Small nucleolar RNAs and RNA-guided post-transcriptional modification

Lauren Lui and Todd Lowe1

Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, U.S.A.

Abstract

snoRNAs (small nucleolar RNAs) constitute one of the largest and best-studied classes of non-coding RNAs that confer enzymatic specificity. With associated proteins, these snoRNAs form ribonucleoprotein complexes that can direct 2′-O-methylation or pseudouridylation of target non-coding RNAs. Aided by computational methods and high-throughput sequencing, new studies have expanded the diversity of known snoRNA functions. Complexes incorporating snoRNAs have dynamic specificity, and include diverse roles in RNA silencing, telomerase maintenance and regulation of alternative splicing. Evidence that dysregulation of snoRNAs can cause human disease, including cancer, indicates that the full scope of snoRNA roles remains an unfinished story. The diversity in structure, genomic origin and function between snoRNAs found in different complexes and among different phyla illustrates the surprising plasticity of snoRNAs in evolution. The ability of snoRNAs to direct highly specific interactions with other RNAs is a consistent thread in their newly discovered functions. Because they are ubiquitous throughout Eukarya and Archaea, it is likely they were a feature of the last common ancestor of these two domains, placing their origin over two billion years ago. In the present chapter, we focus on recent advances in our understanding of these ancient, but functionally dynamic RNA-processing machines.

Keywords:
2′-O-methylation, post-transcriptional modification, pseudouridylation, ribosome biogenesis, small nucleolar RNA.

1To whom correspondence should be addressed (email lowe@soe.ucsc.edu).
Introduction

Although snoRNAs (small nucleolar RNAs) are still primarily characterized by their ability to direct post-transcriptional modification of RNAs, recent research has revealed diverse functions, expression strategies and genomic organization. Excellent reviews exist on snoRNAs that detail their best-understood molecular roles [1–8]. The present chapter seeks to highlight new discoveries and the functional elasticity of this class of ncRNA (non-coding RNA). snoRNAs have had a long history of revealing unexpected new functions – scientists did not characterize their defining role in post-transcriptional modification until three decades after their discovery. First observed in the nucleolus of mammalian cells in the late 1960s and early 1970s ([9,10], reviewed in [11]), snoRNAs were not yet clearly divided into the two major families we understand today. Research in the late 1980s and early 1990s confirmed earlier predicted roles for these RNAs in directing pre-ribosomal RNA cleavage ([12–14], reviewed in [1,8]). Later in the mid-1990s, scientists used conserved sequence features to divide snoRNAs into the families that are currently accepted today [15,16]. In 1996, two separate studies demonstrated that ‘C/D box’ RNAs can direct 2′-O-methylation (Figures 1A and 1B) of rRNA (ribosomal RNA) [17,18]. A year later, two groups also demonstrated that ‘H/ACA box’ RNAs direct pseudouridylation (Figures 1C and 1D) of rRNA [19,20]. Follow-up studies showed that the vast majority of C/D box RNAs in *Saccharomyces cerevisiae* are indeed required for almost all 2′-O-methyl modifications in yeast rRNA [21].

In 2000, biochemical and computational experiments surprisingly confirmed the existence of C/D box sRNAs (sno-like RNAs) in Archaea, the ‘third domain’ of life [22,23]. Later experiments confirmed *in vitro* methylation activity of target sites on rRNA and tRNA [24,25]. Although archaeal organisms are prokaryotic, and therefore lack a nucleus and nucleolus, the presence of snoRNA machinery contributes to the body of evidence indicating that archaea are more closely related to eukaryotes than bacteria. Bacteria also have pseudouridylated and 2′-O-methylated RNAs, but do not have snoRNAs. Instead, they use site- or region-specific protein enzymes to modify ribonucleotides [26]. Archaea and eukaryotes also have similar DNA replication, transcription and translation systems compared with those in bacteria. The existence of C/D box and H/ACA box RNAs, along with their associated orthologous proteins in archaea and eukaryotes, suggests that this system of RNA modification is an ancient one, estimated at over two billion years old [27–29]. New evidence indicates snoRNAs have evolved additional, somewhat unexpected, cellular roles since their origin.

Studies in the past decade linking snoRNA expression to cancer highlight how incomplete our understanding remains. In 2002, Chang et al. [30] provided evidence that an H/ACA box RNA is significantly down-regulated in human brain cancer, the first report of a relationship between snoRNA dysregulation and cancer. Since then, dysregulation of snoRNAs has been linked to lymphoma, breast, prostate, lung, and head and neck cancers (reviewed in [31,32]). Some snoRNAs have tumour suppressor roles, whereas the up-regulation of others may contribute to tumorigenesis. Mutations of snoRNA-associated proteins or snoRNAs that modify rRNA can also cause ribosome dysfunction, leading to a variety of cancers and other diseases [33,34]. Studies of snoRNAs and snoRNA host genes linked to cancer suggest that snoRNA host genes, previously thought to have no protein-coding value, may indeed have roles in cell development and homeostasis (reviewed in [32]). Because some snoRNAs have differential expression patterns in specific cancer types, the levels of particular snoRNAs in
the blood could be used as a non-invasive method to detect and type specific cancers [31]. Cancer research is expanding our current understanding of how snoRNAs and snoRNA host genes affect cell state, further underlining new roles to be studied.

Although the focus of the present chapter is to highlight new functions of snoRNAs, the study of snoRNAs has been driven by the importance of 2′-O-methylation and pseudouridylation in non-coding RNAs. 2′-O-Methylation and pseudouridylation are the most commonly found RNA modifications and play important roles in the stability and folding of RNA [26,35]. Pseudouridines are so numerous they have been called the ‘fifth nucleoside’ [36]. Over 100
sites in human rRNA are targeted for methylation, and nearly that many are pseudouridylated [26,37]. Most snoRNAs target rRNAs, but in eukaryotes these RNAs also target snRNAs (small nuclear RNAs) involved with the spliceosome [38] and tRNAs in archaea [39,40]. Modifications by snoRNAs often cluster around functionally important regions in rRNA, snRNAs and tRNAs, suggesting that modification plays a role in the function of these RNAs (reviewed in [2,41]). RNA modification by snoRNAs represents a fine-tuning mechanism for a molecule that is initially built with only four different units.

