Nucleic Acids Research, 2005, Vol. 33, No. 9 ***2781–2791***
doi:10.1093/nar/gki574

Unusual features of fibrillarin cDNA and gene structure in *Euglena gracilis*: evolutionary conservation of core proteins and structural predictions for methylation-guide box C/D snoRNPs throughout the domain Eucarya

Anthony G. Russell*, Yoh-ichi Watanabe, J. Michael Charette and Michael W. Gray

Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada B3H 1X5

Received March 21, 2005; Revised and Accepted April 25, 2005 DDBJ/EMBL/GenBank accession nos[+]

ABSTRACT

Box C/D ribonucleoprotein (RNP) particles mediate $O^{2'}$-methylation of rRNA and other cellular RNA species. In higher eukaryotic taxa, these RNPs are more complex than their archaeal counterparts, containing four core protein components (Snu13p, Nop56p, Nop58p and fibrillarin) compared with three in Archaea. This increase in complexity raises questions about the evolutionary emergence of the eukaryote-specific proteins and structural conservation in these RNPs throughout the eukaryotic domain. In protists, the primarily unicellular organisms comprising the bulk of eukaryotic diversity, the protein composition of box C/D RNPs has not yet been extensively explored. This study describes the complete gene, cDNA and protein sequences of the fibrillarin homolog from the protozoon *Euglena gracilis*, the first such information to be obtained for a nucleolus-localized protein in this organism. The *E.gracilis* fibrillarin gene contains a mixture of intron types exhibiting markedly different sizes. In contrast to most other *E.gracilis* mRNAs characterized to date, the fibrillarin mRNA lacks a spliced leader (SL) sequence. The predicted fibrillarin protein sequence itself is unusual in that it contains a glycine-lysine (GK)-rich domain at its N-terminus rather than the glycine-arginine-rich (GAR) domain found in most other eukaryotic fibrillarins. In an evolutionarily diverse collection of protists that includes *E.gracilis*, we have also identified putative homologs of the other core protein components of box C/D RNPs, thereby providing evidence that the protein composition seen in the higher eukaryotic complexes was established very early in eukaryotic cell evolution.

INTRODUCTION

Euglena gracilis, a flagellated protozoon, has served as a model laboratory organism for many years, yet little information is available about nuclear gene structure or modes of nuclear gene expression in this protist. This state of affairs is surprising considering that the few reported gene and cDNA sequences have revealed interesting and novel features of mRNA expression and processing in *Euglena*.

It appears, for example, that the majority of *Euglena* mRNA transcripts undergo *trans*-splicing (1). This post-transcriptional process adds a capped 28-nt leader sequence, the spliced leader (SL), to the 5′ ends of pre-mRNAs using a donor RNA, the SL RNA. SL sequences are a universal feature of the 5′ ends of mRNA transcripts in trypanosomatids (2,3), fellow members with the euglenids in the phylum Euglenozoa. SL sequences have also been found on mRNAs in nematodes (4), trematodes (5), chordates (6) and cnidarians (7), organisms very distantly related to the euglenids. In these latter organisms, *trans*-splicing occurs on pre-mRNAs that also undergo conventional spliceosomal *cis*-splicing.

*To whom correspondence should be addressed. Tel: +1 902 494 7035; Fax: +1 902 494 1355; Email: russella@dal.ca
Present address:
Yoh-ichi Watanabe, Department of Biomedical Chemistry, Graduate School of Medicine, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

[+] AF110181, AY925002, AY950656 and AY950657–AY950662

Trypanosomes contain very few *cis*-spliceosomal introns, the first identified being in a poly(A) polymerase gene (8). Most, if not all, protein-coding genes in trypanosomatids are first expressed as polycistronic precursor transcripts. Each individual protein-coding region requires *trans*-splicing and polyadenylation as underlying mechanisms for their liberation and maturation. Whether polycistronic mRNA transcription is a common gene expression strategy in *E.gracilis* is not known. Of note, several nucleus-encoded, chloroplast-targeted proteins in *Euglena* are post-translationally processed into protein subunits from polyprotein precursor molecules (9–11).

Unlike the situation in trypanosomes, protein-coding genes in *E.gracilis* appear to contain many introns, the first ones to be identified residing in nuclear genes encoding chloroplast-targeted proteins. Remarkably, the pre-mRNAs transcribed from these genes were predicted to contain only non-conventional (i.e. non-spliceosomal) introns. These intronic sequences were not present in the corresponding cDNA sequences obtained from the expressed genes, evidently having been removed via an undefined splicing mechanism. The only features common to all of these novel introns are repetitive sequence elements located at the splicing junctions and a potential for base-pairing interactions between sequences at the boundaries of each intron. The first *cis*-spliceosomal introns in *Euglena* were identified in a portion of the gene encoding fibrillarin, an abundant nucleolar protein (12). Additional spliceosomal introns were later identified in genes encoding the α, β and γ tubulins (13).

To date, most of the examined *Euglena* nuclear genes are ones that specify proteins that function in the chloroplast; in contrast, no genes encoding proteins localized to the *Euglena* nucleolus have as yet been completely sequenced and characterized. Fibrillarin is a core component of box C/D RNPs, the ribonucleoprotein complexes containing the box C/D RNAs that specify $O^{2'}$-methylation sites in the rRNAs (and other RNAs) of eukaryotic and archaeal organisms. The fibrillarin protein sequence is highly conserved between the domains Eucarya and Archaea, with the notable absence of the eukaryote-specific N-terminal glycine-arginine-rich region (GAR domain) in the archaeal fibrillarin homologs. The box C/D RNAs are defined by conserved sequence elements consisting of the C box, RUGAUGA, and the D box, CUGA, located near the 5′ and 3′ ends of these RNAs, respectively. Internal degenerate copies of these sequence elements, designated boxes C′ and D′, are also usually present in these RNAs.

The function of box C/D methylation-guide RNAs is conserved between archaeal and eukaryotic organisms; however, the archaeal box C/D RNPs have a simpler protein composition. In higher eukaryotes (i.e. animals and yeast), the box C/D RNPs contain the core proteins Nop56p, Nop58p and Snu13p (human '15.5 kDa') proteins in addition to fibrillarin (14–16). Intriguingly, the Snu13p protein is also found as a protein constituent of the spliceosome of higher eukaryotes (17). In contrast, archaeal box C/D RNPs contain a single Nop56p/Nop58p equivalent (Nop5p), fibrillarin and L7Ae (18,19), a multi-functional protein that is also a component of archaeal ribosomes, and likely a constituent of pseudouridine-guide RNPs as well (20,21). Higher eukaryotes have a distinct ribosomal L7a protein, which has also been designated L4 or L8A in *Saccharomyces cerevisiae* (22), with sequence relatedness to Snu13p/human 15.5 kDa protein and the archaeal L7Ae protein. The latter eukaryotic proteins all appear to be part of a larger gene family, which includes the eukaryotic box H/ACA RNP component, Nhp2p, and some other eukaryotic ribosomal proteins (23).

We have been examining the structure and function of pseudouridine-guide RNPs (24) and box C/D RNPs (A. G. Russell, unpublished data) in *E.gracilis*, in part because this protist is generally considered to be a relatively deep-branching eukaryote and a distant but specific evolutionary relative of the trypanosomatid protozoa. In light of this phylogenetic position, we expect that *Euglena* will provide insight into the evolution of these RNPs. Previously, we had obtained a partial cDNA and gene sequence for *Euglena* fibrillarin. In the work reported here, we present a complete analysis of the *E.gracilis* fibrillarin cDNA and gene structure, the first such information to be obtained for a nucleolus-localized protein in this organism. We also identify homologs of additional predicted protein components of box C/D RNPs in *Euglena* and in other protists. We relate these results to the evolution of the structure of box C/D RNPs and the spliceosome in the domain Eucarya.

MATERIALS AND METHODS

Oligonucleotides

Oligoribonucleotide P-1R (5′-AAUAAAGCGGCCGCGGAUCCAA-3′) was obtained from Dalton Chemical Laboratory Inc. (North York, ON, Canada). Oligodeoxyribonucleotides P-9 [5′-GTNTA(T/C)GCNGTNGA(A/G)TT(T/C)TCNCA-3′], P-10 [5′-GTNTA(T/C)GCNGTNGA(A/G)TT(T/C)AG(T/C)-CA-3′], P-11 [5′-GC(T/C)TG(A/G)TCNGG(T/C)TGNGC-NAC(A/G)TC-3′], P-19 [5′-CATGAAAATGCAATCAACC-ATCCC-3′], P-4 [5′-AATAAAGCGGCCGCGGATCCAA-3′], dTP-4 [5′-AATAAAGCGGGCCGCGGATCCAA(T_{16})-3′] and P-16 [5′-AATAAAGCGGGCCGCGGATCCAA(T_{16})(A/G/C)N-3′] were purchased from Gibco-BRL, ID Lab Biotechnology or MWG Biotech.

Fibrillarin cDNA sequence

E.gracilis DNA, RNA and poly(A)$^+$-enriched RNA were isolated using the methods described previously (12,25). Degenerate RT–PCR was performed to obtain a portion of the fibrillarin cDNA sequence. Random hexamer oligonucleotides served as primers for reverse transcription of *Euglena* poly(A)$^+$ RNA using Superscript II RT (Gibco BRL), followed by PCR amplification of the cDNA template with degenerate primer pairs P-9 (forward) and P-11 (reverse) or P-10 (forward) and P-11 (reverse). Primer design was based on conserved fibrillarin peptide sequences [VYAVEFS(Q/H) for P-9 and P-10, DVAQPDQA for P-11]. From the cDNA sequences obtained, fibrillarin-specific primers (sequences available on request) were then designed for use in both 5′ and 3′ rapid amplification of cDNA ends (RACE) experiments.

