (Rfam: http://rfam.sanger.ac.uk/).

The sequence and secondary structure of TtnuCD26 supported its placement in the U8 family (Figure 2A). Also, the 5′-end of TtnuCD26 was complementary to the 5′-end of 26S rRNA similar to an interaction known from *Drosophila*, *Xenopus* and human (Figure 2A). Interestingly, a hairpin (hp2 in the human U8) conserved from Cnidarian to human U8 was absent in the *Tetrahymena* sequence (Figure 2A). Furthermore, although the stem in hp3 in TtnuCD26 was conserved in comparison with human U8 hp4, the LSm binding motif in the loop of this hairpin (37) was not conserved. In contrast, the loop sequence of human U8 hp5 shared with *Caenorhabditis elegans* and *Xenopus laevis* (38) is conserved in *Tetrahymena* TtnuCD26 but was localized in a small hairpin (hp5) toward the very 3′-end of TtnuCD26 (Figure 2A).

U14 is an unusual box C/D snoRNA in yeast and humans since it displays a dual function and both guides methylation, as well as being involved in rRNA processing events. U14 has a conserved A-domain that was shown to be essential for U14-mediated rRNA processing function and cell survival in yeast (39). An A-domain sequence could also be identified in TtnuCD25 (Figure 2B) and its

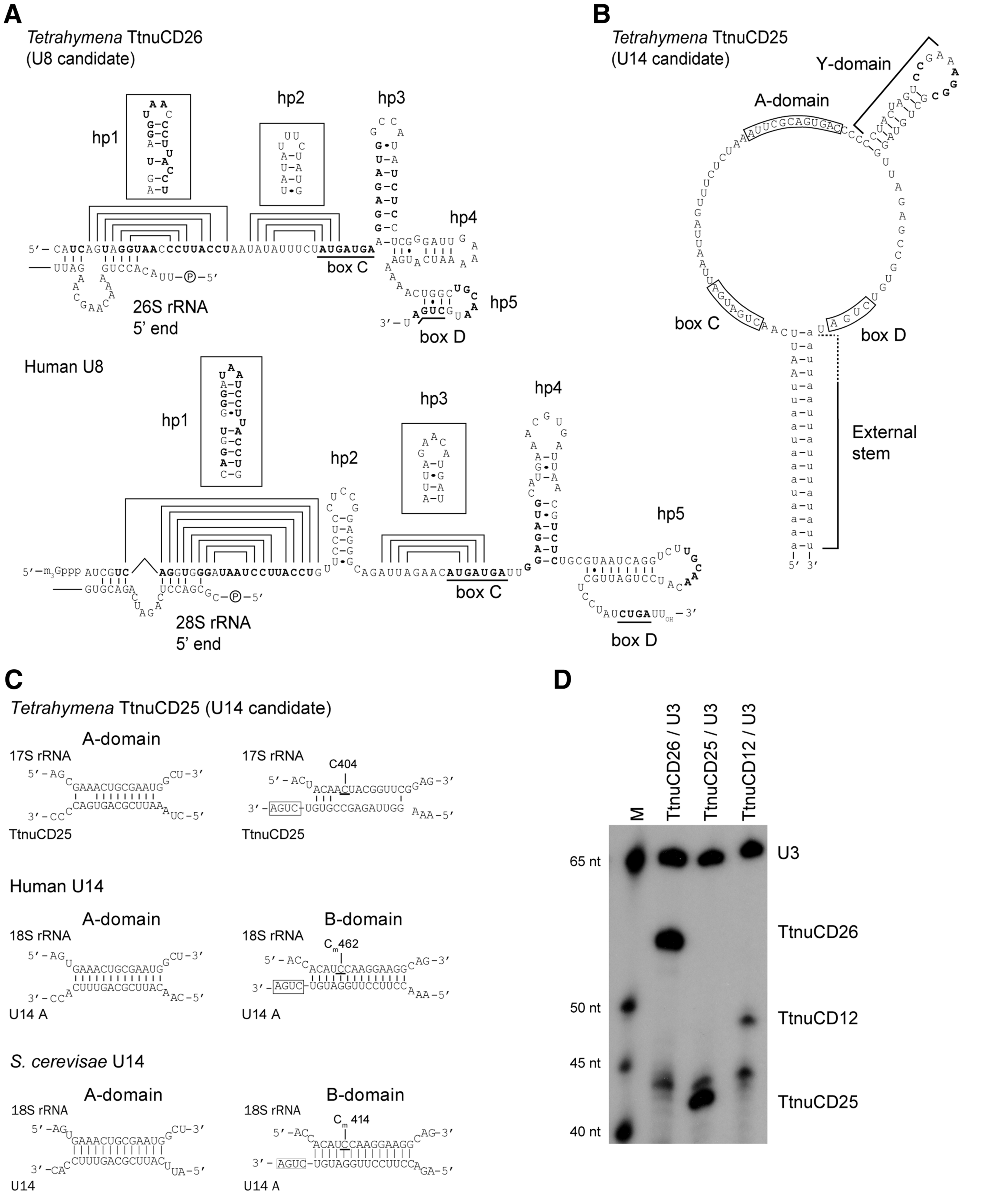

Figure 2. Candidates for *Tetrahymena* U8 and U14 snoRNAs. (**A**) Comparison of the secondary structure of *Tetrahymena* U8 candidate TtnuCD26 with human U8 snoRNA. Both RNAs are base-paired to the very 5′-end of the large ribosomal RNA (26S and 28S rRNA, respectively). Nucleotides of the box C and D are underlined. Hairpins are numbered hp1–5 in each structure. Alternative hairpin structure (hp1/hp2 in TtnuCD26 and hp1/hp3 in human U8) are indicated by lines and shown above in boxes. (**B**) A secondary structure model of *Tetrahymena* TtnuCD25 (U14 candidate). Box C, box D and the U14 specific A-domain sequences are framed and the Y-domain indicated. A predicted external stem-structure, primarily consisting of nucleotides not included in the mature snoRNA, is included in the structure in lower case letters. (**C**) Comparison of the base-pairing between the 17S/18S rRNA with the A-domains of human U14A, yeast U14 and the putative A-domain of TtnuCD25. The B-domain interactions of human and yeast U14 with 18S rRNA and the corresponding sequences of *Tetrahymena* RNAs are also shown. Modifications guided by U14 are indicated above the B-domain base-pairing and the corresponding position in *Tetrahymena* 17S rRNA is also highlighted. (**D**) Primer extension expression analysis of TtnuCD25 and TtnuCD26 in comparison with a canonical modification guide snoRNA TtnuCD12. *Tetrahymena* snoRNA U3 is used as an internal control in all lanes. The position of the 5′-end of the various RNA species and and sizes of molecular marker DNA oligos are given at the sides.

interaction with 17S rRNA corresponding to the interaction found between human and *Saccharomyces cerevisiae* U14 with 18S rRNA could also be formed in *Tetrahymena* (Figure 2C). However, the other canonical U14 element, the B-domain just upstream of the box D motif that is directing a modification of 18S rRNA in humans and yeast (human C462, *Tetrahymena* C404) could not be confirmed in TtnuCD25 (Figure 2C). This resembles the situation in *Drosophila* where the B-element was absent in the identified U14. Instead, another box C/D snoRNA was found to guide methylation of the corresponding nucleotide in flies thus decoupling the U14-guided modification and processing (40). Contrary to the study in flies we did not identify another box C/D snoRNA that could complement U14, but such a snoRNA could be absent from the library due to lack of coverage. A hairpin structure, the Y-domain, localized between the A-domain and B-domain is found in yeasts and plants but is not conserved in vertebrate U14. A Y-domain structure is also found in TtnuCD25 at the corresponding position (Figure 2B). The length of the Y-domain stem (8 bp) is conserved between yeast and plant, but the loop sequence in the two groups show no obvious relationship (yeast consensus AMGAACCY-AU versus plant consensus **CC - -** YGCC**RGGC**U, where M: A/C and R: U/C) (41). The putative Y-domain in TtnuCD25 can likewise be drawn with a stem of 8 bp though with a single nucleotide bulge, and the loop sequence of **CCG**AA**AGGC** resembles the consensus sequence in plants (bolded nucleotides in text and Figure 2B). Similar to many *Tetrahymena* box C/D snoRNAs TtnuCD25 did not exhibit a canonical box C/D snoRNA terminal stem, but had the potential to form a long stem of the adjacent 5′ and 3′ external sequences (Figure 2B).