To fully understand the emerging functional roles of snoRNAs, potential synthetic applications, and the promise of accelerated detection of new snoRNA genes by computational methods, it is helpful to review their biogenesis and defining sequence features. We focus mainly on the largest subset of C/D box RNAs and H/ACA box RNAs which direct the 2′-O-methylation and pseudouridylation of target RNAs respectively [17–20].

**Composition and structure of snoRNPs**

As with many RNA families, both primary sequence and structural features play important roles in snoRNA function. snoRNAs serve as scaffolds in the formation of the snoRNP (small nucleolar ribonucleoprotein) complex, as well as acting as base-pairing guides to target specific ribonucleotides for modification. As computational biologists, we closely observe the natural variation in conserved sequence and secondary structural features, as this enables the development of sensitive computational models to detect this class of RNAs and their targets.

**C/D box RNPs**

C/D box RNAs are ~50–300 nt in length and contain characteristic fairly well-conserved C (RUGAGA) and D (CUGA) box elements, and similar but less conserved C′ and D′ box elements (Figure 1B, see [4]). Archaeal C/D box sRNAs are shorter than their eukaryotic cousins (~50–70 nt), but also contain C, D, C′ and D′ boxes with remarkably similar sequence motifs as their eukaryotic counterparts. The C and D boxes, located respectively at the 5′ and 3′ termini of the RNA, form the ‘kink-turn’, a stem-bulge-stem RNA structure that is critical to the formation of the snoRNP [42]. L7Ae (archaea) or the 15.5 kDa protein (eukaryotes) binds to the kink-turn, allowing the rest of the proteins associated with the C/D box RNP to bind. In archaeal species, the internal C′ and D′ boxes also form a ‘K-loop’, a version of the kink-turn with one stem replaced by a loop, and which is also bound by L7Ae [42]. The conserved C and D box features and formation of the kink-turn are hallmarks of this class of snoRNA.

Typically, eukaryotic C/D box RNAs contain a 10–21 nt guide sequence complementary to the target RNA to be modified by the methyltransferase fibrillarin. This guide sequence is typically shorter in archaeal C/D box sRNAs, generally ranging from 8 to 12 nt in length. Budding yeast snoRNAs tend to be longer than those in most other studied species, with additional sequence in the region between the C′ and D′ boxes. Recently, a study of yeast C/D box RNAs found that many of these extra ‘middle region’ sequences have additional complementarities with their target RNAs. These additional sites of interaction flank the target methylation site, and can stimulate methylation efficiency up to 5-fold [43].

An evolutionarily well-conserved set of core proteins associate with C/D box RNAs to form a unique RNA-directed molecular machine with the striking ability to add 2′-O-methyl groups to a theoretically unbounded number of different, but highly specific, positions across

© The Authors Journal compilation © 2013 Biochemical Society
any number of RNA targets (reviewed in [44]). The range of specificity of this enzymatic complex is only limited by the number of different snoRNAs encoded within the genome. In eukarya, the 15.5 kDa snoRNP protein binds to the kink-turn formed by the C and D boxes. Next, Nop56, Nop58 and fibrillarin bind to complete the RNP. In archaea, the snoRNP is simpler, comprised of the C/D box RNA bound to L7Ae (a homologue with the 15.5 kDa protein), Nop 56/58 (related to both eukaryotic Nop56 and Nop58), and fibrillarin. Although these core proteins are clearly homologous between Archaea and Eukarya, the assembly of the RNP differs between the two domains. Recent research on the assembly of C/D box RNPs is discussed in further detail below.

**H/ACA box RNPs**

RNAs that belong to the H/ACA box RNA family share hallmark secondary structure and primary sequence elements that are quite unique from C/D box snoRNAs. Typical H/ACA box RNAs are ~60–150 nt in length (longer in yeast) and are named for a hinge sequence with a consensus of ANANNA and an ACA element at the 3′-end (Figure 1D) [15,19]. The terminal ACA motif is something of a misnomer in some species, as it can commonly vary, leading it to also be referred to as the ‘ANA’ terminal motif. For example, in trypanosomes, the ACA can be AGA [45], and in some unusual archaeal thermophiles, the ACA can be omitted altogether [46]. Structurally, H/ACA box RNAs are often comprised of two hairpins separated by the hinge element, but the number of hairpins can vary. Eukaryotic H/ACA box RNAs have two hairpins, except in trypanosomes where they can be singlets [47]. In yeast, both hairpin structures are needed for modification even if the H/ACA box RNA only has one pseudouridylation target [48]. Single hairpins have also been found in the protist *Euglena gracilis* [49]. H/ACA box RNAs with one, two or three hairpins have been found in archaea [50,51].

Each hairpin is comprised of a lower stem, a central bulge dubbed the ‘pseudouridylation pocket’, an upper stem and the apical loop. The pseudouridylation pocket is analogous to the guide sequence of C/D box RNAs, but is discontinuous with the ‘left guide’ and ‘right guide’ regions that are interrupted by the upper stem and apical loop. Within the context of its tertiary structure, the left and right guides are in close proximity, yet the separation in the linear sequence of the two complementary regions makes computational detection of H/ACA box RNA genes and their targets much more difficult. When the left and right guide sequences within the pseudouridylation pocket base pair with the target RNA, two nucleotides of the target RNA remain unpaired at the base of the upper stem. The 5′-most nucleotide of this pair is converted from uridine into pseudouridine. Often in archaea, the upper stem forms a kink-turn that is recognized by L7Ae. In eukaryotes this kink-turn is not known to form, but NHP2, the eukaryotic homologue of L7Ae, is structurally similar to L7Ae and may interact with the upper stem of H/ACA box RNAs [52]. Despite the conservation of secondary structure across species, some H/ACA box RNAs in the archaeal genus *Pyrobaculum* are missing the lower stem [46], suggesting that minimal forms of H/ACA box RNAs could exist undetected in other species as well.