For 5′ RACE experiments, oligo-capping (26) was used to ligate the RNA oligonucleotide P-1R to the 5′ ends of mRNAs. Fibrillarin-specific oligo P-19 was then used for cDNA synthesis followed by PCR amplification of the template using oligo P-4 and fibrillarin cDNA-specific primers. PCR amplification was enhanced by the addition of 1.3 M betaine and 5% dimethylsulfoxide (v/v).

3′ RACE experiments were performed as described previously (25) using oligo P-16 for total cDNA synthesis from poly(A)$^+$-enriched RNA. This step was followed by PCR amplification of the product using fibrillarin cDNA-specific primers and oligo P-4. The above procedure was repeated except using oligo dTP-4 for cDNA synthesis to confirm the fibrillarin cDNA 3′ ends and poly(A) sites. PCR amplification of the fibrillarin cDNA template was also performed using primers specific for the extreme 5′ and 3′ ends of the fibrillarin cDNA to verify the linearity of our reported fibrillarin mRNA sequence and to confirm the sequence constituting the target annealing sites for the initial degenerate PCR primers and primers used in the RACE experiments.

Fibrillarin gene sequence

A library of *E.gracilis* genomic DNA (DNA kindly provided by David F. Spencer) was constructed in the BlueSTAR λ vector (Novagen) according to the manufacturer's protocol, except that *Euglena* genomic DNA fragments were not size-fractionated prior to cloning. The *in vitro* packaging reaction utilized the MaxPlax Packaging Extract (Epicentre Technology). Lambda clones containing fibrillarin gene fragments were detected by plaque hybridization with PCR probes generated from known fibrillarin cDNA sequences. Two unique fibrillarin gene-containing clones were detected and these were subsequently subcloned into the pBluescript® II KS(+) vector. The subcloned gene fragments comprised a 3.9 kb BamHI fragment corresponding to the 5′ end of the gene and a 1.5 kb EcoRI fragment containing the 3′ end. To obtain complete sequences of each subclone, primer walking strategies were used and nested deletions were created.

Internal sequence of the fibrillarin gene was determined using a PCR genomic walking strategy (25,27), with primers designed from known cDNA and genomic sequences as they became available. None of the primers was designed to anneal within any of the repeat regions (illustrated in Figure 2), minimizing the possibility of PCR-induced gaps in the fibrillarin genomic sequence reported here. In the case of some larger PCR products, subclones were prepared by shotgun cloning of restriction fragments generated by endonuclease CviJI digestion under relaxed conditions (28). For DNA templates that proved to be particularly difficult to sequence accurately using conventional protocols, a transcriptional sequencing method was employed (29). For transcriptional sequencing using modified RNA polymerases in these problematic regions, we used both CUGA3 and CUGA7 sequencing kits. All sequence data were analyzed and assembled with Sequencher Version 3.1.1 software.

Identification of homologs of box C/D RNP proteins

Sequence data used for the protein alignments were acquired from the National Center for Biotechnology Information (NCBI), The Institute for Genomic Research (TIGR) website at http://www.tigr.org and the Protist EST Program (PEP) (R.F. Watkins and M.W. Gray, unpublished data). Protein sequence alignments were generated using Clustal X Version 1.8 (30), and putative protein homologs were identified using both BLASTP and TBLASTN searches of the available sequence databases. Sequences obtained from genome databases were trimmed and translated using DNASIS V2.5.

RESULTS AND DISCUSSION

Euglena fibrillarin cDNA and protein structure

To obtain the complete *E.gracilis* fibrillarin cDNA sequence, degenerate oligonucleotide primers were designed based on conserved peptide motifs found in other known fibrillarin sequences. A combination of RT–PCR, 5′ RACE and 3′ RACE was then used to generate the complete cDNA sequence shown in Figure 1. Two cDNA species differing only in the length of

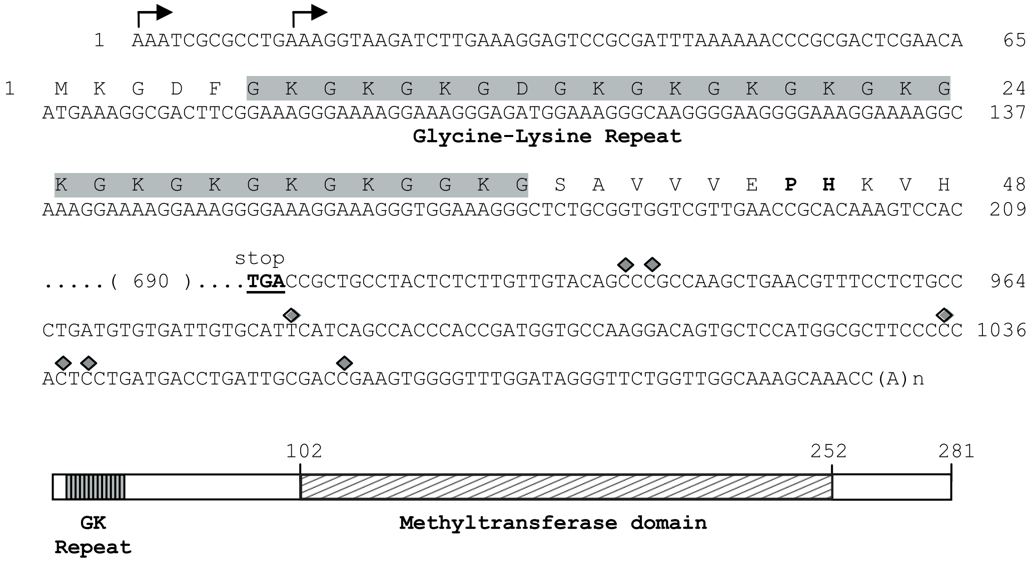

1 AAATCGCGCCTGAAAGGTAAGATCTTGAAAGGAGTCCGCGATTTAAAAAACCCGCGACTCGAACA 65
1 M K G D F G K G K G K G D G K G K G K G K G K G 24
ATGAAAGGCGACTTCGGAAAGGGAAAAGGAAAGGGAGATGGAAAGGGCAAGGGGAAGGGGAAAGGAAAAGGC 137
Glycine-Lysine Repeat
K G K G K G K G K G G K G S A V V V E P H K V H 48
AAAGGAAAAGGAAAGGGGAAAGGAAAGGGTGGAAAGGGCTCTGCGGTGGTCGTTGAACCGCACAAAGTCCAC 209
stop
.....(690)....TGACCGCTGCCTACTCTCTTGTTGTACAGCCCGCCAAGCTGAACGTTTCCTCTGCC 964
CTGATGTGTGATTGTGCATTCATCAGCCACCCACCGATGGTGCCAAGGACAGTGCTCCATGGCGCTTCCCCC 1036
ACTCCTGATGACCTGATTGCGACCGAAGTGGGGTTTGGATAGGGTTCTGGTTGGCAAAGCAAACC(A)n
102
252
281
GK
Repeat
Methyltransferase domain

Figure 1. Sequence and structure of the cDNA encoding *E.gracilis* fibrillarin (NCBI accession no. AF110181). Predicted positions of transcription initiation are indicated with arrows. The position of the translation stop codon is bolded and underlined. Additional detected sites of mRNA polyadenylation are shown as filled diamonds. One-letter amino acid abbreviations for the sequence at the amino terminal end of the protein are shown and the GK-rich region is highlighted in gray. The conserved PH dipeptide sequence discussed in the text is bolded. At the bottom of the sequence is a schematic diagram of the fibrillarin protein domain structure. The highly conserved methyltransferase domain, located between amino acid positions 102 and 252, is depicted by the hatched box and the GK repeat by the striped box.

their respective 5′ ends were detected in 5′ RACE experiments. The alternative 5′ ends may reflect different transcription initiation sites. Neither of the *Euglena* fibrillarin cDNAs contains a *trans*-SL sequence at its 5′ end. We previously identified a SL at the 5′ end of the *Euglena* Cbf5 mRNA, which also encodes a nucleolus-localized protein (25). It appears that in *E.gracilis* alternative 5′ end processing events discriminate among different pre-mRNAs with respect to addition (or not) of different SL sequences (1). The determinants of such discrimination are not yet known.

Based on the predicted translation start codon, the length of the 5′-untranslated region (5′-UTR) is 65 nt for the longest of the fibrillarin mRNA transcripts (Figure 1), similar in size to the 72 nt 5′-UTR of the *Euglena* Cbf5p-encoding mRNA. We also observed heterogeneity in the position of polyadenylation at the 3′ end of the cDNA, implying different possible polyadenylation sites for a fibrillarin transcript. Independently, from data generated by random sequencing of an *E.gracilis* cDNA library under the auspices of PEP, we assembled an expressed sequence tag (EST) cluster that contains a portion of the fibrillarin coding sequence. The sequence of this PEP EST cluster and the amino acid sequence inferred from it are identical to those reported here.