Primer extension analysis of TtnuCD25 and TtnuCD26 in comparison with TtnuCD12, a snoRNA expected to be involved in guiding modifications of rRNA showed the U8 and U14 candidates to be as abundant as U3 and considerably more abundant than TtnuCD12 (Figure 2D). Since the expression level of snoRNAs involved in rRNA cleavage is expected to be higher than for modification guide snoRNAs, the results supported a role for TtnuCD25 and TtnuCD26 in pre-rRNA processing. Finally, northern blot analysis showed these RNAs to be present at all tested cellular conditions (Figure 3). U8 appears to decrease during starvation and to be less abundant in stationary phase, two conditions where ribosome synthesis is slowed down. However, these observations need to be confirmed by additional approaches.

Expression of ncRNAs

Primer extension was used to confirm expression of a subset of ncRNAs in the cDNA library (Figure 3A–C) and to verify the expression of the few snoRNAs that were found by inspection of genomic sequence in the vicinity of experimentally identified snoRNAs (Figure 3B; TtnuCD7, TtnuCD12 and TtnuCD15). The analysis confirmed the expression of all tested box C/D and H/ACA snoRNAs. In a few cases two or more strong signals were detected (Figure 3A and B; TtnuHACA11, TtnuHACA23 and TtnuCD25). This could be due to premature primer extension stops caused by structural features but could also indicate that the snoRNA genes were transcribed or processed into two different length variants. To determine this, we analyzed the candidates by northern blot analysis using RNA isolated from different cellular conditions and in one case we found evidence of two variants of different length (Figure 3D; TtnuHACA11). For several snoRNAs, faint primer extension signals that represented products longer than the main signal could be discerned (Figure 3A and B; TtnuHACA23 and TtnuCD10). These signals could originate from processing intermediates of precursors, so we tested if the presence or absence of these signals was correlated to the genomic organization of the snoRNA gene. No such correlation was found and the origin of these signals remains unknown.

TtnuCD10 which was identified as a *Tetrahymena* U15 was also analyzed by northern blot analysis. It resided as the fourth RNA in the largest snoRNA cluster identified and was the longest of the box C/D-like snoRNAs found in *Tetrahymena* (117 nt compared with the general 55–77 nt). The northern blot showed this to be abundant but did not reveal dramatic differences in its expression (Figure 3D).

We also analyzed the expression of unassigned ncRNAs by primer extension (Figure 3C). Primer extension analysis with TtnuUkn1, 3 and 5 primers gave a single clear signal. The TtnuUkn4 results indicated two different length variants. This observation was paralleled by northern blotting that showed one RNA (230 nt) found equally expressed at all cellular conditions, and one (285 nt) that was differentially expressed (Figure 3D). It should be noted that the TtnuUnk4 cDNA clone matched the *Tetrahymena* genome at three different locations each of which could give rise to different RNAs. Primer extension analysis showed several signals for TtnuUkn2 and TtnuUkn6 and no signal for TtnuUkn7 (Figure 3C). TtnuUkn2 was detected by northern blotting as a single RNA of a size consistent with the longest product in the primer extension analysis (Figure 3D).

Prediction of snoRNA targets and their conservation in rRNA

Most snoRNAs guide nucleotide modification of other cellular RNAs, in particular rRNA. They function through base-pairing of a snoRNA guide sequence to the target molecule that direct associated proteins to methylate (box C/D) or pseudouridylate (box H/ACA), a single nucleotide. Each snoRNA have the potential to target one nucleotide per box D/D′ in box C/D snoRNAs and at least one per pseudouridylation pocket in box H/ACA snoRNAs, respectively (32). We used the search algorithms SnoScan and SnoGPS to predict the targets of the identified snoRNAs. A detailed description of the approach and the results are presented in Supplementary Tables S2 and S3. Not surprisingly, the majority of snoRNAs were predicted to target rRNA. In total, we predicted 46 N_ms and 31 Ψs in rRNA. When compared to snoRNA-guided rRNA modifications in yeast, plant and human, 20 N_m and 20 Ψ modifications were conserved between *Tetrahymena* and one or more of the other model organisms. This implies that almost half of the predicted modifications namely 26 N_ms and 11 Ψs were specific for *Tetrahymena* in this comparison. The modifications and their conservation in yeast, plant and humans were superimposed onto the secondary structure of *Tetrahymena* rRNA large (LSU) and small subunit (SSU) (Figure 4 and Supplementary Figure S3). This is most likely a non-exhaustive dataset limited by the relative low depth of the sequencing approach and more snoRNAs and thus, modifications are expected to be found by future efforts. The distribution of modifications in rRNA is not random and modified nucleotides are generally clustered in