The four core proteins of H/ACA box RNPs are Cbf5 (also known as dyskerin/DKC1 in mammals or NAP57), Nop10, Gar1 and L7Ae (Figure 2). The key protein in the complex, Cbf5, is a pseudouridine synthase which catalyzes the isomerization of the target uridine to pseudouridine. It contains a catalytic domain and a PUA (pseudouridine and archeosine transglycosylase) domain. The PUA domain has a dual role as an RNA-binding motif and a localization signal to the nucleus (reviewed [36]). Mutations of this domain are linked to dyskeratosis congenita,
a rare inherited bone marrow disease [4]. The ACA box and lower stem of H/ACA box RNAs are bound by the PUA domain, which positions the guide regions by the catalytic domain. Studies of Cbf5 alone and complexed in the H/ACA box RNP indicate that it is structurally similar to TruB, the bacterial pseudouridine synthase which does not harness an RNA cofactor for specificity [6]. In some species, Cbf5 isomerizes uridines of tRNAs without the presence of other H/ACA box RNA proteins [53,54], which also also supports its common ancestry with TruB.

**scaRNPs (small Cajal body-specific RNAs)**
scaRNAs are a related hybrid type of snoRNA that associate with C/D box and H/ACA box RNA proteins, and direct both 2′-O-methylation and pseudouridylation of snRNAs [55]. These RNAs contain CAB boxes (Cajal body box, consensus UGAG) in the apical loop of the hairpin. CAB boxes localize the scaRNA to Cajal bodies, nuclear organelles that function in snRNP maturation [3].

**Functions of snoRNAs beyond ribosomal processing and modification**

Although the majority of C/D box and H/ACA box RNAs known today are annotated for their roles in RNA modification, early studies recognized and detailed their essential role in rRNA processing during ribosome biogenesis (reviewed in [1]). Among eukaryotes, uridine-rich
snoRNAs such as U3, U14, U22, snR10 and snR30 act as integral components of the processing machinery by guiding endonucleolytic cleavage of pre-rRNA and folding of the mature rRNA (reviewed in [8,56]). In both eukaryotes and archaea, the proximal base-pairing sites between 2′-O-methyl guide RNAs and rRNA further support their general chaperone role in rRNA folding [8,40]. Recent studies also indicate that RNA helicases regulate the association of snoRNPs with rRNA [8].

There is now ample evidence that other classes of RNAs interact directly with snoRNAs. For years, snoRNA studies have noted C/D box or H/ACA box snoRNAs that have potential guide sequences with no clear modification targets among rRNAs or other known modification targets (spliceosomal RNAs in eukaryotes [38]; tRNAs in archaea [39,40]). The presence of these ‘orphan’ guide snoRNA genes in most species suggests there are additional target RNAs to be identified [57,58]. Computational search algorithms, expression screens and high-throughput sequencing have greatly aided traditional RNA biochemical approaches in identifying these new interactions.

**snoRNAs and RNA silencing**

Studies indicate that some snoRNAs are processed into miRNAs (microRNAs), another important and incompletely understood class of small ncRNAs. First discovered in 1993 [59], miRNAs are ~18–24 nt long RNAs that associate with RNA silencing machinery to regulate the stability and translation of messenger RNAs [60]. The development of high-throughput sequencing methods for small RNAs has accelerated our appreciation of the complexity and dynamics of small RNA populations, including snoRNAs and miRNAs.

In many cases, miRNAs derived from snoRNAs (termed sno-miRNAs or sno-derived RNAs) were discovered by a combination of bioinformatics and analysis of deep sequencing of RNAs physically associated with core proteins of the RNA silencing pathway, such as Argonaute [61,62]. sno-miRNAs derived from both C/D and H/ACA box snoRNAs have been found in a variety of eukaryotes, including human, mouse, fruit fly, plants and fission yeast [61–65]. C/D box snoRNAs that act as miRNA precursors have also been found in the parasitic protozoan *Giardia lamblia* [66] and the Epstein–Barr virus [67]. Sequencing libraries that have been enriched for small RNAs by size fractionation also aid the search for snoRNAs and sno-miRNAs [65,68]. Determining snoRNA database entries that overlap with miRNAs has also yielded new discoveries [65].

How snoRNAs are processed into sno-miRNAs is still unclear, but the mechanism may include traditional snoRNA processing machinery [7] and RNAi (RNA interference) proteins [61]. Some sno-miRNAs retain the traditional functionality of snoRNAs and are localized in the nucleolus [64,69]. H/ACA box snoRNAs that act as miRNA precursors have been shown to bind to snoRNA-associated proteins, including dyskerin [64] and GAR-1 [62]. Populations of sno-derived small RNAs are also affected by the loss of components of the RNAi pathway, such as Dicer [62].

This new role of snoRNAs in gene silencing may partially explain the presence of ‘orphan guides’ in some cases. The targets of these conserved guides may be mRNAs to be silenced, instead of rRNAs to be modified. Most of the RNAs derived from snoRNAs that have been studied to date appear to have miRNA-like functions, but derived RNAs longer than 22 nt may have other functions [7]. The emerging relationship between snoRNAs and RNA silencing may reveal more about the evolution of RNA in gene regulation.
Telomerase: a specialized H/ACA box RNA in vertebrates

Telomeres, the ends of eukaryotic chromosomes, are maintained by the RNP complex telomerase which consists of a protein reverse transcriptase [TERT (telomerase reverse transcriptase)] and TR (telomerase RNA; also abbreviated to TER or TERC). Without telomerase, telomeres would shorten and eventually be lost due to incomplete replication of telomere ends by conventional DNA polymerases [70]. In vertebrates, the TR is approximately 450 nt long and has H/ACA box RNA features at its 3′-end which are required for proper localization and accumulation. One of these features, the CAB box, directs telomerase to Cajal bodies [3]. Vertebrate TR interacts with all of the core proteins of H/ACA box RNAs, but interestingly, there is no evidence of a pseudouridylation target [3]. Mutations in TR or its associated proteins can lead to the syndrome dyskeratosis congenita [4].