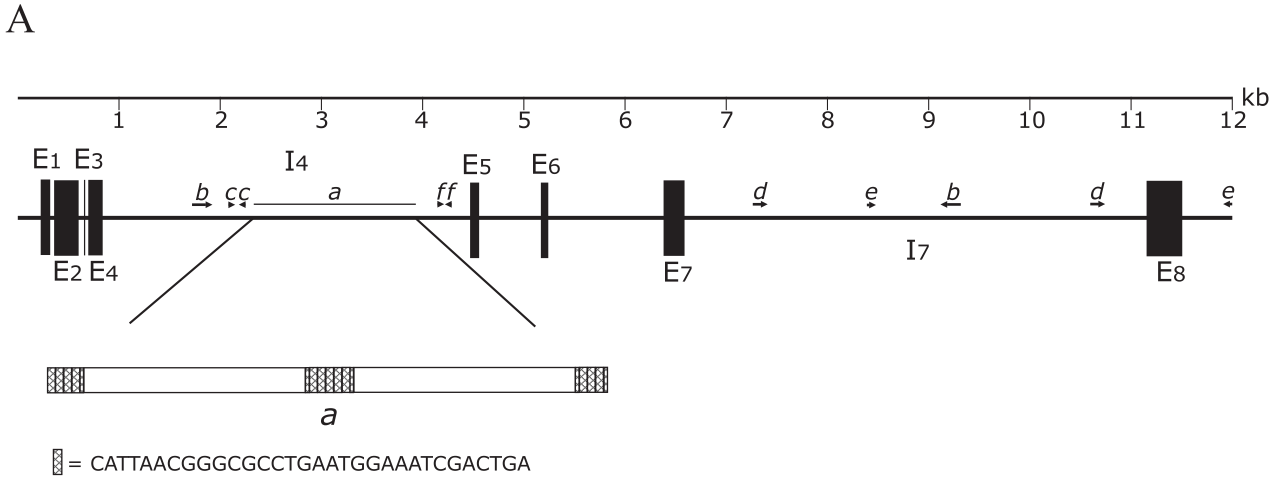
The fibrillarin cDNA sequence is predicted to encode a 281 amino acid, 30 kDa protein that contains the conserved methyltransferase domain found in all other fibrillarin homologs identified to date. Unlike most eukaryotic fibrillarins, which exhibit a distinctive N-terminal GAR domain, the N-terminal region in *Euglena* fibrillarin is instead glycine-lysine-rich (i.e. a ‘GK’ domain, Figure 1). A GK domain has also been reported in *Tetrahymena thermophila* (ciliate) fibrillarin (31); however, the overall primary structure of the GK-rich region is different in these two proteins. The *Euglena* protein has GK repeats and a very short GAR-like domain, which is bounded at its C-terminal extremity by a conserved PH dipeptide sequence (Figure 1, bold) that defines the end of fibrillarin GAR domains in other eukaryotes. In contrast, the *Tetrahymena* GAR-like domain exhibits limited sequence repetition, many interspersed proline residues and a KH dipeptide boundary sequence. While it is known that arginine residues in eukaryotic fibrillarin GAR domains can be dimethylated (32,33), the role of these modifications in fibrillarin function and the overall role of the GAR domain remain enigmatic. It will be interesting to determine whether there are amino acid modifications within the *Euglena* and *Tetrahymena* GK domains and what functional significance the R-to-K substitutions may have. It is noteworthy that whereas the human and *Xenopus* fibrillarin proteins are able to substitute functionally for the *S.cerevisiae* homolog (Nop1p) in a yeast knockout strain, the *Tetrahymena* protein is not (31).

Euglena fibrillarin gene structure

In an earlier study reporting the first spliceosomal introns in a *Euglena* species, we had characterized only the extreme 5′ end of the *E.gracilis* fibrillarin gene (12). Continuing this analysis, we have now obtained the complete fibrillarin gene sequence, using the fibrillarin cDNA sequence, to design PCR primers to amplify genomic DNA segments and to generate appropriate hybridization probes for screening a *Euglena* genomic DNA library. Two λ clones, each containing a genomic fragment corresponding to either the 5′ or 3′ end portion of the gene, as well as PCR products spanning the intervening genomic segments, indicate that the *E.gracilis* fibrillarin gene is ∼12 kb in size (Figure 2A). The gene contains seven introns of vastly different length. The first three introns, previously described (12), are small (44–68 bp) and of the conventional (‘Con’) spliceosomal-type. The remaining four introns, reported here, are much larger (620–4631 bp). Because a coding sequence for a box C/D small nucleolar (sno) RNA has been identified in an intron in some plant fibrillarin genes (34), we inspected the *Euglena* fibrillarin introns for any potential snoRNA-like sequences. This survey did not indicate any obvious snoRNA candidate sequences but did reveal several repeated sequences of unknown function (Figure 2A). Most of these repeats are imperfect; some are in an inverted orientation and so may form base-paired stem structures in the pre-mRNA, either within intron 4 (repeats *c* and *f*) or between introns 4 and 7 (repeat *b*). An exceptionally long repeat is found in region *a* of intron 4; this structure contains two direct copies of a large repeat sequence flanked by variable numbers of complete and partial copies of a second, short repeat element.

An examination of the sequences of the seven fibrillarin introns indicates that six are conventional spliceosomal introns containing characteristic GT ... AG boundary sequences (Figure 2B). Intron 5, however, contains GA instead of GT at the 5′ intron boundary and the sequences at the intron–exon junctions (boxed) are direct repeats, making the precise positions of intron excision and exon ligation uncertain. These unusual properties were previously observed in the so-called intermediate (‘Int’) type introns identified in tubulin genes in *E.gracilis* (13). Adjacent to the repetitive sequence elements containing the splice sites are complementary intronic sequences (Figure 2B, underlined) capable of forming a 10 bp stem structure that would bring the 5′ and 3′ ends of intron 5 into close proximity. Base-pairing potentials of similar length and at similar relative intron positions were also observed in the Int-type introns and in the non-conventional (‘Non’) introns (which do not possess similarity to GT ... AG intron boundary sequences) found in the α and β tubulin genes. Significantly, as seen in the *Euglena* fibrillarin gene, a single tubulin gene may contain a mixture of intron types. These combined observations raise the question of whether some components of the *E.gracilis* spliceosome might be utilized during the removal of Int-type introns. As we previously noted for conventional spliceosomal introns in *Euglena* (12), 5′ exon–intron junctions have the potential to base pair with the U1 snRNA. If the Int-type introns are in fact processed by *Euglena* spliceosomal components, U1 snRNA may not participate in their removal because these introns do not exhibit the same base-pairing potential evident in the conventional spliceosomal introns.

A few other cases of removal of non-conventional introns from mRNAs have been reported. In yeast, an intron is removed from the *HAC1* mRNA by a spliceosome-independent mechanism that utilizes a multi-functional endoribonuclease, Ire1p, for splice site cleavage (35). Following intron excision, exons are joined by tRNA ligase (36). In three archaeal species, a conserved intron has been identified in the gene encoding the archaeal homolog of Cbf5p (37). In this case, the intron is likely removed from the Cbf5 mRNA



B

Intron	Type	Size	Boundary
1	Con	46	GAAA/**gt**atgaaaacttcag......aaattattccattttt**ag**/GGGA
2	Con	68	AGAA/**gt**atgatctgcacct......ttcaaccgcccttgc**ag**/ACG
3	Con	44	ACG/**gt**gtgttgtcttctc......ctccttcaacttcat**ag**/ACGG
4	Con	3636	CTCG/**gt**ctgctctaccttg......gaattcatcctctgc**ag**/GTGC
5	Int	620	TCAG **g**atcagagtgttgct......tgcaacaatctaggg**ag** GACG
6	Con	1141	TGCG/**gt**gtgccctgttcag......tcgtcatgcttcctc**ag**/GATG
7	Con	4631	CAAG/**gt**gtccctgccaggg......tccactgctaatttc**ag**/GCTC
	Conventional Consensus		R/**gt**rtg ... y**ag**/R

Figure 2. Structure of the *E.gracilis* fibrillarin gene (NCBI accession no. AY950662). (**A**) Composite structure of the ~12 kb fibrillarin gene drawn approximately to scale. Exons (E) are illustrated as black boxes and introns (I) as black lines. Segments containing imperfect repeated sequences are indicated using directional arrows and the repeated sequence pairs have lowercase letter (italics) designations. Region *a* in intron 4, which contains numerous repeat elements, is also shown in expanded form. Small cross-hatched boxes denote segments representing all or part of a small imperfect repeat unit. The larger boxes each contain a single copy of a long imperfect repeat sequence. (**B**) Structure of fibrillarin introns. The sizes as well as the actual nucleotide sequences at the exon–intron junctions of each intron are shown. Exon sequences are shown in uppercase letters, introns in lowercase lettering. Introns are classified as conventional spliceosomal-type (Con) if they contain canonical intron boundary nucleotides (indicated in bold). At its exon–intron boundaries, intron 5 (designated an intermediate type, Int) contains repeated sequences (boxed), with regions of sequence complementarity underlined. The consensus sequence of nucleotide positions at the exon–intron boundaries of the conventional introns is indicated at the bottom of Figure 2B.

via the intron excision-endonuclease/ligase mechanism employed in the processing of intron-containing archaeal pre-tRNAs. This inference is based on the observation that the intron–exon boundaries are predicted to fold into the requisite structure for this splicing pathway. At present, insufficient information is available to allow us to deduce whether related mechanisms may be used in *Euglena* for the removal of non-conventional or Int-type introns. We note that, as yet, there are no clear orthologs of Ire1p in the *Euglena* or other protist EST databases, let alone predicted protein sequences exhibiting significant amino acid similarity specifically to the ribonuclease domain of Ire1p.

The relationship between the different intron types in *Euglena* is perplexing given that several examples are now known of genes containing a mixture of intron types. Non-conventional introns predominate in genes encoding chloroplast-targeted proteins and are prevalent in the *gapC* gene, which encodes a cytoplasmic GAPDH protein of apparent eubacterial origin (38). Because these genes were likely integrated into the *Euglena* nuclear genome via an endosymbiotic source (39), it is possible that the different intron types have different evolutionary origins. In this regard, the Int-type introns are particularly pertinent because they may represent a transition state between non-conventional and spliceosomal introns. We note that intron 1 of the *gapC* gene appears to be another example of an Int-type intron because it has potential GT . . . AG intron boundary sequences in addition to the capacity for base pairing between the intron sequences adjacent to the 5′ and 3′ splice sites, a conserved feature of the non-conventional type of introns. A more comprehensive investigation of intron types in additional *Euglena* genes may shed light on these intriguing possibilities.