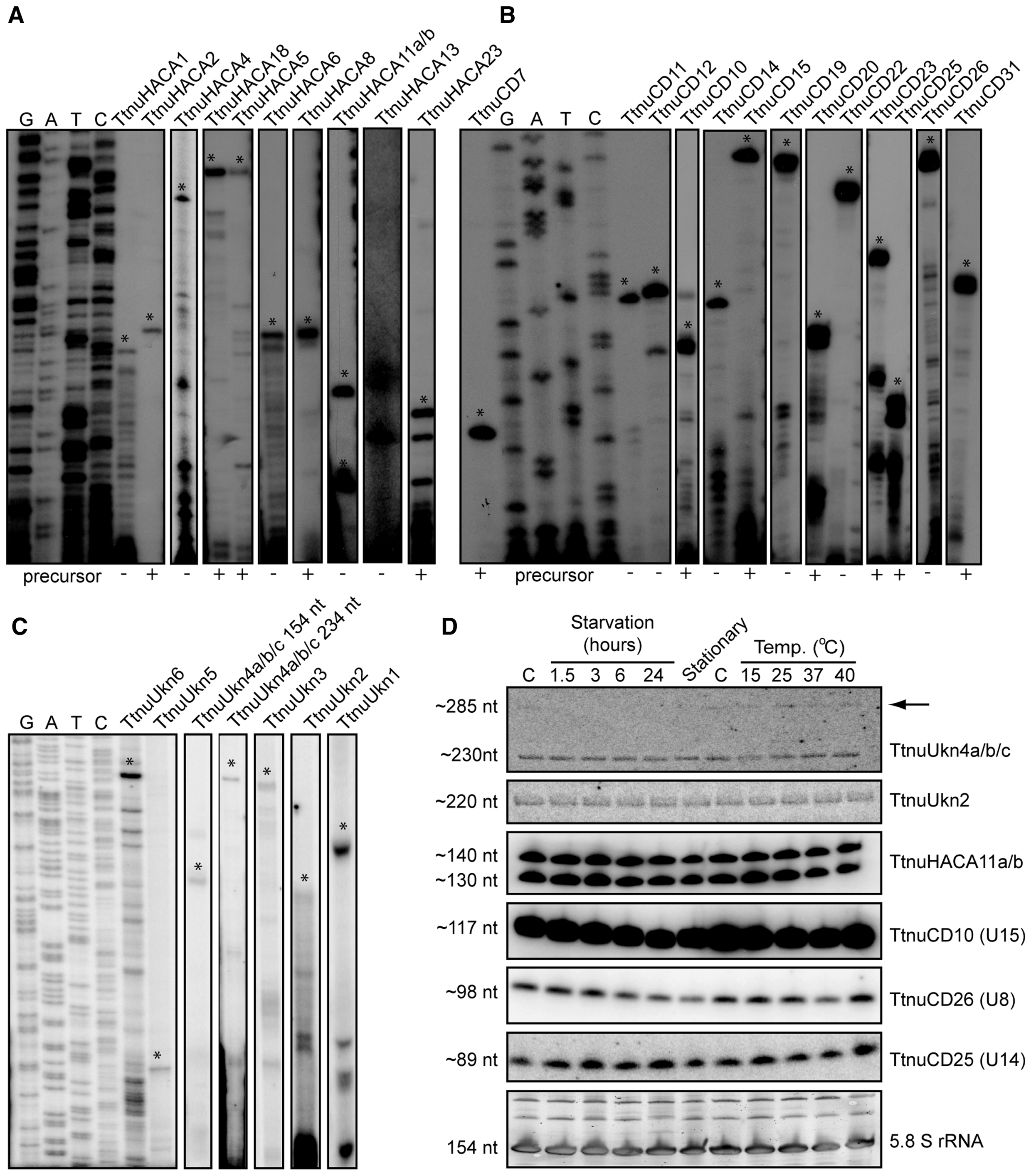

Figure 3. Expression studies of selected ncRNAs. (**A**) Primer extension analysis of nuclear RNA with primers targeted against various box H/ACA snoRNAs (TtnuHACAs) as marked above the lanes. Primer extension reactions were run next to a known but unrelated sequencing reaction. Signals corresponding to 5′-ends are marked with an asterisk. Below the lanes are indicated whether a faint signal from a putative precursor molecule was observed when the signals were overexposed (+) or if no precursor signal was observed (−). (**B**) Similar analysis of box C/D snoRNAs (TtnuCDs). (**C**) Similar analysis of unknown ncRNAs (TtnuUkn's). (**D**) Northern blot analysis of RNA isolated from cells at different conditions with probes representing selected ncRNAs. The lowest panel, serving as loading control, shows 5.8S rRNA visualized by SYBR Gold staining. The arrow indicates a low abundant, differentially expressed RNA.

functionally important domains (42). The predicted modifications, including the species-specific *Tetrahymena* modifications, followed this pattern. Thus, very few modifications were observed in domain I and the upper part of domain II of the LSU (Supplementary Figure S3). Likewise, LSU domain III was apparently unmodified, as observed in yeast. Conversely, the LSU domains IV and V that are heavily modified in the yeast, plant and human

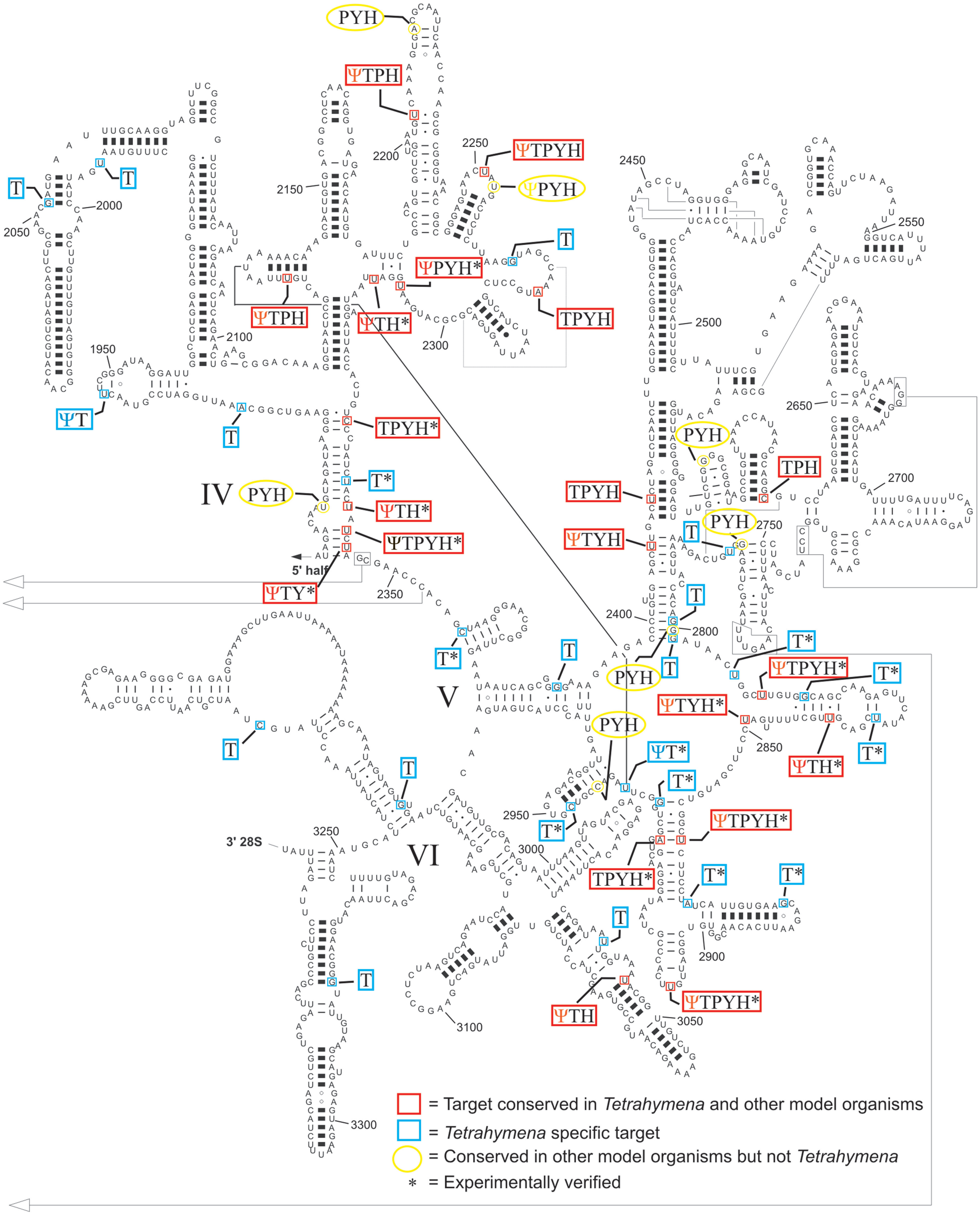

Figure 4. Map of predicted and experimentally determined nucleotide modifications on the secondary structure of the 3′ half of 26S rRNA (similar results for 26S 5′ half and 17S rRNA are depicted in Supplementary Figure S3). Putative and verified target nucleotides of the identified snoRNAs are framed and labeled according to their conservation in *Tetrahymena* (T), plant (*Arabidopsis thaliana*; P), yeast (*S. cerevisiae*; Y) and human (H). Pseudouridylations are marked by Ψ and experimentally verified modifications by an asterisk. The rRNA secondary structures were adopted from www.rna.ccbb.utexas.edu (65) and information regarding modifications from plant, yeast and human was extracted from the 3D ribosomal modification map website http://people.biochem.umass.edu/fournierlab/3dmodmap/main.php (66). Note that the data presented in the figure is a non-exhaustive compilation of modifications and that modifications in plant, yeast and human are included only at sites of modification in *Tetrahymena* or if modifications are found in all three reference organisms.

rRNA were also predicted to be heavily modified in *Tetrahymena* (Figure 4). Interestingly, several predicted *Tetrahymena*-specific modifications were located in the peripheral 5′ half of LSU domain IV where no modifications have been observed in yeast, plant or humans. LSU domain VI has no modifications in yeast and plant rRNA, but has both methylated and pseudouridylated residues in humans. This domain was likewise predicted to contain modifications in *Tetrahymena* rRNA.