Spliced leader mRNAs in trypanosomes

Maturation of all mRNAs in trypanosomes requires the incorporation of a 39 nt spliced leader exon that is added in trans [47,71]. A trypanosome H/ACA box RNA has been shown to guide pseudouridylation of the spliced leader RNA [47], providing an important example of mRNA modification by snoRNAs, rather than the typical rRNA, tRNA and snRNA targets.

Artificial uses of snoRNAs

Scientists have utilized the main features of snoRNAs, such as the ability to target specific RNA sequences and localization to the nucleolus, for synthetic biology purposes. C/D box RNAs have been used to regulate gene expression [72,73], regulate alternative splicing [74,75], localize ribozymes to the nucleolus in yeast [76] and guide site-specific modification [77]. Modified H/ACA box RNAs have also been used to convert translation termination codons into sense codons to suppress translation termination. Not only does this method suggest that mRNAs may be naturally modified, it indicates a potential way to treat premature termination codon diseases, such as cystic fibrosis and Duchenne muscular dystrophy [78,79]. The ability to synthesize C/D box RNAs that affect alternative splicing and gene expression by only changing guide sequences [72,75], together with current stories of non-canonical snoRNA function [7,80] suggest that natural snoRNAs could have already evolved these functions. High-throughput sequencing technologies to detect post-transcriptional modifications in mRNAs is rapidly improving, which should aid in searches for novel mRNA–snoRNA interactions.

An unexpected pathological role of snoRNAs in human disease

In the late 1990s and early in the new millennium, it was found that dysregulation of snoRNA expression can cause human disease [81,82]. The best understood example is the genetic disorder PWS (Prader–Willi syndrome), a congenital disease characterized by mental retardation, hyperphagia leading to obesity and short stature. It affects approximately 1 in 8000–20000 people and is the most common genetic cause for Type II diabetes [80,83].

Investigation of the genetic defects associated with PWS led to the discovery of tissue-specific snoRNAs in the brain, and implicates processed snoRNAs in the regulation of
alternative splicing. Minimal deletions in the SNURF-SNRPN locus on paternal chromosome 5q11-q13 were genetically linked to PWS [84]. The SNURF-SNRPN pre-mRNA contains a heterogeneous cluster of intronic orphan snoRNAs, and is maternally imprinted (only expressed from the paternal allele). This atypical snoRNA gene cluster includes multiple copies of SNORD115 (also known as HBII-52), which only occurs in the brain [85]. SNORD115 is processed into shorter forms that can bind to the alternative exon 5b of the serotonin-2C receptor mRNA [80]. The hybridization of the snoRNA with the pre-mRNA masks a splicing silencer, allowing for the incorporation of the alternative exon in the mature mRNA [80,85]. Bioinformatics analysis also indicates that the mouse form of the SNORD115 (MBII-52) may regulate alternative splicing of mRNAs other than that of the serotonin-2C receptor [80]. Yet another snoRNA in the SNURF-SNRPN locus, SNORD116/ HBII-85, may regulate alternative splicing of other mRNA targets [86]. The continuing studies of PWS suggest that there are undiscovered functional roles of snoRNAs outside of RNA modification.

### Diversity in expression and organization among species

The varied genomic organization of C/D and H/ACA box RNAs within individual genomes and across different phyla illustrates the dynamic transcriptional units and processing pathways associated with this RNA family. snoRNA genes may occur as independently transcribed genes, within introns of protein-coding genes or IncRNAs (long ncRNAs), or in clusters that are transcribed as a polycistronic transcript (reviewed in [4,5,87]).

All species studied have at least some fraction of their snoRNAs driven by independent promoters, either as monocistrons or polycistronic transcripts (Figure 3). Most eukaryotic snoRNAs that fall into this class are transcribed by RNA Pol II (RNA polymerase II). The few that are transcribed by RNA Pol III have additional transcription elements such as internal box A and box B motifs, or the upstream sequence element in plants. Also in plants are discistronic tRNA-snoRNA genes that are transcribed by RNA Pol III from the tRNA promoter [5]. C/D box RNAs processed from polycistrons are trimmed by exonucleases (reviewed in [4,7]).

With the exception of protozoans, most sequenced eukaryotic genomes contain intron-encoded snoRNAs. In a few notable instances, snoRNAs are found in the introns of IncRNAs in mammals [88], and in select archaea there is an instance of a C/D box sRNA encoded within the intron of a tRNA gene [39]. The snoRNAs located in introns can occur singly or in clusters. These clusters can be homogenous (copies of same snoRNA gene) or heterogenous; in some cases, introns may even be composed of both C/D and H/ACA box RNAs. Some of the most highly conserved snoRNAs across phyla are commonly found in core genes such as those encoding ribosomal proteins (e.g. RPL7A), but in other cases, host genes do not appear to have a significant protein product (e.g. GAS5) [4].

snoRNAs found within polycistrons or introns require processing by nucleases before they are functional (reviewed in [7]). In yeast, polycistronic snoRNAs are cleaved preferentially in hairpin structures containing AGNN tetraloops by Rnt1p, an RNase III homologue. Extra sequence is cleaved away by 5' → 3' and 3' → 5' exonucleases (Table 1). Intronic snoRNAs can be processed from introns released as lariats or, more rarely, excised by endonucleases. Intronic snoRNAs originating from lariat-structured introns are also trimmed by exonucleases. In
eukaryotes, snoRNAs mature in Cajal bodies, although the nucleolus is the final destination of most snoRNAs.

Evidence supports the propagation of snoRNAs by transposable elements and other types of duplications. In plants, the large presence of snoRNAs is attributed to polyploidy, large chromosomal rearrangements and tandem duplications (reviewed in [87,89]). Bioinformatics analyses indicate that mammalian snoRNAs result from retrotransposition events [90,91]. In the human genome, hundreds of snoRNAs and snoRNA-related molecules appear to be derived from transposons. These snoRNAs have features of retrotransposons, such as an A-rich tail and are flanked by a target site duplication sequence that corresponds to their insertion site. Diversity of genomic location provides a challenge to the computational prediction methods used to find and annotate snoRNAs, in addition to variation in sequence and structure.