Identification of homologs of box C/D RNP proteins in *E.gracilis* and other protists

As mentioned earlier, box C/D RNPs of higher eukaryotes are more complex than their archaeal counterparts. The most

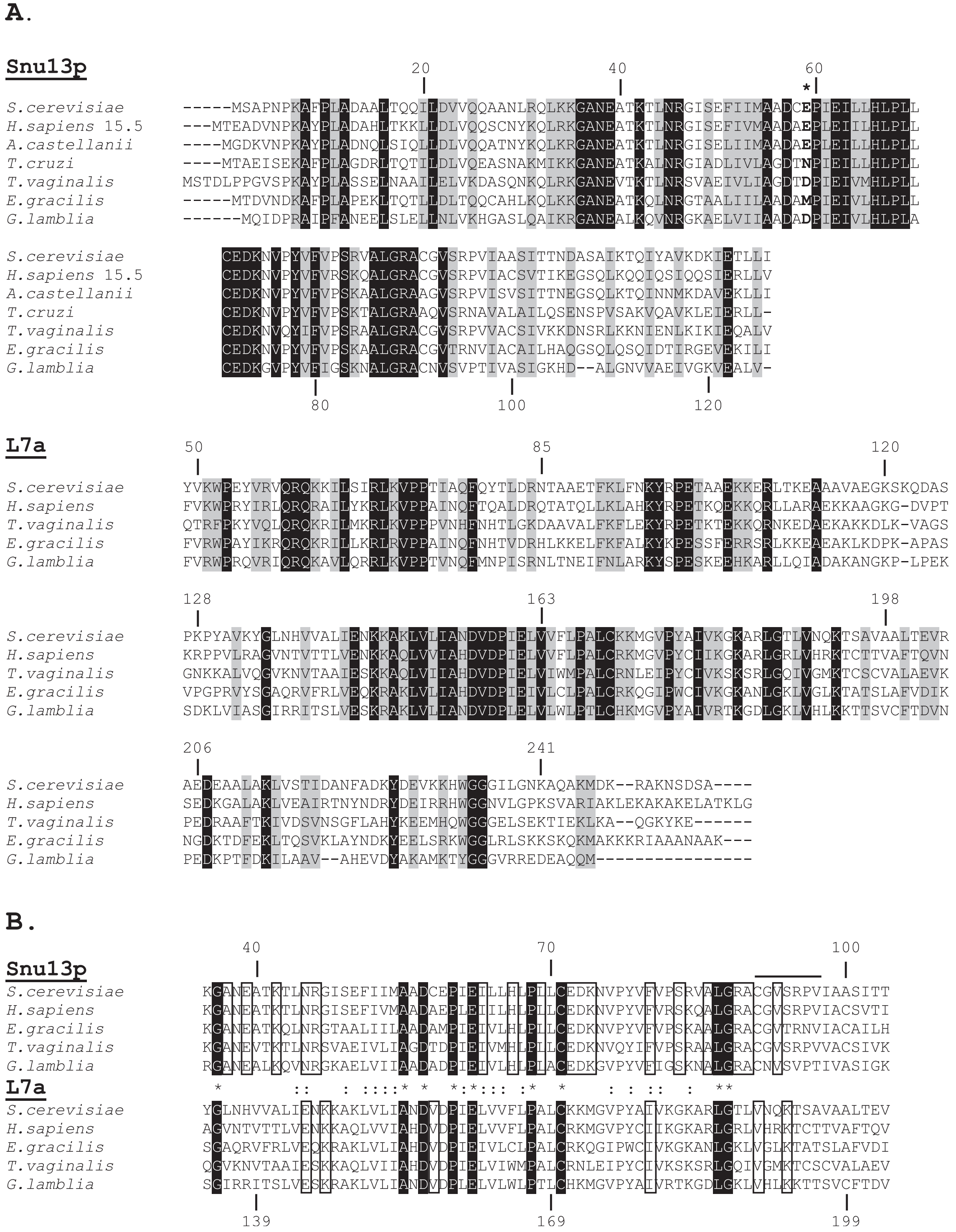

A.
Snu13p
20 40 60
*
S.cerevisiae -----MSAPNPKAFPLADAALTQQILDVVQQAANLRQLKKGANEATKTLNRGISEFIIMAADCEPIEILLHLPLL
H.sapiens 15.5 ---MTEADVNPKAYPLADAHLTKKLLDLVQQSCNYKQLRKGANEATKTLNRGISEFIVMAADAEPLEIILHLPLL
A.castellanii ----MGDKVNPKAYPLADNQLSIQLLDLVQQATNYKQLKRGANEATKTLNRGISELIIMAADAEPLEILLHLPLL
T.cruzi ----MTAEISEKAFPLAGDRLTQTILDLVQEASNAKMIKKGANEATKALNRGIADLIVLAGDTNPIEILLHLPLL
T.vaginalis MSTDLPPGVSPKAYPLASSELNAAILELVKDASQNKQLRKGANEVTKTLNRSVAEIVLIAGDTDPIEIVMHLPLL
E.gracilis -----MTDVNDKAFPLAPEKLTQTLLDLTQQCAHLKQLKKGANEATKQLNRGTAALIILAADAMPIEIVLHLPLL
G.lamblia ------MQIDPRAIPFANEELSLELLNLVKHGASLQAIKRGANEALKQVNRGKAELVIIAADADPIEIVLHLPLA
S.cerevisiae CEDKNVPYVFVPSRVALGRACGVSRPVIAASITTNDASAIKTQIYAVKDKIETLLI
H.sapiens 15.5 CEDKNVPYVFVRSKQALGRACGVSRPVIACSVTIKEGSQLKQQIQSIQQSIERLLV
A.castellanii CEDKNVPYVFVPSKAALGRAAGVSRPVISVSITTNEGSQLKTQINNMKDAVEKLLI
T.cruzi CEDKNVPYVFVPSKTALGRAAQVSRNAVALAILQSENSPVSAKVQAVKLEIERLL-
T.vaginalis CEDKNVQYIFVPSRAALGRACGVSRPVVACSIVKKDNSRLKKNIENLKIKIEQALV
E.gracilis CEDKNVPYVFVPSKAALGRACGVTRNVIACAILHAQGSQLQSQIDTIRGEVEKILI
G.lamblia CEDKGVPYVFIGSKNALGRACNVSVPTIVASIGKHD--ALGNVVAEIVGKVEALV-
80 100 120
L7a
50 85 120
S.cerevisiae YVKWPEYVRVQRQKKILSIRLKVPPTIAQFQYTLDRNTAAETFKLFNKYRPETAAEKKERLTKEAAAVAEGKSKQDAS
H.sapiens FVKWPRYIRLQRQRAILYKRLKVPPAINQFTQALDRQTATQLLKLAHKYRPETKQEKKQRLLARAEKKAAGKG-DVPT
T.vaginalis QTRFPKYVQLQRQKRILMKRLKVPPPVNHFNHTLGKDAAVALFKFLEKYRPETKTEKKQRNKEDAEKAKKDLK-VAGS
E.gracilis FVRWPAYIKRQRQKRILLKRLRVPPAINQFNHTVDRHLKKELFKFALKYKPESSFERRSRLKKEAEAKLKDPK-APAS
G.lamblia FVRWPRQVRIQRQKAVLQRRLKVPPTVNQFMNPISRNLTNEIFNLARKYSPESKEEHKARLLQIADAKANGKP-LPEK
128 163 198
S.cerevisiae PKPYAVKYGLNHVVALIENKKAKLVLIANDVDPIELVVFLPALCKKMGVPYAIVKGKARLGTLVNQKTSAVAALTEVR
H.sapiens KRPPVLRAGVNTVTTLVENKKAQLVVIAHDVDPIELVVFLPALCRKMGVPYCIIKGKARLGRLVHRKTCTTVAFTQVN
T.vaginalis GNKKALVQGVKNVTAAIESKKAQLVIIAHDVDPIELVIWMPALCRNLEIPYCIVKSKSRLGQIVGMKTCSCVALAEVK
E.gracilis VPGPRVYSGAQRVFRLVEQKRAKLVLIAHDVDPIEIVLCLPALCRKQGIPWCIVKGKANLGKLVGLKTATSLAFVDIK
G.lamblia SDKLVIASGIRRITSLVESKRAKLVLIANDVDPLELVLWLPTLCHKMGVPYAIVRTKGDLGKLVHLKKTTSVCFTDVN
206 241
S.cerevisiae AEDEAALAKLVSTIDANFADKYDEVKKHWGGGILGNKAQAKMDK--RAKNSDSA----
H.sapiens SEDKGALAKLVEAIRTNYNDRYDEIRRHWGGNVLGPKSVARIAKLEKAKAKELATKLG
T.vaginalis PEDRAAFTKIVDSVNSGFLAHYKEEMHQWGGGELSEKTIEKLKA--QGKYKE------
E.gracilis NGDKTDFEKLTQSVKLAYNDKYEELSRKWGGLRLSKKSKQKMAKKKRIAAANAAK---
G.lamblia PEDKPTFDKILAAV--AHEVDYAKAMKTYGGGVRREDEAQQM----------------
B.
40 70 100
Snu13p
S.cerevisiae KGANEATKTLNRGISEFIIMAADCEPIEILLHLPLLCEDKNVPYVFVPSRVALGRACGVSRPVIAASITT
H.sapiens KGANEATKTLNRGISEFIVMAADAEPLEIILHLPLLCEDKNVPYVFVRSKQALGRACGVSRPVIACSVTI
E.gracilis KGANEATKQLNRGTAALIILAADAMPIEIVLHLPLLCEDKNVPYVFVPSKAALGRACGVTRNVIACAILH
T.vaginalis KGANEVTKTLNRSVAEIVLIAGDTDPIEIVMHLPLLCEDKNVQYIFVPSRAALGRACGVSRPVVACSIVK
G.lamblia RGANEALKQVNRGKAELVIIAADADPIEIVLHLPLACEDKGVPYVFIGSKNALGRACNVSVPTIVASIGK
L7a
* :: : ::::* * *:*::: :* * : : :: : **
S.cerevisiae YGLNHVVALIENKKAKLVLIANDVDPIELVVFLPALCKKMGVPYAIVKGKARLGTLVNQKTSAVAALTEV
H.sapiens AGVNTVTTLVENKKAQLVVIAHDVDPIELVVFLPALCRKMGVPYCIIKGKARLGRLVHRKTCTTVAFTQV
E.gracilis SGAQRVFRLVEQKRAKLVLIAHDVDPIEIVLCLPALCRKQGIPWCIVKGKANLGKLVGLKTATSLAFVDI
T.vaginalis QGVKNVTAAIESKKAQLVIIAHDVDPIELVIWMPALCRNLEIPYCIVKSKSRLGQIVGMKTCSCVALAEV
G.lamblia SGIRRITSLVESKRAKLVLIANDVDPLELVLWLPTLCHKMGVPYAIVRTKGDLGKLVHLKKTTSVCFTDV
139 169 199