In addition to the prediction of targets in rRNA, we found a surprisingly large number of snoRNAs that seem to target RNAs other than the mature rRNA species. A total of 18 targets in tRNA (9), snRNA (6) and pre-rRNA (3) were predicted for box C/D RNAs and 15 targets in pre-rRNA (8), tRNA (5), snRNA (1) and SRP-RNA (1) for box H/ACA RNAs (Supplementary Tables S2 and S3). Finally, 5 box D/D's in 4 box C/D snoRNAs and 8 pseudouridylation pockets in 5 box H/ACA snoRNAs could not be associated with a target (Supplementary Tables S2 and S3).

Experimental mapping of Nm and Ψ in rRNA and other cellular RNAs

The modification repertoire of the identified snoRNAs included a surprisingly large number of *Tetrahymena*-specific modifications in both rRNA and other cellular RNAs. To investigate if this was an artifact of the prediction method, or if *Tetrahymena* exhibits an alternative modification pattern, some of the predicted modifications were tested experimentally. To map N_ms we applied primer extension analysis with varying dNTP concentrations. At lower dNTP concentrations, reverse transcriptase pauses at methylated residues, which results in a signal 1 nt before the position of the modified residue as read on a sequencing ladder run in parallel. Pseudouridines were experimentally mapped by CMC-treatment followed by detection by primer extension. The results from the primer extension analysis with four different oligos are presented in Figure 5A–D and experimentally detected modifications are marked with an asterisk in Figure 4 and Supplementary Figure S3. In total, we verified seven

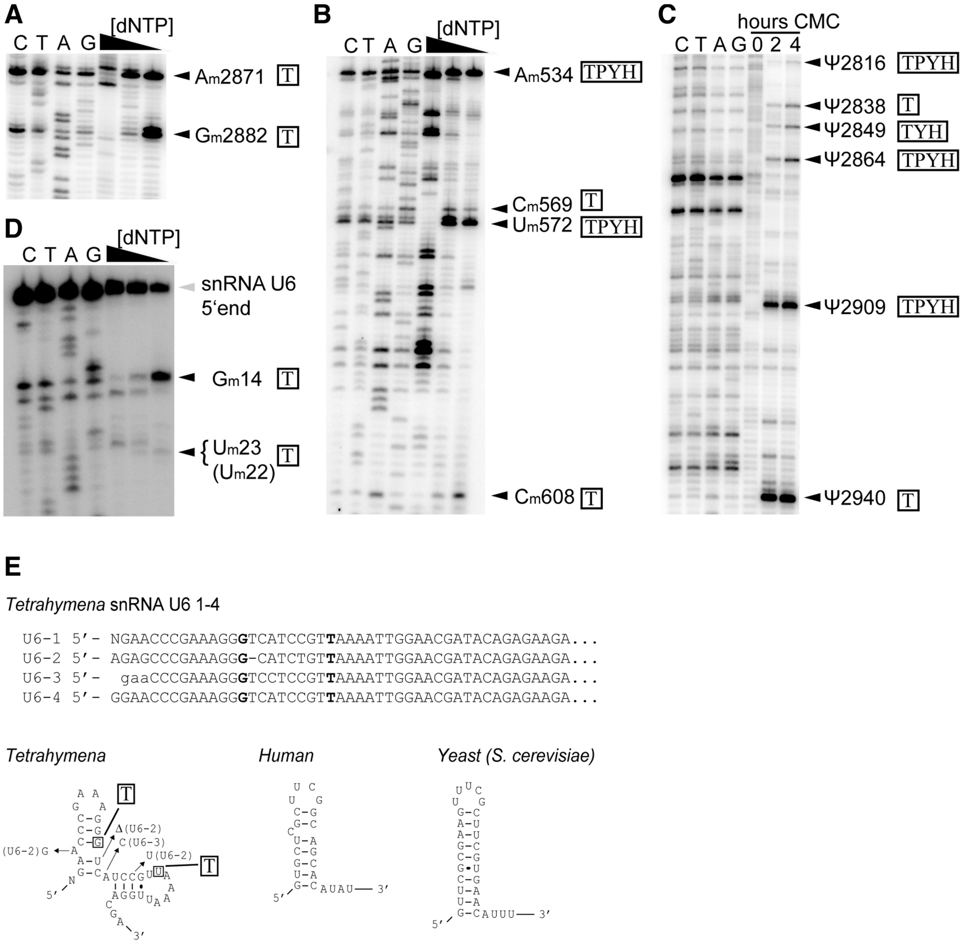

Figure 5. Experimental verification of nucleotide modifications by primer extension analysis. (**A**) 2′-*O*-methylations in 26S rRNA. (**B**) 2′-*O*-methylations in 17 S rRNA. (**C**) Pseudouridylations in 26S rRNA. (**D**) 2′-*O*-methylations in snRNA U6-1–4. The primer extension reactions were run next to the appropriate sequence obtained by dideoxy RNA sequencing. The oligo applied in the primer extension analysis in (D) could base-pair with an identical sequence in all four U6 (1–4) RNAs, but U6-2 has a single nucleotide deletion. Thus, the U_m23 is equal to U_m22 in U6-2. (**E**) Sequence alignment of the 5′-ends of *Tetrahymena U6* snRNA genes and secondary structures of the 5′ stem-loop of U6 snRNA from *Tetrahymena*, human and yeast. Methylated nucleotides are indicated by bold in the alignment and boxed in the secondary structures. Nucleotide differences between U6 variants are indicated with arrows and the relevant snRNAs are given in parentheses.

predicted *Tetrahymena*-specific rRNA methylations and two pseudouridylations. Further, we verified eight methylations and twelve pseudouridylations conserved between *Tetrahymena* and at least one of the other model organisms. In the process, an additional nine *Tetrahymena*-specific rRNA methylations, for which we did not identify an associated snoRNA guide, were detected. Finally, four methylated and one pseudouridylated nucleotide conserved between yeast, plant and humans, but not predicted in our analysis, was experimentally shown to be present also in *Tetrahymena*.

TtnuCD21 and TtnuCD30 were predicted to target U_m23 in U6-1, U6-3 and U6-4 snRNA (U_m22 in U6-2) and C_m18 in U6-2 snRNA, respectively. When U6 snRNA was analyzed experimentally two modifications were detected (Figure 5D and E). One was at U_m22/t U_m23 as predicted by Ttnu21 box D′. The other modification at G_m14 was not predicted. In contrast to the U6 snRNA modifications, we were unable to analyze the predicted modifications in tRNA by primer extension because of technical limitations.

DISCUSSION

Since the detection of most classes of ncRNAs constitutes a specific challenge due to their diversity and relative poor sequence conservation many genome projects are underrepresented in ncRNA annotation. However, systematic searches for smaller ncRNAs by bioinformatics and experimental 'RNomics' have been carried out for several model organisms, e.g. mouse, *Drosophila* and *C. elegans* and have expanded the number of known ncRNAs greatly (30,40,43). In this study, we have experimentally identified ncRNAs from the macronucleus of the ciliate *T. thermophila* in the size range of 40–500 nt. Not surprisingly, the majority of our findings could be classified to highly expressed classes of ncRNAs such as snRNAs and snoRNAs.