### Computational search methods and databases

Computational methods can greatly accelerate the search for plausible snoRNA candidates relative to traditional biochemical isolation experiments, but a fair number of challenges remain.

---

**Figure 3. Types of genomic organization of snoRNA genes with distribution within species**

Animal, plant, and yeast numbers are based on data collected by [5], G. lamblia data from [117], T. brucei data from [118] and archaeal data from [22,23,40]. MRP RNA genes are not included in this table. The diversity of snoRNAs is illustrated by the variety of genomic locations in which they are found. snoRNA genes (coloured block arrows) can be independently transcribed with their own promoters or found in the introns of protein and (rarely) RNA genes. Clusters may be composed of both C/D and H/ACA box RNAs, such as in plants and trypanosomes [47]. Typically mammalian host genes are involved in ribosomal biogenesis or protein synthesis [98], but several do not have a significant protein product. snoRNAs independently transcribed or encoded within introns are transcribed by RNA Pol II unless otherwise indicated. aFor C. elegans, the number of snoRNA coding units is reported based on data from [5]. These units have elements indicating that they are potentially recognized by RNA Pol III and are already included in the other categories of this table based on their genomic location. bOnly the U3 snoRNA homologue is known to be independently transcribed by RNA Pol III in trypanosomes [119]. cSome archaean species that belong to the phylum Euryarchaeota, such as Pyrococcus furiosus, have a C/D box RNA encoded within tRNA-TRP [25].
Table 1. Processing, assembly and core proteins of snoRNPs

HUGO gene IDs for vertebrate proteins are included after the gene name for reference. Note that in archaea snoRNA processing proteins are not known. The Table is modified from [7], with additions from [8].

<table>
<thead>
<tr>
<th>Protein</th>
<th>Vertebrate</th>
<th>Yeast</th>
<th>Archaea</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core C/D box RNP proteins</td>
<td>15.5 kDa (NHP2L1)</td>
<td>Snu13p</td>
<td>L7Ae</td>
<td>Bind to kink-turn formed by C/D motifs (and C'/D' motif in archaea), nucleation of snoRNP assembly</td>
</tr>
<tr>
<td>Naf1 (NAF1)</td>
<td></td>
<td></td>
<td></td>
<td>Shuttle protein, accumulation of H/ACA box RNAs, placeholder for GAR1</td>
</tr>
<tr>
<td>Shq1 (SHQ1)</td>
<td></td>
<td></td>
<td></td>
<td>Shuttle protein, placeholder for H/ACA box RNA</td>
</tr>
<tr>
<td>Bcd1 (KLF6)</td>
<td></td>
<td></td>
<td></td>
<td>Accumulation of C/D box RNAs, coupling of transcription and processing of snoRNA</td>
</tr>
<tr>
<td>IBP160 (AQR)</td>
<td></td>
<td></td>
<td></td>
<td>Coupling between pre-mRNA splicing and snoRNA assembly</td>
</tr>
<tr>
<td>NOP17 (PIH1D1)</td>
<td></td>
<td></td>
<td></td>
<td>C/D box RNA assembly</td>
</tr>
<tr>
<td>Core H/ACA box RNP proteins</td>
<td>NHP2 (NHP2)</td>
<td>Nhp2p</td>
<td>L7Ae</td>
<td>Bind to kink-turn (archaea), nucleation of snoRNP assembly</td>
</tr>
<tr>
<td>NAP57/dyskerin (DKC1)</td>
<td>Cbf5p</td>
<td>Cbf5</td>
<td></td>
<td>Pseudouridine synthase</td>
</tr>
<tr>
<td>GAR1 (GAR1)</td>
<td>Gar1p</td>
<td>Gar1</td>
<td></td>
<td>Target RNA turnover</td>
</tr>
<tr>
<td>NOP10 (NOP10)</td>
<td>Nop10p</td>
<td>Nop10</td>
<td></td>
<td>Interaction with L7Ae and core RNA</td>
</tr>
<tr>
<td>Assembly proteins</td>
<td>TIP48 (RUVBL2); TIP49 (RUVBL1)</td>
<td>TIP48p; TIP49p</td>
<td></td>
<td>Bridge interactions between the 15.5 kDa and the NOP56 and NOP58 proteins</td>
</tr>
<tr>
<td>(Continued)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Continued)
<table>
<thead>
<tr>
<th>Protein</th>
<th>Vertebrate</th>
<th>Yeast</th>
<th>Archaea</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAF9 (TAF9)</td>
<td></td>
<td></td>
<td></td>
<td>C/D box RNA assembly</td>
</tr>
<tr>
<td>NUFIP (NUFIP1)</td>
<td></td>
<td></td>
<td></td>
<td>Regulation of TIP48 and TIP49 interactions with core C/D box RNA proteins</td>
</tr>
<tr>
<td>Hsp90 (HSP90AA1)</td>
<td></td>
<td></td>
<td></td>
<td>Assists folding of NHP2 and 15.5 kDa proteins</td>
</tr>
<tr>
<td>RNA-processing proteins</td>
<td>Xrn1 (XRN1); Xrn2 (XRN2)</td>
<td>Xrn1p; Rat1p/ Xrn2p</td>
<td></td>
<td>5′ → 3′ exonuclease trimming</td>
</tr>
<tr>
<td>Rrp6 (EXOSC10); Rrp46 (EXOSC5)</td>
<td>Rrp6p; Rrp46p</td>
<td></td>
<td></td>
<td>3′ → 5′ exonuclease trimming</td>
</tr>
<tr>
<td>Rnt1p</td>
<td></td>
<td></td>
<td></td>
<td>Excision of snoRNA from hairpin (RNase III activity)</td>
</tr>
<tr>
<td>Lsm; La (SBB)</td>
<td>Lsmp; Lhp1p</td>
<td></td>
<td></td>
<td>Stabilization of mature 3′-end of snoRNA</td>
</tr>
<tr>
<td>TGS1 (TGS1)</td>
<td></td>
<td></td>
<td></td>
<td>Hypermethylates m1G-gap of snoRNAs from independent genes</td>
</tr>
<tr>
<td>Prp43; Has1p</td>
<td></td>
<td></td>
<td></td>
<td>Release/association of snoRNAs with rRNA (helicase activity)</td>
</tr>
<tr>
<td>Localization proteins</td>
<td>PHAX (PHAX)</td>
<td></td>
<td></td>
<td>Transport of C/D box RNAs to Cajal bodies</td>
</tr>
<tr>
<td>CRM1 (XPO1)</td>
<td>CRM1</td>
<td></td>
<td></td>
<td>Transport of C/D box RNAs from Cajal bodies to nucleoli</td>
</tr>
<tr>
<td>Ran (RAN)</td>
<td></td>
<td></td>
<td></td>
<td>snoRNA trafficking, nuclear export</td>
</tr>
<tr>
<td>CBP20 (NCBP2); CBP80 (NCBP1)</td>
<td></td>
<td></td>
<td></td>
<td>Cap-binding complex proteins</td>
</tr>
<tr>
<td>Nopp140 (NOLC1)</td>
<td>Srp40p</td>
<td></td>
<td></td>
<td>Nucleocytoplasmic transport factor</td>
</tr>
</tbody>
</table>
Despite conservation of some sequence and structural features, the variation of snoRNAs in structure, length and genomic context across different phyla has hampered efforts to create general universally effective computational search methods for the two major types of snoRNA. On the basis of the presence of associated proteins in all eukaryotic and archaeal species, and careful study of a small number of model species, we estimate that the majority of snoRNA genes have not been identified or annotated properly for most species in public databases [92]. BLAST [93], a popular sequence similarity search algorithm, is poor at identifying homologues because of the rapid divergence of guide sequences [40] and the relatively short length of the conserved sequence features. Although a portion of methylation sites are widely conserved, many are conserved only between closely related species or within phyla [40]. Recent discoveries that stretch the possibilities of functional H/ACA and C/D box RNA structure [46,94,95] illustrate the need for computational programs that are specialized by clades (Table 2).