Figure 3. (**A**) Alignments of homologous eukaryotic Snu13p and L7a protein sequences. Amino acid positions in both proteins are numbered relative to the corresponding *S.cerevisiae* sequences. Positions of amino acid identity are indicated as white letters on a black background, strongly conserved positions (as defined by the Gonnet Pam250 matrix) are highlighted in gray. The variable N-terminal lysine-rich portion of L7a has been omitted. The asterisk marks the position of *S.cerevisiae* Snu13p amino acid residue Glu59 (discussed in the text). (**B**) Composite alignment of Snu13p and L7a sequences showing the region of amino acid similarity shared between the paralogs. Positions of amino acid identity are displayed as white letters on a black background and are also indicated with asterisks. Strongly conserved positions are marked by colons. Amino acids that are uniquely conserved in only one paralog (L7a or Snu13p only) are indicated with open boxes. Amino acid positions discussed in the text are highlighted by the horizontal line. Amino acid numbering corresponds to part A. Sources of the Snu13p sequences are: *E.gracilis* (NCBI accession no. AY950657); *A.castellanii* (AY950656); *S.cerevisiae* (NP_010888); *Homo sapiens* 15.5 kDa protein (AF155235); *G.lamblia* (EAA41217); *T.cruzi* [AAP49574 and The *T.cruzi* Genome Database (www.tigr.org/tdb/e2k1/tca1/), ID: 7667.m00012]; and *T.vaginalis* [The *T.vaginalis* Genome Sequencing project (www.tigr.org/tdb/e2k1/tva1/), ID: 41877.m00079]. Sources of the L7a sequences are: *E.gracilis* (accession no. AY925002), *S.cerevisiae* L7A-1 (AAB65045), *H.sapiens* L7a (CAI12832), *G.lamblia* L7a (AACB01000019) and *T.vaginalis* [The *T.vaginalis* Genome Sequencing Project (www.tigr.org/tdb/e2k1/tva1/), ID: 43310.m00099].

parsimonious explanation for these differences is that sometime during eukaryotic cell evolution, gene duplication events led to the emergence of these additional protein constituents, with the resultant paralogous proteins retaining or acquiring specialized functions. It was, therefore, of interest to search for homologs of these proteins in a more diverse collection of eukaryotic organisms, specifically in deeply branching organisms, such as *E.gracilis* and *Giardia lamblia*. Our 2-fold objective was first to determine the extent to which these proteins are conserved throughout the eukaryotic domain, and then to predict some of the structural properties of the box C/D RNPs in the organisms in question. These are important first steps toward understanding and comparing the detailed functional mechanism of eukaryotic box C/D RNPs in the $O^{2'}$-methylation reaction in different eukaryotes.

Snu13p homologs in protists and a K-turn motif in the *E.gracilis* U4 snRNA

We first examined the phylogenetic distribution of Snu13p (15.5 kDa protein), as it is a primary RNA-binding protein of higher eukaryotic box C/D RNPs (40). The analogous RNA-binding function is performed by L7Ae in archaeal box C/D RNPs (41). Using the entire *S.cerevisiae* Snu13p sequence or conserved peptide sequences therein as query in a BLAST search of the *E.gracilis* and *Acanthamoeba castellanii* (amoebozoan) PEP EST databases, we identified putative homologs of the paralogous Snu13p and L7a proteins. Additionally, searches of the complete or partial genome databases of other relevant eukaryotic taxa revealed candidate Snu13p sequences in addition to L7a. Some of these Snu13p and L7a sequences have been aligned and are shown in Figure 3A. Both proteins show a high degree of amino acid sequence conservation across diverse eukaryotic taxa. The conservation of amino acid sequence and the ability to clearly identify putative homologs of both proteins in distantly related eukaryotic organisms indicates that some of the inferred gene duplication events that generated the additional eukaryotic proteins likely occurred very early in eukaryotic cell evolution, especially considering the presence of both L7a and Snu13p homologs in other primitive and putatively deep-branching protists, such as *G.lamblia* (diplomonad) and *Trichomonas vaginalis* (parabasalian). The high degree of amino acid sequence conservation between these Snu13p homologs indicates that they are most likely core components of box C/D RNP complexes in most, if not all, eukaryotes.

In aligning the eukaryotic Snu13p-like sequences, we noted that residue Glu59 (*S.cerevisiae* amino acid sequence numbering), which has been shown to contribute to the RNA-binding specificity of the yeast protein (42), is not evolutionarily conserved when taxa outside of the opisthokont (animals+fungi) and amoebozoan lineages are considered (Figure 3A, asterisk). In the *S.cerevisiae* protein, a Glu59Ala substitution created a lethal yeast phenotype (42), with this single-amino-acid substitution differentially disrupting the binding of the mutant protein to the U4 snRNA more than it affects its interaction with the U3 box C/D snoRNA. The Snu13p protein binds to a kink-turn (K-turn) motif, an RNA structural element that is present in both the U4 5′ stem–loop structure and in box C/D RNAs (including U3) of higher eukaryotes (17,43,44). Using a partial sequence for the *E.gracilis* U4 snRNA (12) and additional sequence obtained by 3′ RACE for this RNA (J. M. Charette, unpublished data), we examined the potential for the 5′ stem–loop in *E.gracilis* U4 snRNA to fold into a K-turn motif. The predicted secondary structure of this portion of the *E.gracilis* U4 snRNA is very similar to that of its human or yeast counterparts (Figure 4). The 5+2 internal bulge containing the sheared G•A pairs, critical Snu13p recognition elements in the K-turn motif, are present. In addition, the length of base pairing within the adjacent extended stem and the location of the bulged A nucleotide are identical.

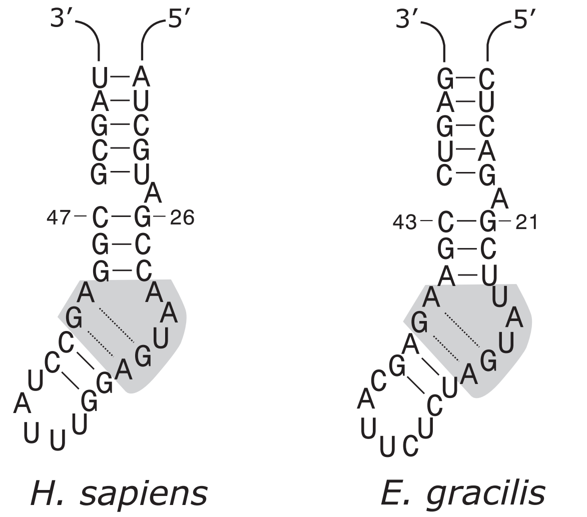


Figure 4. Structures of the K-turn motifs present in the 5′ stem-loops of the U4 snRNAs from human (*H.sapiens*) and *E.gracilis*. Nucleotide positions within each RNA are indicated. The regions in gray highlight the asymmetric 5+2 internal bulges containing the non-canonical sheared G•A pairs.

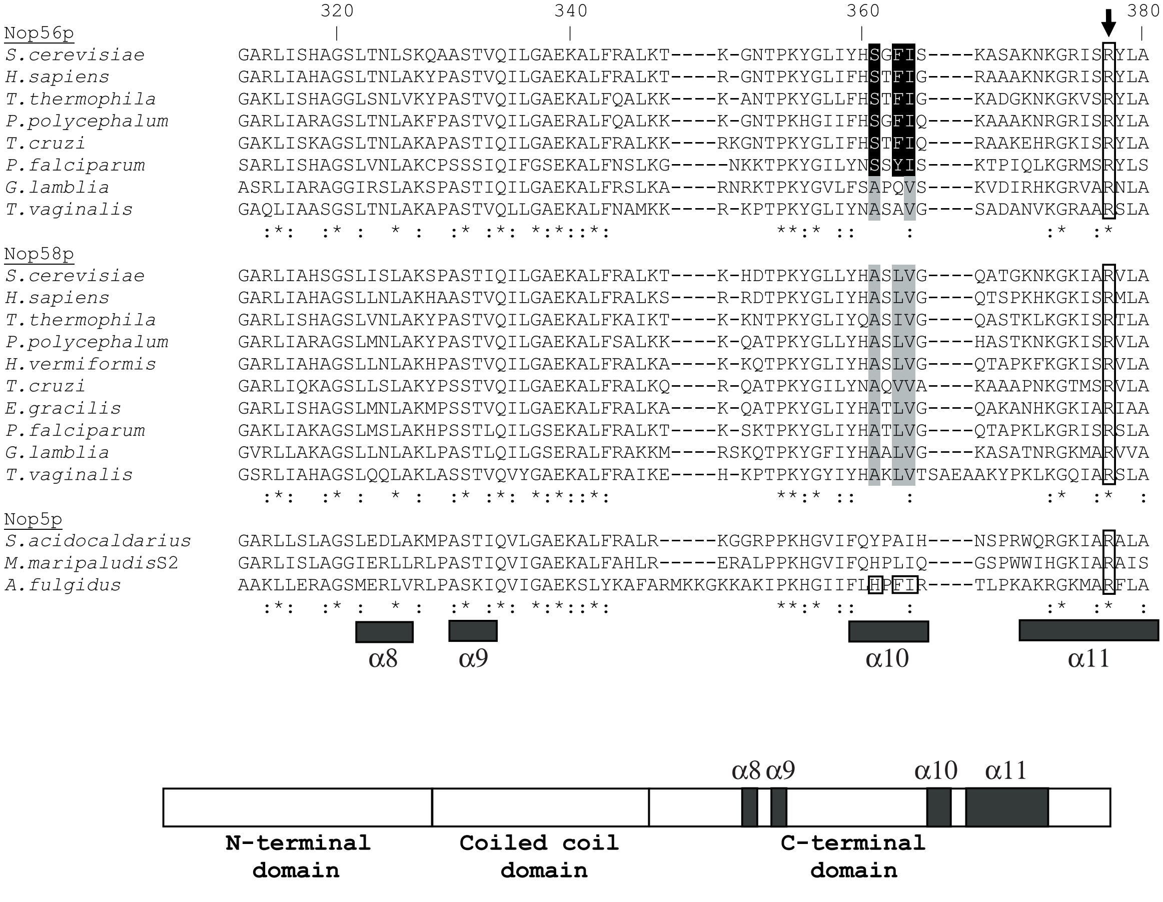
Conservation of the K-turn structural element in *E.gracilis* U4 snRNA strongly suggests that the *E.gracilis* Snu13p homolog interacts with U4 snRNA in this region and, thus, is a component of the *E.gracilis* spliceosome, despite the fact that the *E.gracilis* Snu13p homolog has a methionine residue at the position corresponding to *S.cerevisiae* Glu59 (Figure 3A). In the 3D structure of the complex between the human 15.5 kDa protein and U4 snRNA (43), Glu59 (*S.cerevisiae* numbering) is located within a tight turn in the polypeptide chain where it makes a peptide backbone hydrogen bond contact to U31 of the K-turn motif in the U4 snRNA.