Tetrahymena snoRNAs

We identified 66 ncRNAs that could be classified as box H/ACA and box C/D snoRNAs based on the presence of canonical sequence signatures and the secondary structure (Figure 1). The box H/ACA snoRNAs were in general similar to the two stem box H/ACA snoRNAs found in vertebrates. No single stem H/ACA-like snoRNAs, as seen in some other protists, e.g. *Euglena* and *Trypanosoma* (44), was observed. Nine of the identified box H/ACA snoRNAs deviated from the hallmark 3′-terminal box ACA in having an alternative sequence AUA, UAA, UCA or AAA. However, since several of these were predicted to guide modifications of rRNA conserved between *Tetrahymena*, yeast, plant and humans we expect them to be functional pseudouridylation guide snoRNAs (Supplementary Table S3). Occasional deviations from the ACA consensus sequence have been known from the first report of these RNAs (29,45) and an AGA motif in place of ACA is the general rule in Trypanosomes (44,46).

The box C/D snoRNAs showed a more systematic deviation, some from the canonical structure. Although harboring the terminal boxes C and D as well as the internal boxes C′ and D′, most *Tetrahymena* box C/D snoRNAs did not have the potential to form the canonical short 5′-, 3′-terminal stem. Instead, several *Tetrahymena* box C/D snoRNAs could form a 5–25 bp stem between the flanking 5′- and 3′-sequences (Supplementary Figure S2). In addition, several TtnuCDs also had the potential to form extensive internal base-pairing. The short 5′-, 3′-terminal stem has been shown to be essential for maturation and accumulation of vertebrate intron encoded, as well as yeast polycistronic box C/D snoRNAs (47). However, some mammalian, several yeast and Trypanosomatids box C/D snoRNAs have also been reported to lack the 5′-, 3′-terminal stem. Instead many of them show, similar to the *Tetrahymena* box C/D snoRNAs, a potential for forming a stem between sequences flanking the mature box C/D snoRNA (48–51). Notably, existence of an external stem structure was shown to support accumulation or correct processing of the snoRNAs with an unusual short or absent canonical terminal stem (49–51). Thus, it seems that the occasional strategy of an external stem in mammals is a more general principle for *Tetrahymena* box C/D snoRNAs. In support of this, it was noted that mammalian box C/D snoRNAs with the external stem featured just two unpaired 3′-terminal nucleotides following the box D, whereas the typical box C/D snoRNA, containing the internal 5′-, 3′-terminal stem, had 5 or 6 nt 3′ of box D (50). Concordantly, in *Tetrahymena* the vast majority of box C/D snoRNAs had just 2 nt 3′ of box D (Supplementary Table S2). It remains to be elucidated how box C/D snoRNAs that apparently lack both the 5′-, 3′ terminal stem and the potential to form an external stem are processed accurately and accumulate.

SnoRNA genomic organization

SnoRNA genes are known to exhibit a large degree of flexibility in genomic organization between and within species. In humans, worms and flies they are mainly individual units confined to introns and co-transcribed with the host protein gene, whereas in plants and in the protist *Trypanosoma* they are clustered either in introns or in intergenic regions (13,44). Based on the snoRNAs analyzed in this article, the genomic organization in *Tetrahymena* resembles the situation in yeast in which the majority of snoRNA genes are found in intergenic regions as independent units with a fraction organized in clusters. In addition, a few are located within introns of protein-coding genes (Figure 1B). Curiously, the clustered snoRNAs in yeast are all of the box C/D snoRNA subclass, and this tendency was also observed in *Tetrahymena,* albeit with a single box H/ACA snoRNA exception.

The snoRNA USE (5′-AAACCCATAA) (Figure 1C) identified in this study was identical to the one identified previously upstream of *Tetrahymena* snRNA genes and in ncRNA genes of *Paramecium* (26,52). It was found upstream of 22 out of 27 intergenic monocistronic snoRNAs. The distance from the USE to the mature gene product was generally longer in *Tetrahymena* snoRNA genes compared with the snRNA genes and showed more variation. This could reflect that intergenic snoRNAs are produced as precursors of varying length before they are processed to the mature RNA. This is consistent with the formation of the external stem structure in *Tetrahymena* box C/D precursors to ensure correct snoRNP assembly and processing. In agreement with this, primer extension products longer than corresponding to the mature snoRNA were demonstrated for several snoRNAs of both the box C/D and the box H/ACA subgroup (Figure 3A and B). Among the clustered snoRNAs, fewer (11 out of 19) had an identifiable USE and the sequence tended to deviate from the consensus suggesting that at least some of these are co-transcribed.

The sequences downstream of clustered and intergenic individual snoRNAs were clearly different and both differed considerably from snRNA genes and the genes encoding snoRNAs U3-1–4 (Figure 1D) previously analyzed in *Tetrahymena* (26). The T-tracts following monocistronic snoRNA genes could indicate that RNA polymerase III is involved in the transcription of this group in *Tetrahymena* although no experimental evidence (e.g. α-amanitin sensitivity experiments) exist to support this. SnoRNA genes are in general, believed to be transcribed by RNA polymerase II but RNA polymerase III has previously been implicated in transcription of intergenic monocistronic snoRNA genes in *C. elegans* and clustered snoRNAs in *Paramecium* (13,52). The absence of T-tracts downstream of clustered snoRNAs suggests that they are transcribed by RNA polymerase II or as polycistronic transcripts. However, a polycistronic transcript is in disagreement with the presence of an USE in several clustered snoRNAs. One explanation could be that transcription of a snoRNA cluster can be initiated at several locations within the cluster. In a preliminary study, it was shown that the genomic organization of some of the snoRNA genes within the largest cluster was not conserved among *Tetrahymena* species (unpublished results). Given this evolutionary dynamic situation, it is possible that some of the observed USE-like elements are degenerate and no longer involved in transcription initiation. It seems that the trend, going from unicellular organisms to plants and metazoans, is a reduction in the number of individual promoters driving snoRNA expression. This is obtained through evolution by snoRNA gene clustering and colonization of introns (13). Our observations suggest that *Tetrahymena* has adopted an intermediary position in this spectrum.

SnoRNA guided modification pattern in *Tetrahymena*

Targets of the snoRNAs were predicted based on the guide sequences in box C/D snoRNAs and the internal loop in stems of box H/ACA snoRNAs. The majority of snoRNAs target rRNA and most of the sites were conserved in one or more of yeast, plant and human (Figure 4 and Supplementary Figure S3). In addition, we predicted a surprisingly large number of modifications that were *Tetrahymena*-specific in this comparison. In order to determine if this was correctly predicted, we verified a subset of the *Tetrahymena*-specific and inter-species conserved, 2′-*O*-methylations and pseudouridylations experimentally (Figures 4 and 5 and Supplementary Figure S3). In doing this, we experimentally mapped additional modifications which were not accounted for by the predictions based on snoRNAs in our cDNA library. This indicates that several snoRNAs remain to be identified. Alternatively, these modifications could be introduced in a snoRNA-independent pathway. Nucleotide modifications in rRNA are generally clustered in conserved and functionally important domains (42) and many modification sites are highly conserved among distantly related species. A high number of species-specific predicted or experimentally determined modifications have also been found in the protozoan *Trypanosoma brucei* (44). However, there was little overlap (three modifications) between *Tetrahymena* and *Trypanosoma* modification patterns, beside highly conserved modifications also seen in other model organisms (20 modifications). One possibility for these differences could be that the species-specific modifications accommodate differences in sequence and secondary structure of variable domains in the rRNA. Both the *Tetrahymena*-specific modifications and the conserved modifications concur with the clustering in functional important regions of the rRNA. As an example, the lower part of domain II in LSU 5′ half is rich in modifications, whereas modified residues are absent in the upper part. Similarly, domain V and the conserved 3′ half of domain IV are heavily modified. In the yeast ribosomal model, the modifications of these three domains are defining a shell around the A- and P-site tRNAs (42). Modifications in the rRNA of *Tetrahymena* were also predicted in the 5′ half of domain IV including the variable region (D8). Although the D8 region is not evolutionary conserved in primary sequence, it has been shown to be essential for rRNA stabilization and/or processing (53). Historically, the function of the Nms and Ψs has been somewhat enigmatic and mainly believed to be a fine-tuning of the ribosome structure. However, recent work has shown significant growth deficiencies in cells with individual box C/D snoRNA deletions, and severe ribosomal performance decrease in response to H/ACA snoRNA depletion. In addition, ribosomal modifications were shown to be significant for translation accuracy and rRNA biogenesis (54–57). These findings, together with new knowledge on ribosome structure (58,59), increase the importance of determining rRNA modification patterns and the factors involved therein in the effort toward a complete understanding of ribosome biology.