The recent availability of high-throughput sequencing data strengthens computational searches by narrowing 'search space' and confirming predictions with transcriptional data. The increasing availability of sequenced genomes also enables the use of comparative genome analysis in the search for snoRNAs [23,64] and ncRNAs in general [96]. RNA sequencing libraries that are enriched for small RNAs has greatly assisted bioinformatics searches for H/ACA and C/D box RNAs [46,51,68,97]. Some existing programs increased specificity by determining antisense targets of guide regions [21,98], but miss orphan guides. With transcriptional data, we can confirm the presence of predicted snoRNAs with unknown targets [68]. In the archaeal species Pyrobaculum aerophilum, the number of C/D box snoRNAs was increased ~37% (65–89 genes) by using transcriptional data and eliminating the requirement for antisense RNA

### Table 2. Computational snoRNA prediction methods and databases

<table>
<thead>
<tr>
<th>Program or database</th>
<th>Description</th>
<th>Website</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>snoTarget</td>
<td>Mammalian C/D box RNA predictor</td>
<td>Server: <a href="http://bpg.utoledo.edu/~dbs/snotarget/">http://bpg.utoledo.edu/~dbs/snotarget/</a></td>
<td>[86]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standalone: <a href="http://www.utoledo.edu/med/depts/bioinfo/snotarget.html">http://www.utoledo.edu/med/depts/bioinfo/snotarget.html</a></td>
<td></td>
</tr>
<tr>
<td>snoscan</td>
<td>C/D box RNA predictor</td>
<td>Server and standalone: <a href="http://lowelab.ucsc.edu/snoscan/">http://lowelab.ucsc.edu/snoscan/</a></td>
<td>[21]</td>
</tr>
<tr>
<td>snoGPS</td>
<td>Human, yeast, archaeal H/ACA RNA predictor</td>
<td>Server and standalone: <a href="http://lowelab.ucsc.edu/snoGPS/">http://lowelab.ucsc.edu/snoGPS/</a></td>
<td>[121]</td>
</tr>
</tbody>
</table>

(Continued)
Table 2. Computational snoRNA prediction methods and databases  (Continued)

<table>
<thead>
<tr>
<th>Program or database</th>
<th>Description</th>
<th>Website</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>snoSeeker</td>
<td>Mammalian C/D and H/ACA box RNA predictor</td>
<td>Server and standalone: <a href="http://genelab.sysu.edu.cn/snoseeker/index.php">http://genelab.sysu.edu.cn/snoseeker/index.php</a></td>
<td>[122]</td>
</tr>
<tr>
<td>ERPIN and RNAMOT</td>
<td>Archael H/ACA RNA prediction method</td>
<td>Method outlined in reference</td>
<td>[123]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ERPIN: <a href="http://tagc.univ-mrs.fr/erpin">http://tagc.univ-mrs.fr/erpin</a></td>
<td></td>
</tr>
<tr>
<td>RNAsnoop</td>
<td>H/ACA box RNA target predictor</td>
<td>Standalone: <a href="http://www.bioinf.uni-leipzig.de/~htafer/RNAsnoop/RNAsnoop.html">http://www.bioinf.uni-leipzig.de/~htafer/RNAsnoop/RNAsnoop.html</a></td>
<td>[124]</td>
</tr>
<tr>
<td>PLEYX</td>
<td>C/D box RNA target predictor</td>
<td>Standalone: <a href="http://www.bioinf.uni-leipzig.de/Software/PLEXY">http://www.bioinf.uni-leipzig.de/Software/PLEXY</a></td>
<td>[125]</td>
</tr>
<tr>
<td>Psiscan</td>
<td>Trypanosome H/ACA box RNA predictor</td>
<td>Method described in reference</td>
<td>[126]</td>
</tr>
<tr>
<td>Plant snoRNA Database</td>
<td>Plant snoRNA sequences and target sites</td>
<td><a href="http://bioinf.scri.sari.ac.uk/cgi-bin/plant_snorna/home">http://bioinf.scri.sari.ac.uk/cgi-bin/plant_snorna/home</a></td>
<td>[127]</td>
</tr>
<tr>
<td>Yeast snoRNA Database at UMass-Amherst</td>
<td>Yeast snoRNA sequences and target sites</td>
<td><a href="http://people.biochem.umass.edu/fournierlab/snornadb/main.php">http://people.biochem.umass.edu/fournierlab/snornadb/main.php</a></td>
<td>[128]</td>
</tr>
<tr>
<td>Rfam Database</td>
<td>Archaeal and eukaryotic snoRNA sequences and families</td>
<td><a href="http://rfam.sanger.ac.uk/">http://rfam.sanger.ac.uk/</a></td>
<td>[102]</td>
</tr>
<tr>
<td>snoRNABase</td>
<td>Human snoRNA sequence and target sites</td>
<td><a href="http://www-snorna.biotoul.fr/">http://www-snorna.biotoul.fr/</a></td>
<td>[37]</td>
</tr>
<tr>
<td>Lowe/Eddy snoRNA Database</td>
<td>Yeast, archaea and Arabidopsis thaliana C/D box RNA sequences and alignments</td>
<td><a href="http://lowelab.ucsc.edu/snoRNAdb/">http://lowelab.ucsc.edu/snoRNAdb/</a></td>
<td>[21,23,127]</td>
</tr>
</tbody>
</table>
targets [68], adding substantially to a list of carefully curated predictions that had existed for 10 years [99].