Considering the variation in amino acid residues found at this position in the protist Snu13p-like sequences, it will be interesting to examine the structure of the relevant U4 snRNAs, once they have been identified in other eukaryotic taxa, for their ability to form the K-turn motif. It will also be important to determine binding specificities of some of these Snu13p proteins for box C/D RNAs, both U3 and methylation-guide snoRNAs, and also for U4 snRNAs, to determine how these amino acid variations at position 59 influence the protist protein structures and binding properties. In this regard, we note that alanine is not present at position 59 in any of the protist sequences we have examined. In considering these observations, particularly the evolutionary conservation of the K-turn motif in the *E.gracilis* U4, we infer that Snu13p (15.5 kDa protein) participation in spliceosome function was established prior to the divergence of *E.gracilis* from the other main eukaryotic lineages.

The region of amino acid similarity shared between the Snu13p and L7a paralogs, depicted in Figure 3B, is within the most highly conserved portion of each of the proteins (Figure 3A). This region of similarity is adjacent to residues within Snu13p (Figure 3B, overlined) that, like Glu59, are also thought to play a role in defining the RNA-binding specificity of this protein [(45); B. Brown II, personal communication] Currently, no detailed structural information is available for any eukaryotic L7a protein and its RNA-binding site (or motif) within the large subunit ribosomal RNA has not been defined. However, it does not appear that the human L7a protein can bind to a kink-turn motif (46). Differences in the amino acid sequence of L7a compared with Snu13p in the abovementioned region (Figure 3B, overlined) are likely to contribute to the different RNA-binding properties displayed by the L7a protein.

Nop56p and Nop58p homologs are both present together throughout Eucarya

An examination of PEP EST databases for *Hartmannella vermiformis* and *Physarum polycephalum* (two amoebozoons) as well as *E.gracilis* revealed cDNAs predicted to encode either or both Nop56p and Nop58p homologs. In a survey of protist taxa for which genomic information is currently available, we found that both Nop56 and Nop58 genes are simultaneously present in the genomes of these protists. An alignment of the most highly conserved portion of the two proteins, corresponding to *S.cerevisiae* Nop56p residues 312–380 and containing part of the predicted RNA-binding domain (47), is shown in Figure 5. The amino acid residues at positions 361, 363 and 364 discriminate the sequence of Nop56p from that of Nop58p in most eukaryotic taxa. In the



```
                           320                 340                      360                   ↓ 380
Nop56p                      |                   |                        |                       |
S.cerevisiae        GARLISHAGSLTNLSKQAASTVQILGAEKALFRALKT----K-GNTPKYGLIYHSGFIS----KASAKNKGRISRYLA
H.sapiens           GARLIAHAGSLTNLAKYPASTVQILGAEKALFRALKT----R-GNTPKYGLIFHSTFIG----RAAAKNKGRISRYLA
T.thermophila       GAKLISHAGGLSNLVKYPASTVQILGAEKALFQALKK----K-ANTPKYGLLFHSTFIG----KADGKNKGKVSRYLA
P.polycephalum      GARLIARAGSLTNLAKFPASTVQILGAERALFQALKK----K-GNTPKHGIIFHSGFIQ----KAAAKNRGRISRYLA
T.cruzi             GAKLISKAGSLTNLAKAPASTIQILGAEKALFRALKK----RKGNTPKYGLIFHSTFIQ----RAAKEHRGKISRYLA
P.falciparum        SARLISHAGSLVNLAKCPSSSIQIFGSEKALFNSLKG-----NKKTPKYGILYNSSYIS----KTPIQLKGRMSRYLS
G.lamblia           ASRLIARAGGIRSLAKSPASTIQILGAEKALFRSLKA----RNRKTPKYGVLFSAPQVS----KVDIRHKGRVARNLA
T.vaginalis         GAQLIAASGSLTNLAKAPASTVQLLGAEKALFNAMKK----R-KPTPKYGLIYNASAVG----SADANVKGRAARSLA
                      :*:  :* :  * :  :* :*: *:*::*:              **:* ::    :           :*  :*

Nop58p
S.cerevisiae        GARLIAHSGSLISLAKSPASTIQILGAEKALFRALKT----K-HDTPKYGLLYHASLVG----QATGKNKGKIARVLA
H.sapiens           GARLIAHAGSLLNLAKHAASTVQILGAEKALFRALKS----R-RDTPKYGLIYHASLVG----QTSPKHKGKISRMLA
T.thermophila       GARLISHAGSLVNLAKYPASTVQILGAEKALFKAIKT----K-KNTPKYGLIYQASIVG----QASTKLKGKISRTLA
P.polycephalum      GARLIARAGSLMNLAKYPASTVQILGAEKALFSALKK----K-QATPKYGLLYHASLVG----HASTKNKGKISRVLA
H.vermiformis       GARLIAHAGSLLNLAKHPASTVQILGAEKALFRALKA----K-KQTPKYGLIYHASLVG----QTAPKFKGKISRVLA
T.cruzi             GARLIQKAGSLLSLAKYPSSTVQILGAEKALFRALKQ----R-QATPKYGILYNAQVVA----KAAAPNKGTMSRVLA
E.gracilis          GARLISHAGSLMNLAKMPSSTVQILGAEKALFRALKA----K-QATPKYGLIYHATLVG----QAKANHKGKIARIAA
P.falciparum        GAKLIAKAGSLMSLAKHPSSTLQILGSEKALFRALKT----K-SKTPKYGLIYHATLVG----QTAPKLKGRISRSLA
G.lamblia           GVRLLAKAGSLLNLAKLPASTLQILGSERALFRAKKM----RSKQTPKYGFIYHAALVG----KASATNRGKMARVVA
T.vaginalis         GSRLIAHAGSLQQLAKLASSTVQVYGAEKALFRAIKE----H-KPTPKYGYIYHAKLVTSAEAAKYPKLKGQIARSLA
                      :*:  :* :  * :  :* :*: *:*::*:              **:* ::    :           :*  :*  :

Nop5p
S.acidocaldarius    GARLLSLAGSLEDLAKMPASTIQVLGAEKALFRALR-----KGGRPPKHGVIFQYPAIH----NSPRWQRGKIARALA
M.maripaludisS2     GARLISLAGGIERLLRLPASTIQVIGAEKALFAHLR-----ERALPPKHGVIFQHPLIQ----GSPWWIHGKIARAIS
A.fulgidus          AAKLLERAGSMERLVRLPASKIQVIGAEKSLYKAFARMKKGKKAKIPKHGIIFLHPFIR----TLPKAKRGKMARFLA
                      :*:  :* :  * :  :* :*: *:*::*:              **:* ::    :           :*  :*  :
```

Figure 5. Clustal X protein sequence alignments of eukaryotic Nop56p and Nop58p paralogs. The most highly conserved portion of each protein (corresponding to *S.cerevisiae* Nop56p amino acid positions 312–380) is shown. The archaeal Nop5p (Nop56p/58p) protein homologs from *Sulfolobus acidocaldarius* (AF201092), *Methanococcus maripaludis* S2 (CAF30152) and *A.fulgidus* (NP_070912) are also included in the alignment. Identical amino acid positions for all shown sequences are indicated with asterisks and positions of strong conservation (Gonnet Pam250 matrix) with colons. Amino acids at *S.cerevisiae* positions 361, 363 and 364 that are discussed in the text are shown as either white letters on a black background (Nop56p) or are shaded in gray (Nop58p). The corresponding positions in the *A.fulgidus* sequence are boxed. A highly conserved arginine residue, *S.cerevisiae* Arg377 (*A.fulgidus* Arg224), is boxed in all sequences and this position is also indicated by the arrow. The locations of some of the α-helical secondary structure elements within the *A.fulgidus* protein, as determined in (47), are indicated below the sequences. A schematic of the domain structure of the *A.fulgidus* protein is illustrated at the bottom. Sources of the eukaryotic sequences are *S.cerevisiae* Nop56p (Q12460), Nop58p (NP_014955); *H.sapiens* Nop56p (NP_006383), Nop58p (NP_057018); *P.polycephalum* Nop56p (AY950661), Nop58p (AY950660); *H.vermiformis* Nop58p (AY950658); *E.gracilis* Nop58p (AY950659); *T.cruzi* Nop56p (ID: 6845.m00001), Nop58p (ID: 7667.m00023), The *T.cruzi* Genome Database (www.tigr.org/tdb/e2k1/tca1/); *T.vaginalis* Nop56p (ID: 51612.m00053), Nop58p (ID: 43306.m00059), The *T.vaginalis* Genome Sequencing project (www.tigr.org/tdb/e2k1/tva1/); *T.thermophila* Nop56p (ID: 7902), Nop58p (ID: 11339), The *T.thermophila* Genome Sequencing project (www.tigr.org/tdb/e2k1/ttg/); *Plasmodium falciparum* Nop56p (NP_701051), Nop58p (NP_700559); *G.lamblia* Nop56p (EAA42149), Nop58p (EAA38450).