In addition to rRNA modifications, we predicted snoRNA guided modification of nucleotides in snRNAs, tRNAs and SRP RNA. Some of these are consistent with a previous analysis of the nucleotide composition of several snRNAs and snoRNAs that revealed many examples of ribose methylations and pseudouridylations (60). SnoRNA guided modifications of snRNAs have been verified in several cases and modifications of nucleotides in snRNA U2 are required for snRNP assembly and pre-mRNA splicing (61). Some modifications of snRNAs are guided by particular box C/D-H/ACA chimeric snoRNAs termed small Cajal body RNAs (scaRNAs).

Other snRNA modifications are guided by regular snoRNAs (62). We did not identify any scaRNAs in the *Tetrahymena* macronuclear cDNA library, but predicted that regular *Tetrahymena* snoRNAs guide modifications of U2, U4, U5 and U6 snRNAs as well as snoRNA U3 (Supplementary Tables S2 and S3). In the case of snRNA U6, we verified the predicted modification of U_m23/U_m22. Additionally, we mapped an unpredicted modification at G_m14 (Figure 5D). These modifications may be of interest because *Tetrahymena* U6 deviates from most other U6 snRNAs in the secondary structure of the 5′-end (Figure 5E). Most of the 200 sequences in the Rfam seed alignment folds into a 5′-UUCG capped single hairpin at the 5′-end. In contrast, *Tetrahymena* U6 snRNA folds into a two-hairpin structure supported by the sequence variation among the four *U6* genes. Furthermore, the 5′ terminal stem is the only of the Rfam seed sequences capped by a 5′-GAAA tetraloop, a loop sequence often involved in tertiary interactions with protein or RNA partners. In addition to having a structurally different 5′ stem-loop structure, the methylations in this domain also sets *Tetrahymena* apart from other organisms, e.g. human and yeast that apparently are unmodified in their 5′ stem-loops. It will be of interest to see if these structural differences are important for splicing of *Tetrahymena* introns that are highly AU-rich. All of the snoRNAs predicted to guide snRNA modification were also predicted to guide modifications of rRNA. This could imply that they are active in two distinct nuclear compartments namely the Cajal body, where most snRNAs are believed to be modified and the nucleolus where rRNA modification takes place (14). SnoRNA guided modification of tRNAs has so far been demonstrated only in Archaea, and although predicted in Eukarya (43) they remain to be verified.

The cDNA library also included one box C/D snoRNA and four box H/ACA snoRNAs where a target could not be assigned in the rRNA, snRNA, SRP RNA and tRNAs included in our target library, as well as five snoRNAs with two guide sequences but only one predicted target. SnoRNAs with just one target are not uncommon. SnoRNAs with no apparent target in rRNA or snRNAs are highly interesting. In humans some 'orphan' snoRNAs have been shown to target mRNA and control alternative splicing and to be involved in the human disorder Prader–Willi syndrome (10,63).

New ncRNAs in *Tetrahymena*

Our study uncovered two groups of *Tetrahymena* RNAs that may be functionally distinct RNAs: the long box C/D RNAs and the TtnUkn's. Due to limited conservation at the primary sequence level of these RNAs, we compared the five long box C/D *Tetrahymena* RNAs with the Rfam database using a structure-based search algorithm. Based on this analysis, we propose that TtnuCD26 is *Tetrahymena* U8 and that TtnuCD25 is *Tetrahymena* U14. U14 is required for early steps in processing of the ribosomal RNA precursor and is essential for cell survival. U8 is involved in the maturation of 5.8S and 28S rRNA in the large ribosomal subunit. U14 is widespread but relatively few examples have been described. In contrast, many U8 sequences are known, but it has previously not been found outside metazoans. Both TtnuCD26 and TtnuCD25 were supported as *Tetrahymena* U8 and U14 by the presence of conserved sequence elements and conserved complementarity to rRNA (Figure 2A, B and C). Also, their abundances were similar to U3 [previously estimated at 4×10^5 molecules/cell (60)] and higher than a typical modification guide snoRNA (Figure 2D). However, TtnuCD26 deviated from human U8 with respect to some secondary structure elements and was missing the sequence motif reported to bind LSm proteins (Figure 2A). TtnuCD25 contained the A-domain and a Y-domain, however, similar to a reported fly U14 (40) it could not form the B-element interaction with 17S rRNA described in most model organisms (Figure 2B and C). This should be viewed in relation to other deviations in *Tetrahymena*, e.g. the apparent absence of a conventional TMG (trimethyl guanosine) cap on several snoRNAs [(60), unpublished results] and suggests some unique characteristics of *Tetrahymena* rRNA processing that remains to be elucidated. Along the same lines, TtnuCD32 could be a new player in ribosomal processing. This RNA is abundant, highly structured and has a long sequence stretch with complementarity to rRNA. Yet, we were unable to identify any homologue in the most studied model organisms. Genetic knock-down strains of this RNA have demonstrated it to be essential in *Tetrahymena* (unpublished results). Further characterization of the molecular defects in these knock-down strains and the possible role of TtnuCD32 in ribosomal RNA processing is underway.

The cDNA library presented in this work included seven sequences that we suggest represent new *Tetrahymena* ncRNAs (TtnuUkn's). We were unable to identify homologues of any of these by searches of Genbank and several ncRNA databases. All but one (TtnuUkn7) appeared to be expressed as evidenced by primer extension although only four (TtnuUkn1, 3, 4 and 5) had well-defined 5′-ends (Figure 3C). Two were detected in northern blot analysis (Figure 3D). TtnuUkn2 was expressed equally at all conditions tested and the size determined by northern blotting analysis (~220 nt) of whole cell RNA corresponded well to the size determined by primer extension analysis (211 nt) of nuclear RNA. Two bands could be recognized by northern blot analysis of TtnuUkn4. One was expressed at equal levels at all conditions tested, but a faint band above the major band was differentially expressed and could be discerned in RNA from exponentially growing cells, in heat shock and cold shock RNA but not in starved or stationary phase RNA. The single clone representing TtnuUnk4 was one of the few clones that mapped to more than one locus in the *Tetrahymena* genome. Thus, it is possible that the two bands represent transcripts from two different loci. However, the existence of a USE upstream of both the two 5′-ends determined by primer extension suggests that the two length variants observed by primer extension could originate from the same locus.

The excision of ~10–20% of the MIC genome, primarily repetitive sequences, in the formation of the transcriptional active MAC leaves *Tetrahymena* as an attractive model for deeper ncRNA sequencing projects. The RNomics approach used here identifies primarily the highly abundant snRNAs and snoRNAs. It is possible that the elimination of MIC genomic sequence, presumably diminishing transcripts from, e.g. transposable elements would make detection of novel low and medium expressed *bona fide* small ncRNAs by deep sequencing approaches more feasible in this organism compared to other model organisms.