A variety of snoRNA databases exist, with most entries originally (or only) identified by bioinformatics screens. The lack of a common nomenclature can lead to confusion when comparing snoRNA homologues and pseudogenes [100], although the nomenclature SNORD, SNORA and SCARNA has been developed for human C/D box RNAs, H/ACA box RNAs and scaRNAs respectively [101]. Many of these databases also include information about predicted target sites. One notable resource, the Rfam database [102] does not annotate predicted targets of snoRNAs due to its general RNA search methodology [103], but it acts as a valuable repository of most classes of RNA, including specific snoRNA genes that are conserved between three or more species.

**snoRNP biogenesis**

snoRNAs form the nucleating backbone of snoRNPs, to which associating proteins bind to form the mature RNP. Study of the assembly of these complexes aids our understanding of diseases affected by snoRNPs and rRNA biogenesis. In addition, the differences between the archaeal and eukaryotic versions reveal subtle changes in the evolutionary interplay between the RNAs and their associated proteins.

**C/D box RNPs**

In archaea, the conventional model of C/D box RNPs is ‘symmetrical’, and involves one snoRNA molecule per complex (Figure 4A; reviewed in [44]). First, the L7Ae protein initiates assembly by binding to the kink-turn formed by the C and D boxes, followed by a second L7Ae protein binding to the K-loop formed by the C′ and D′ boxes. Next, two Nop56/58 (also known as Nop5) proteins bind, one at each pole of the snoRNA. Two fibrillarin proteins then complete the symmetrical assembly with a total of six proteins bound to a single RNA. Interestingly, loss of L7Ae K-loop binding does not abolish its incorporation into the C′/D′ portion of the RNP, nor does it destroy methylation of the target nucleotide of the D′-guide. Likely protein–protein interactions are an important factor in C/D box RNP assembly, not just initiation by L7Ae [104].

Follow-up studies of the archaeal snoRNP have offered a new alternative model. In 2009, Bleichert et al. [105] proposed a dimeric C/D box RNP structure based on EM (electron microscopy) data of an archaeal C/D sRNP (Figure 4B). This ‘di-sRNP’ consists of two C/D box RNAs and four of each core protein, contrary to the conventional model, and can be extended to the eukaryotic model (Figure 4D). As a counterpoint to the di-sRNP model, a 2011 study by Lin et al. [106] synthesized two RNAs that could form kink-turn structures and act as a scaffold for Nop5, L7Ae and fibrillarin [106]. This study resulted in a crystal structure of a monomeric RNP that also had proven in vitro activity. However, the structure in the Lin et al. [106] study did not contain a K-loop between the C′ and D′ boxes, which is typical of archaeal C/D box RNAs and thus may call into question the applicability of the results. To investigate the importance of the K-loop, Bower-Phipps and Taylor [107] demonstrated in 2012 that nearly all C/D box sRNPs with an internal loop in their C/D sRNA do adopt the di-sRNP architecture. By testing C/D box sRNAs from several widely divergent archaeal species, this study concluded that the di-sRNP structure is likely to be conserved across the archaeal
domain. It remains unclear if computationally predicted (but functionally unverified) C/D box sRNAs that lack a K-loop could potentially assemble naturally into the monomeric RNP form.

Less is known about eukaryotic C/D box RNP structure, but it appears to be 'asymmetric' since the 15.5 kDa protein only binds to the kink-turn formed by the terminal C and D boxes (Figure 4C). On the basis of in vivo cross-linking experiments in oocytes of the African clawed frog, *Xenopus laevis*, Nop56 is predicted to bind to the C'/D' motifs and Nop58 to the C/D motifs [108]. Fibrillarin is predicted to bind both sets of motifs. In contrast, a study in the yeast *S. cerevisiae* indicates that all four core proteins associate with both the C/D and C'/D' motifs, on the basis of results of co-immunoprecipitation experiments [109]. This work

**Figure 4. C/D box sRNP models**

Fully assembled C/D box sRNPs differ between Archaea and Eukarya, and various models exist based on crystal structures and EM data. The model illustrated here is based on Figures in [129]. L7Ae/15.5 kDa, yellow; Nop56/58 and Nop56, light blue; Nop58, purple; fibrillarin, green; and C/D box RNA, dark blue. (A) Conventional 'symmetrical' archaeal C/D box RNP model supported by crystallographic data [106]. (B) Proposed archaeal di-sRNP model. The di-sRNP model is based on an EM-reconstruction of archaeal C/D box RNPs from [107]. (C) Conventional 'asymmetrical' eukaryotic C/D box RNP model. This architecture is supported by biochemical data [108,130]. In the 'asymmetrical' eukaryotic model, Nop58 binds to the C/D motifs and Nop56 to the C'/D' motifs and the 15.5 kDa protein (homologous with L7Ae) only binds to the kink-turn formed by the C and D boxes. (D) Proposed eukaryotic di-sRNP model.
demonstrated that the individual C/D and C′/D′ RNPs are coupled spatially and functionally for methylation activity [109]. The association of the 15.5 kDa protein with the C′/D′ motif, despite its inability to bind the K-loop formed by the motif, suggests the possibility that the eukaryotic C/D box RNP structure is symmetrical, like the archaeal version. This interaction echoes the ability of archaeal L7Ae to associate with the C′/D′ motif even if it is mutated so it can no longer bind to the K-loop [104]. Similar to archaeal C/D box RNP assembly, protein–protein interactions play a large role in the formation of the snoRNP complex.