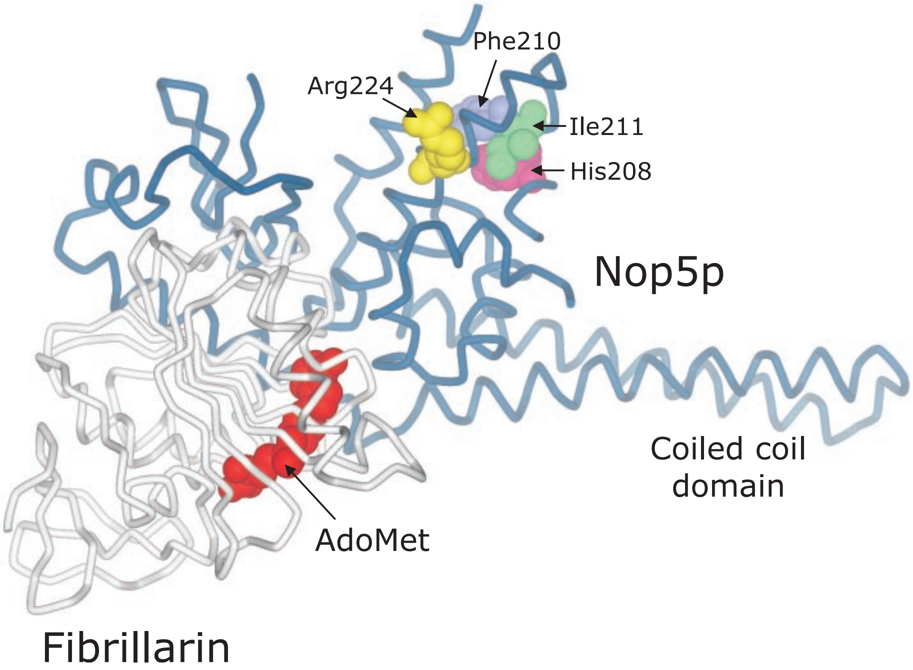


Figure 6. Structure of a Nop5p-fibrillarin complex from *A.fulgidus* as solved by Aittaleb *et al*. (47). In the protein backbone trace, Nop5p is depicted in blue and fibrillarin in white. The bound cofactor *S*-adenosyl-L-methionine (AdoMet) is shown as red–orange in a space-filling representation. The location of the coiled coil domain of Nop5p is indicated. The Nop5p amino acid residues that are discussed in the text and highlighted in Figure 5 are also shown in space-filling mode: Arg224 (yellow), His208 (pink), Phe210 (light blue) and Ile211 (green). The graphical representation of the structure was generated by DS ViewerPro 6.0 (Accelrys) using coordinates from PDB entry 1NT2.

E.gracilis PEP database, two distinct EST sequences are present: one encoding all but the extreme C-terminal portion of an apparent Nop58p homolog (Figure 5) and a second predicted to encode the N-terminal segment of the Nop56p homolog (region not shown in the alignment). At present, there are only a few ESTs corresponding to each homolog, which accounts for the fact that as yet we do not have EST clusters encoding the complete open reading frame (ORF) for either of these moderately large proteins. Importantly, two ORFs in the genomes of *G.lamblia* and *T.vaginalis* are predicted to encode both a Nop56p and a Nop58p homolog. It is apparent that the amino acid sequences of the two paralogs in these organisms are somewhat more homogenous (i.e. at residues 361, 363 and 364) relative to the situation in most other eukaryotes.

Currently, no structural information is available to pinpoint the residues in either Nop56p or Nop58p that directly interact with the box C/D RNA. In higher eukaryotes, it has been shown that the binding of the core proteins to a box C/D RNA occurs asymmetrically (48,49). Cross-linking experiments revealed contacts between Nop56p and the box C′/D′ motif, and between Nop58p and the C/D motif (49); however, the factors that contribute to the discrimination of these proteins for their respective binding partners are not known. As mentioned above, the most highly conserved region of these proteins is shown in Figure 5, suggesting that this segment represents a functional core within these paralogous proteins. We also note that the archaeal homolog, Nop5p (Nop56p/58p), which appears to be distributed symmetrically in the archaeal box C/D RNP, does not show strict evolutionary conservation of the same amino acid residues at positions 361 and 363 (Figure 5).

The structure of an archaeal Nop5p-fibrillarin complex from *Archaeoglobus fulgidus* has been solved by X-ray crystallography (47) and is depicted in Figure 6. The amino acid residues in the Nop5p component of the complex corresponding to *S.cerevisiae* Nop56p positions 361, 363 and 364 (His208, Phe210 and Ile211, respectively) are in close proximity to *A.fulgidus* Arg224, a residue whose substitution to alanine (R224A) has a detrimental effect on the RNA-binding ability of the *A.fulgidus* Nop5p–fibrillarin complex (47). Assuming that the eukaryotic Nop56p and Nop58p proteins adopt a similar fold to that of the archaeal Nop5p homolog, the residues corresponding to *S.cerevisiae* Nop56p residues 361, 363 and 364 may play a structural role that influences the RNA-binding properties of the eukaryotic proteins. It is possible that these residues contribute to the differential distribution of Nop56p and Nop58p within the eukaryotic box C/D RNP complex.

In conclusion, the gene duplication event that likely generated the Nop56p and Nop58p paralogs must also have occurred very early in eukaryotic evolution because organisms considered to be deep-branching have both paralogs. These combined results further suggest that in many protist box C/D snoRNP complexes, there may be asymmetric assembly of proteins on the C/D and C′/D′ motifs of the box C/D RNA, as has been observed in higher eukaryotes. This feature distinguishes eukaryotic from archaeal box C/D RNPs.

ACKNOWLEDGEMENTS

The authors thank Murray N. Schnare for critically reading the manuscript. The authors also thank Bernard A. Brown II (Wake Forest University) for his comments on the manuscript and for assisting with the preparation of Figure 6. Preliminary sequence data for *T.vaginalis*, *T.thermophila* and *Trypanosoma cruzi* were obtained from The Institute for Genomic Research website at http://www.tigr.org. Sequencing of the *T.vaginalis* and *T.cruzi* genomes was funded by the National Institute of Allergy and Infectious Diseases (NIAID). We thank Marlena Dlutek for performing DNA sequencing at Dalhousie University. Kiyoshi Kita generously permitted the use of his sequencing facility at the University of Tokyo. Nippon GeneTech (Japan) kindly provided CUGA sequencing kits. This work was supported by funding to M.W.G. from the Canadian Institutes of Health Research (operating grant MOP-11212) and Genome Canada through Genome Atlantic and the Atlantic Innovation Fund (Protist EST Program). M.W.G. is pleased to acknowledge salary support from the Canada Research Chairs Program and the Canadian Institute for Advanced Research. Funding to pay the Open Access publication charges for this article was provided by the Canadian Institutes of Health Research.

Conflict of interest statement. None declared.