ACCESSION NUMBERS

JF909302, EF503647, JF909303, EF503648, JF909304, JF909305, JF909306, EF503645, JF909307, JF909308, JF909309, JF909310, JF909311, JF909312, JF909313, EF503644, JF909314, JF909315, JF909316, JF909317, JF909318, JF909319, EF503641, JF909320, EF503642, EF503643, JF909321, JF909322, EF503646, EF503640, JF929905, JF909323, JF909324, JF909325, EF503649, EF503650, JF909326, JF909327, JF909328, JF909329, JF909330, EF503653, JF909331, JF909332, JF909333, JF909334, JF909335, EF503656, EF503658, JF909336, JF909337, EF503651, EF503652, EF503654, JF909338, EF503655, EF503657, EF503659, JF909339, JF909340, JF909341, JF909342, JF909343, JF909344 and JF909345.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.: Supplementary Figures S1–S3, Supplementary Tables S1–S4.

ACKNOWLEDGEMENTS

We wish to thank Franz Frenzel for technical assistance, Elfar Torarinsson for Linux and snoGPS support, Christian Anthon and Jan Gorodkin for the analysis using Infernal, as well as Lis Hasholt for critical reading of the article.

FUNDING

Faculty of Health Science, University of Copenhagen (to K.L.A.) and a grant from The Danish Natural Science Research Council (to H.N.). Funding for open access charge: Faculty of Health Science, University of Copenhagen.

Conflict of interest statement. None declared.

REFERENCES

1. Willingham,A.T. and Gingeras,T.R. (2006) TUF love for 'junk' DNA. *Cell*, **125**, 1215–1220.
2. Wilusz,J.E., Sunwoo,H. and Spector,D.L. (2009) Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.*, **23**, 1494–1504.
3. Prasanth,K.V. and Spector,D.L. (2007) Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. *Genes Dev.*, **21**, 11–42.
4. Farazi,T.A., Juranek,S.A. and Tuschl,T. (2008) The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. *Development*, **135**, 1201–1214.
5. Plath,K., Mlynarczyk-Evans,S., Nusinow,D.A. and Panning,B. (2002) Xist RNA and the mechanism of X chromosome inactivation. *Annu. Rev. Genet.*, **36**, 233–278.
6. Ganot,P., Bortolin,M.L. and Kiss,T. (1997) Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. *Cell*, **89**, 799–809.
7. Kiss-Laszlo,Z., Henry,Y., Bachellerie,J.P., Caizergues-Ferrer,M. and Kiss,T. (1996) Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. *Cell*, **85**, 1077–1088.
8. Granneman,S. and Baserga,S.J. (2004) Ribosome biogenesis: of knobs and RNA processing. *Exp. Cell Res.*, **296**, 43–50.
9. Cavaille,J., Buiting,K., Kiefmann,M., Lalande,M., Brannan,C.I., Horsthemke,B., Bachellerie,J.P., Brosius,J. and Huttenhofer,A. (2000) Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *Proc. Natl Acad. Sci. USA*, **97**, 14311–14316.
10. Sahoo,T., del,G.D., German,J.R., Shinawi,M., Peters,S.U., Person,R.E., Garnica,A., Cheung,S.W. and Beaudet,A.L. (2008) Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. *Nat. Genet.*, **40**, 719–721.
11. Ender,C., Krek,A., Friedlander,M.R., Beitzinger,M., Weinmann,L., Chen,W., Pfeffer,S., Rajewsky,N. and Meister,G. (2008) A human snoRNA with microRNA-like functions. *Mol. Cell*, **32**, 519–528.
12. Saraiya,A.A. and Wang,C.C. (2008) snoRNA, a novel precursor of microRNA in Giardia lamblia. *PLoS Pathog.*, **4**, e1000224.
13. Dieci,G., Preti,M. and Montanini,B. (2009) Eukaryotic snoRNAs: a paradigm for gene expression flexibility. *Genomics*, **94**, 83–88.
14. Filipowicz,W. and Pogacic,V. (2002) Biogenesis of small nucleolar ribonucleoproteins. *Curr. Opin. Cell Biol.*, **14**, 319–327.
15. Kim,S.H., Spensley,M., Choi,S.K., Calixto,C.P., Pendle,A.F., Koroleva,O., Shaw,P.J. and Brown,J.W. (2010) Plant U13 orthologues and orphan snoRNAs identified by RNomics of RNA from Arabidopsis nucleoli. *Nucleic Acids Res.*, **38**, 3054–3067.
16. Li,D., Wang,Y., Zhang,K., Jiao,Z., Zhu,X., Skogerboe,G., Guo,X., Chinnusamy,V., Bi,L., Huang,Y. *et al.* (2011) Experimental RNomics and genomic comparative analysis reveal a large group of species-specific small non-message RNAs in the silkworm Bombyx mori. *Nucleic Acids Res.*, **39**, 3792–3805.
17. Kruger,K., Grabowski,P.J., Zaug,A.J., Sands,J., Gottschling,D.E. and Cech,T.R. (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell*, **31**, 147–157.
18. Greider,C.W. and Blackburn,E.H. (1987) The telomere terminal transferase of Tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity. *Cell*, **51**, 887–898.
19. Doerder,F.P., Deak,J.C. and Lief,J.H. (1992) Rate of phenotypic assortment in Tetrahymena thermophila. *Dev. Genet.*, **13**, 126–132.
20. Eisen,J.A., Coyne,R.S., Wu,M., Wu,D., Thiagarajan,M., Wortman,J.R., Badger,J.H., Ren,Q., Amedeo,P., Jones,K.M. *et al.* (2006) Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote. *PLoS. Biol.*, **4**, e286.
21. Iwamura,Y., Sakai,M. and Muramatsu,M. (1982) Rearrangement of repeated DNA sequences during development of macronucleus in Tetrahymena thermophila. *Nucleic Acids Res.*, **10**, 4279–4291.
22. Couvillion,M.T., Lee,S.R., Hogstad,B., Malone,C.D., Tonkin,L.A., Sachidanandam,R., Hannon,G.J. and Collins,K. (2009) Sequence, biogenesis, and function of diverse small RNA classes bound to the Piwi family proteins of Tetrahymena thermophila. *Genes Dev.*, **23**, 2016–2032.
23. Lee,S.R. and Collins,K. (2006) Two classes of endogenous small RNAs in Tetrahymena thermophila. *Genes Dev.*, **20**, 28–33.
24. Mochizuki,K., Fine,N.A., Fujisawa,T. and Gorovsky,M.A. (2002) Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in tetrahymena. *Cell*, **110**, 689–699.