In eukaryotes, a growing number of proteins have been implicated in the biogenesis of C/D box RNPs (see Table 1). Among them are the proteins IBP160 and Bcd1 (box C/D RNA 1) which are implicated as assembly factors [4]. IBP160 is a general splicing factor that binds upstream of intronic C/D box RNAs and starts the biogenesis of the C/D box RNP. Bcd1 is necessary for the accumulation of C/D box RNAs, but not a member of the final active C/D box RNP. The role of other identified assembly and localization factors such as CRM1 is unclear and the subject of current research (reviewed in [8]).

H/ACA box RNPs

Owing to differences between archaeal L7Ae and the eukaryotic NHP2, the assembly of archaeal and eukaryotic H/ACA box RNPs differs (reviewed in [36]). NHP2 belongs to the same protein family as L7Ae, but does not bind kink-turns like its archaeal homologues. For the archaeal complex, L7Ae and Cbf5 bind directly to the guide RNA. Nop10 and Gar1 bind to Cbf5 independently, forming a stable heterotrimeric complex. In contrast in eukaryotes, NHP2 does not bind RNA directly, and must be recruited to the H/ACA box RNP by protein–protein interactions, forming a complex with dyskerin (homologue with Cbf5) and Nop10. This protein complex binds to the hairpin of the H/ACA box RNA. Finally, Gar1, named for the GAR (glycine-arginine rich) domains flanking the central domain of the protein, binds to dyskerin.

Compared with archaeal H/ACA box RNP assembly, eukaryotic H/ACA box RNP assembly is complex, requiring multiple assembly factors (reviewed in [8]). Naf1 (nuclear assembly factor 1) appears to be involved in the biogenesis of H/ACA box RNPs, but is not necessary for the active RNP, where Gar1 takes the place of Naf1. Gar1 contains a C-terminal extension that regulates substrate turnover and allows Cbf5 to bind tighter than Naf1 [110]. Naf1 is needed for the accumulation of H/ACA box RNPs, so the exchange of Naf1 for Gar1 may be a key step in regulating the activity of this complex. Like Naf1, Shq1 is a nucleoplasmic shuttle protein and may act as a placeholder for the H/ACA box RNA during RNP assembly [111].

Evolution of L7Ae and kink-turn-binding proteins

snoRNPs of both eukaryotes and archaea contain kink-turn-binding proteins or homologues that belong to the L7Ae/L30e protein family. The RNA kink-turn motif is characterized by stacked sheared GA base pairs flanked by two stems, or by a stem and a loop in the case of K-loops. This motif acts as the binding site for L7Ae and its homologues (e.g. 15.5 kDa protein), nucleating the assembly of the RNP. In the case of eukaryotic H/ACA box RNPs, NHP2 belongs to the L7Ae/L30e protein family, but does not appear to have binding specificity to kink-turns.

The existence of only two kink-turn-binding proteins in archaea and six in eukaryotes indicates diversification of context and/or function across the tree of life [44]. Other
eukaryotic members of this family occur in specialized RNPs; for example, the SBP2 protein is required for selenocysteine incorporation, mediated by binding a kink-turn structure in the 3′ UTR (untranslated region) of selenoprotein mRNAs. Recently, YbxF and YlxQ were demonstrated to be bacterial homologues of L7Ae [112]. These proteins bind kink-turns, but not as strongly as L7Ae, and do not bind K-loops. In eukaryotes, studies suggest that Hsp (heat-shock protein) 90 works in concert with conserved cofactors to control the biogenesis of RNPs with kink-turn-binding proteins [113]. L7Ae in archaea has also unexpectedly been found to be part of the tRNA-processing RNase P complex [114]. These findings invite broader questions about possible regulatory cross-talk between diverse RNA processing pathways, including snoRNPs, which appear to compete for a common component.

Conclusions

Previously thought to have a singular role in ribosome biogenesis, we now have diverse evidence that snoRNAs play a broader role in the cell. This is exemplified by cases of dysregulation of these RNAs that have unanticipated effects on the organism. Protein translation, premRNA splicing, telomere stability and thus cell viability depend on the function of snoRNAs. The various roles of snoRNAs in translation suggest that snoRNAs may have originated from an early translation system [29]. Future elucidation of the versatility of snoRNA function will advance from a mix of biochemical, computational and high-throughput sequencing approaches.

Summary

- Guide C/D box RNPs and H/ACA box RNPs are defined by their ability to carry out precise 2′-O-methylation and pseudouridylation of target nucleotides.
- The assembly of snoRNPs (small nucleolar RNAs) is complex and differs in some aspects between the eukaryotic and archaeal domains.
- snoRNAs are transcribed from a variety of genomic contexts, illustrating evolutionary diversity and maturation pathways.
- snoRNAs have functions and roles beyond post-transcriptional modification, such as ribosomal RNA cleavage, regulation of gene expression by RNA silencing and alternative splicing, and telomerase maintenance.
- Disruption and mutations to snoRNPs can cause human diseases.
- Computational search methods exist for snoRNAs, but the diversity in sequence motifs strongly suggests that other forms of snoRNAs probably exist that current methods do not detect.

References


25. Clouet d’Orval, B., Bortolin, M.L., Gaspin, C. and Bachellerie, J.P. (2001) Box C/D RNA guides for the ribose methylation of archaeal tRNAs. The tRNA\textsubscript{Trp} intron guides the formation of two ribose-methylated nucleosides in the mature tRNA\textsubscript{Trp}. Nucleic Acids Res. 29, 4518–4529