REFERENCES

1. Tessier,L.-H., Keller,M., Chan,R.L., Fournier,R., Weil,J.-H. and Imbault,P. (1991) Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by *trans*-splicing in *Euglena*. *EMBO J*., **10**, 2621–2625.
2. Liang,X.-H., Haritan,S., Uliel,S. and Michaeli,S. (2003) *Trans* and *cis* splicing in trypanosomatids: mechanism, factors, and regulation. *Eukaryot. Cell*, **2**, 830–840.
3. Agabian,N. (1990) *Trans* splicing of nuclear pre-mRNAs. *Cell*, **61**, 1157–1160.
4. Blumenthal,T. (1998) Gene clusters and polycistronic transcription in eukaryotes. *BioEssays*, **20**, 480–487.
5. Rajkovic,A., Davis,R.E., Simonsen,J.N. and Rottman,F.M. (1990) A spliced leader is present on a subset of mRNAs from the human parasite *Schistosoma mansoni*. *Proc. Natl Acad. Sci. USA*, **87**, 8879–8883.
6. Vandenberghe,A.E., Meedel,T.H. and Hastings,K.E.M. (2001) mRNA 5′-leader *trans*-splicing in the chordates. *Genes Dev*., **15**, 294–303.
7. Stover,N.A. and Steele,R.E. (2001) Trans-spliced leader addition to mRNAs in a cnidarian. *Proc. Natl Acad. Sci. USA*, **98**, 5693–5698.
8. Mair,G., Shi,H., Li,H., Djikeng,A., Aviles,H.O., Bishop,J.R., Falcone,F.H., Gavrilescu,C., Montgomery,J.L., Santori,M.I. *et al*. (2000) A new twist in trypanosome RNA metabolism: *cis*-splicing of pre-mRNA. *RNA*, **6**, 163–169.
9. Nowitzki,U., Gelius-Dietrich,G., Schwieger,M., Henze,K. and Martin,W. (2004) Chloroplast phosphoglycerate kinase from *Euglena gracilis*: endosymbiotic gene replacement going against the tide. *Eur. J. Biochem*., **271**, 4123–4131.
10. Tessier,L.H., Paulus,F., Keller,M., Vial,C. and Imbault,P. (1995) Structure and expression of *Euglena gracilis* nuclear rbcS genes encoding the small subunits of the ribulose 1,5-bisphosphate carboxylase/oxygenase: a novel splicing process for unusual intervening sequences? *J. Mol. Biol*., **245**, 22–33.
11. Muchhal,U.S. and Schwartzbach,S.D. (1994) Characterization of the unique intron–exon junctions of *Euglena* gene(s) encoding the polyprotein precursor to the light-harvesting chlorophyll a/b binding protein photosystem II. *Nucleic Acids Res*., **22**, 5737–5744.
12. Breckenridge,D.G., Watanabe,Y.-i., Greenwood,S.J., Gray,M.W. and Schnare,M.N. (1999) U1 small nuclear RNA and spliceosomal introns in *Euglena gracilis*. *Proc. Natl Acad. Sci. USA*, **96**, 852–856.
13. Canaday,J., Tessier,L.H., Imbault,P. and Paulus,F. (2001) Analysis of *Euglena gracilis alpha*-, *beta*- and *gamma*-tubulin genes: introns and pre-mRNA maturation. *Mol. Genet. Genomics*, **265**, 153–160.
14. Newman,D.R., Kuhn,J.F., Shanab,G.M. and Maxwell,E.S. (2000) Box C/D snoRNA-associated proteins: two pairs of evolutionarily ancient proteins and possible links to replication and transcription. *RNA*, **6**, 861–879.
15. Lafontaine,D.L. and Tollervey,D. (1999) Nop58p is a common component of the box C+D snoRNPs that is required for snoRNA stability. *RNA*, **5**, 455–467.
16. Gautier,T., Berges,T., Tollervey,D. and Hurt,E. (1997) Nucleolar KKE/D repeat proteins Nop56p and Nop58p interact with Nop1p and are required for ribosome biogenesis. *Mol. Cell. Biol*., **12**, 7088–7098.
17. Watkins,N.J., Segault,V., Charpentier,B., Nottrott,S., Fabrizio,P., Bachi,A., Wilm,M., Rosbash,M., Branlant,C. and Lührmann,R. (2000) A common core RNP structure shared between the small nucleolar box C/D RNPs and the spliceosomal U4 snRNP. *Cell*, **103**, 457–466.
18. Kuhn,J.F., Tran,E.J. and Maxwell,E.S. (2002) Archaeal ribosomal protein L7 is a functional homolog of the eukaryotic 15.5kD/Snu13p snoRNP core protein. *Nucleic Acids Res*, **30**, 931–941.
19. Omer,A.D., Lowe,T.M., Russell,A.G., Ebhardt,H., Eddy,S.R. and Dennis,P.P. (2000) Homologs of small nucleolar RNAs in Archaea. *Science*, **288**, 517–522.
20. Rozhdestvensky,T.S., Tang,T.H., Tchirkova,I.V., Brosius,J., Bachellerie,J.-P. and Hüttenhofer,A. (2003) Binding of L7Ae protein to the K-turn of archaeal snoRNAs: a shared RNA binding motif for C/D and H/ACA box snoRNAs in Archaea. *Nucleic Acids Res*., **31**, 869–877.
21. Zago,M.A., Dennis,P.P. and Omer,A.D. (2005) The expanding world of small RNAs in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Mol. Microbiol*., **55**, 1812–1828.
22. Mager,W.H., Planta,R.J., Ballesta,J.-P.G., Lee,J.C., Mizuta,K., Suzuki,K., Warner,J.R. and Woolford,J. (1997) A new nomenclature for the cytoplasmic ribosomal proteins of *Saccharomyces cerevisiae*. *Nucleic Acids Res*., **25**, 4872–4875.
23. Tran,E., Brown,J. and Maxwell,E.S. (2004) Evolutionary origins of the RNA-guided nucleotide-modification complexes: from the primitive translation apparatus? *Trends Biochem. Sci*., **29**, 343–350.
24. Russell,A.G., Schnare,M.N. and Gray,M.W. (2004) Pseudouridine-guide RNAs and other Cbf5p-associated RNAs in *Euglena gracilis*. *RNA*, **10**, 1034–1046.
25. Watanabe,Y.-i. and Gray,M.W. (2000) Evolutionary appearance of genes encoding proteins associated with box H/ACA snoRNAs: Cbf5p in *Euglena gracilis*, an early diverging eukaryote, and candidate Gar1p and Nop10p homologs in archaebacteria. *Nucleic Acids Res*., **28**, 2342–2352.
26. Maruyama,K. and Sugano,S. (1994) Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene*, **138**, 171–174.
27. Siebert,P.D., Chenchik,A., Kellogg,D.E., Lukyanov,K.A. and Lukyanov,S.A. (1995) An improved PCR method for walking in uncloned genomic DNA. *Nucleic Acids Res*., **23**, 1087–1088.
28. Fitzgerald,M.C., Skowron,P., Van Etten,J.L., Smith,L.M. and Mead,D.A. (1992) Rapid shotgun cloning utilizing the two base recognition endonuclease *Cvi*JI. *Nucleic Acids Res*., **20**, 3753–3762.
29. Sasaki,N., Izawa,M., Watahiki,M., Ozawa,K., Tanaka,T., Yoneda,Y., Matsuura,S., Carninci,P., Muramatsu,M., Okazaki,Y. *et al*. (1998) Transcriptional sequencing: A method for DNA sequencing using RNA polymerase. *Proc. Natl Acad. Sci. USA*, **95**, 3455–3460.
30. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res*., **31**, 3497–3500.
31. David,E., McNeil,J.B., Basile,V. and Pearlman,R.E. (1997) An unusual fibrillarin gene and protein: structure and functional implications. *Mol. Biol. Cell*, **8**, 1051–1061.
32. Lischwe,M.A., Ochs,R.L., Reddy,R., Cook,R.G., Yeoman,L.C., Tan,E.M., Reichlin,M. and Busch,H. (1985) Purification and partial characterization of a nucleolar scleroderma antigen (Mr = 34,000; pI, 8.5) rich in NG,NG-dimethylarginine. *J. Biol. Chem*., **260**, 14304–14310.
33. Xu,C., Henry,P.A., Setya,A. and Henry,M.F. (2003) *In vivo* analysis of nucleolar proteins modified by the yeast arginine methyltransferase Hmt1/Rmt1p. *RNA*, **6**, 746–759.
34. Barneche,F., Steinmetz,F. and Echeverria,M. (2000) Fibrillarin genes encode both a conserved nucleolar protein and a novel small nucleolar RNA involved in ribosomal RNA methylation in *Arabidopsis thaliana*. *J. Biol. Chem*., **275**, 27212–27220.
35. Sidrauski,C. and Walter,P. (1997) The transmembrane kinase Ire1p is a site-specific endonuclease that initiates mRNA splicing in the unfolded protein response. *Cell*, **90**, 1031–1039.
36. Sidrauski,C., Cox,J.S. and Walter,P. (1996) tRNA ligase is required for regulated mRNA splicing in the unfolded protein response. *Cell*, **87**, 405–413.

37. Watanabe,Y.-i., Yokobori,S.-I., Inaba,T., Yamagishi,A., Oshima,T., Kawarabayasi,Y., Kikuchi,H. and Kita,K. (2002) Introns in protein-coding genes in Archaea. *FEBS Lett.*, **510**, 27–30.
38. Henze,K., Badr,A., Wettern,M., Cerff,R. and Martin,W. (1995) A nuclear gene of eubacterial origin in *Euglena gracilis* reflects cryptic endosymbioses during protist evolution. *Proc. Natl Acad. Sci. USA*, **92**, 9122–9126.
39. Archibald,J.M. and Keeling,P.J. (2002) Recycled plastids: a 'green movement' in eukaryote evolution. *Trends Genet.*, **18**, 577–584.
40. Watkins,N.J., Dickmanns,A. and Lührmann,R. (2002) Conserved stem II of the box C/D motif is essential for nucleolar localization and is required, along with the 15.5K protein, for the hierarchical assembly of the box C/D snoRNP. *Mol. Cell. Biol*, **22**, 8342–8352.
41. Omer,A.D., Ziesche,S., Ebhardt,H. and Dennis,P.P. (2002) *In vitro* reconstitution and activity of a C/D box methylation guide ribonucleoprotein complex. *Proc. Natl Acad. Sci. USA*, **99**, 5289–5294.
42. Dobbyn,H.C. and O'Keefe,R.T. (2004) Analysis of Snu13p mutations reveals differential interactions with the U4 snRNA and U3 snoRNA. *RNA*, **10**, 308–320.
43. Vidovic,I., Nottrott,S., Hartmuth,K., Lührmann,R. and Ficner,R. (2000) Crystal structure of the spliceosomal 15.5 kD protein bound to a U4 snRNA fragment. *Mol. Cell*, **6**, 1331–1342.
44. Nottrott,S., Hartmuth,K., Fabrizio,P., Urlaub,H., Vidovic,I., Ficner,R. and Lührmann,R. (1999) Functional interaction of a novel 15.5kD [U4/U6.U5] tri-snRNP protein with the 5′ stem-loop of U4 snRNA. *EMBO J.*, **18**, 6119–6133.
45. Hamma,T. and Ferré-D'Amaré,A.R. (2004) Structure of protein L7Ae bound to a K-turn derived from an archaeal box H/ACA sRNA at 1.8Å resolution. *Structure (Camb.)*, **12**, 893–903.
46. Russo,G., Cuccurese,M., Monti,G., Russo,A., Amoresano,A., Pucci,P. and Pietropaolo,C. (2005) Ribosomal protein L7a binds RNA through two distinct RNA-binding domains. *Biochem. J.*, **385**, 289–299.
47. Aittaleb,M., Rashid,R., Chen,Q., Palmer,J.R., Daniels,C.J. and Li,H. (2003) Structure and function of archaeal box C/D sRNP core proteins. *Nature Struct. Biol.*, **10**, 256–263.
48. Szewczak,L.B.W., DeGregorio,S.J., Strobel,S.A. and Steitz,J.A. (2002) Exclusive interaction of the 15.5 kD protein with the terminal box C/D motif of a methylation guide snoRNP. *Chem. Biol.*, **9**, 1095–1107.
49. Cahill,N.M., Friend,K., Speckmann,W., Li,Z.-H., Terns,R.M., Terns,M.P. and Steitz,J.A. (2002) Site-specific cross-linking analyses reveal an asymmetric protein distribution for a box C/D snoRNP. *EMBO J.*, **21**, 3816–3828.