25. Yao,M.C., Fuller,P. and Xi,X. (2003) Programmed DNA deletion as an RNA-guided system of genome defense. *Science*, **300**, 1581–1584.
26. Orum,H., Nielsen,H. and Engberg,J. (1992) Structural organization of the genes encoding the small nuclear RNAs U1 to U6 of Tetrahymena thermophila is very similar to that of plant small nuclear RNA genes. *J. Mol. Biol.*, **227**, 114–121.
27. Orum,H., Nielsen,H. and Engberg,J. (1991) Spliceosomal small nuclear RNAs of Tetrahymena thermophila and some possible snRNA-snRNA base-pairing interactions. *J. Mol. Biol.*, **222**, 219–232.
28. Atzorn,V., Fragapane,P. and Kiss,T. (2004) U17/snR30 is a ubiquitous snoRNA with two conserved sequence motifs essential for 18S rRNA production. *Mol. Cell Biol.*, **24**, 1769–1778.
29. Nielsen,H., Orum,H. and Engberg,J. (1992) A novel class of nucleolar RNAs from Tetrahymena. *FEBS Lett.*, **307**, 337–342.
30. Huttenhofer,A., Kiefmann,M., Meier-Ewert,S., O'Brien,J., Lehrach,H., Bachellerie,J.P. and Brosius,J. (2001) RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J.*, **20**, 2943–2953.
31. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
32. Kiss,T. (2001) Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *EMBO J.*, **20**, 3617–3622.
33. Lowe,T.M. and Eddy,S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, **283**, 1168–1171.
34. Schattner,P., Decatur,W.A., Davis,C.A., Ares,M. Jr, Fournier,M.J. and Lowe,T.M. (2004) Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the Saccharomyces cerevisiae genome. *Nucleic Acids Res.*, **32**, 4281–4296.
35. Maden,B.E. (2001) Mapping 2'-O-methyl groups in ribosomal RNA. *Methods*, **25**, 374–382.
36. Ofengand,J., Del,C.M. and Kaya,Y. (2001) Mapping pseudouridines in RNA molecules. *Methods*, **25**, 365–373.
37. Tomasevic,N. and Peculis,B.A. (2002) Xenopus LSm proteins bind U8 snoRNA via an internal evolutionarily conserved octamer sequence. *Mol. Cell Biol.*, **22**, 4101–4112.
38. Hokii,Y., Sasano,Y., Sato,M., Sakamoto,H., Sakata,K., Shingai,R., Taneda,A., Oka,S., Himeno,H., Muto,A. *et al.* (2010) A small nucleolar RNA functions in rRNA processing in Caenorhabditis elegans. *Nucleic Acids Res.*, **38**, 5909–5918.
39. Jarmolowski,A., Zagorski,J., Li,H.V. and Fournier,M.J. (1990) Identification of essential elements in U14 RNA of Saccharomyces cerevisiae. *EMBO J.*, **9**, 4503–4509.
40. Yuan,G., Klambt,C., Bachellerie,J.P., Brosius,J. and Huttenhofer,A. (2003) RNomics in Drosophila melanogaster: identification of 66 candidates for novel non-messenger RNAs. *Nucleic Acids Res.*, **31**, 2495–2507.
41. Samarsky,D.A., Schneider,G.S. and Fournier,M.J. (1996) An essential domain in Saccharomyces cerevisiae U14 snoRNA is absent in vertebrates, but conserved in other yeasts. *Nucleic Acids Res.*, **24**, 2059–2066.
42. Decatur,W.A. and Fournier,M.J. (2002) rRNA modifications and ribosome function. *Trends Biochem. Sci.*, **27**, 344–351.
43. Zemann,A., op de,B.A., Kiefmann,M., Brosius,J. and Schmitz,J. (2006) Evolution of small nucleolar RNAs in nematodes. *Nucleic Acids Res.*, **34**, 2676–2685.
44. Liang,X.H., Uliel,S., Hury,A., Barth,S., Doniger,T., Unger,R. and Michaeli,S. (2005) A genome-wide analysis of C/D and H/ACA-like small nucleolar RNAs in Trypanosoma brucei reveals a trypanosome-specific pattern of rRNA modification. *RNA.*, **11**, 619–645.
45. Ganot,P., Caizergues-Ferrer,M. and Kiss,T. (1997) The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. *Genes Dev.*, **11**, 941–956.
46. Liang,X.H., Liu,L. and Michaeli,S. (2001) Identification of the first trypanosome H/ACA RNA that guides pseudouridine formation on rRNA. *J. Biol. Chem.*, **276**, 40313–40318.
47. Cavaille,J. and Bachellerie,J.P. (1996) Processing of fibrillarin-associated snoRNAs from pre-mRNA introns: an exonucleolytic process exclusively directed by the common stem-box terminal structure. *Biochimie*, **78**, 443–456.
48. Liang,X.H., Hury,A., Hoze,E., Uliel,S., Myslyuk,I., Apatoff,A., Unger,R. and Michaeli,S. (2007) Genome-wide analysis of C/D and H/ACA-like small nucleolar RNAs in Leishmania major indicates conservation among trypanosomatids in the repertoire and in their rRNA targets. *Eukaryot. Cell*, **6**, 361–377.
49. Villa,T., Ceradini,F. and Bozzoni,I. (2000) Identification of a novel element required for processing of intron-encoded box C/D small nucleolar RNAs in Saccharomyces cerevisiae. *Mol. Cell Biol.*, **20**, 1311–1320.
50. Darzacq,X. and Kiss,T. (2000) Processing of intron-encoded box C/D small nucleolar RNAs lacking a 5',3'-terminal stem structure. *Mol. Cell Biol.*, **20**, 4522–4531.
51. Xu,Y., Liu,L., Lopez-Estrano,C. and Michaeli,S. (2001) Expression studies on clustered trypanosomatid box C/D small nucleolar RNAs. *J. Biol. Chem.*, **276**, 14289–14298.
52. Chen,C.L., Zhou,H., Liao,J.Y., Qu,L.H. and Amar,L. (2009) Genome-wide evolutionary analysis of the noncoding RNA genes and noncoding DNA of Paramecium tetraurelia. *RNA.*, **15**, 503–514.
53. Sweeney,R., Chen,L. and Yao,M.C. (1994) An rRNA variable region has an evolutionarily conserved essential role despite sequence divergence. *Mol. Cell Biol.*, **14**, 4203–4215.
54. Esguerra,J., Warringer,J. and Blomberg,A. (2008) Functional importance of individual rRNA 2'-O-ribose methylations revealed by high-resolution phenotyping. *RNA*, **14**, 649–656.
55. Baudin-Baillieu,A., Fabret,C., Liang,X.H., Piekna-Przybylska,D., Fournier,M.J. and Rousset,J.P. (2009) Nucleotide modifications in three functionally important regions of the Saccharomyces cerevisiae ribosome affect translation accuracy. *Nucleic Acids Res.*, **37**, 7665–7677.
56. Liang,X.H., Liu,Q. and Fournier,M.J. (2007) rRNA modifications in an intersubunit bridge of the ribosome strongly affect both ribosome biogenesis and activity. *Mol. Cell*, **28**, 965–977.
57. Liang,X.H., Liu,Q. and Fournier,M.J. (2009) Loss of rRNA modifications in the decoding center of the ribosome impairs translation and strongly delays pre-rRNA processing. *RNA.*, **15**, 1716–1728.
58. Ben-Shem,A., Jenner,L., Yusupova,G. and Yusupov,M. (2010) Crystal structure of the eukaryotic ribosome. *Science*, **330**, 1203–1209.
59. Rabl,J., Leibundgut,M., Ataide,S.F., Haag,A. and Ban,N. (2010) Crystal structure of the eukaryotic 40s ribosomal subunit in complex with initiation factor 1. *Science*, **331**, 730–736.
60. Pedersen,N., Hellung-Larsen,P. and Engberg,J. (1985) Small nuclear RNAs in the ciliate Tetrahymena. *Nucleic Acids Res.*, **13**, 4203–4224.
61. Yu,Y.T., Shu,M.D. and Steitz,J.A. (1998) Modifications of U2 snRNA are required for snRNP assembly and pre-mRNA splicing. *EMBO J.*, **17**, 5783–5795.
62. Bachellerie,J.P., Cavaille,J. and Huttenhofer,A. (2002) The expanding snoRNA world. *Biochimie*, **84**, 775–790.
63. Kishore,S. and Stamm,S. (2006) The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science*, **311**, 230–232.
64. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
65. Cannone,J.J., Subramanian,S., Schnare,M.N., Collett,J.R., D'Souza,L.M., Du,Y., Feng,B., Lin,N., Madabusi,L.V., Muller,K.M. *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC. Bioinformatics.*, **3**, 2.
66. Piekna-Przybylska,D., Decatur,W.A. and Fournier,M.J. (2008) The 3D rRNA modification maps database: with interactive tools for ribosome analysis. *Nucleic Acids Res.*, **36**, D178–D183.