# Nematode sbRNAs: homologs of vertebrate Y RNAs

*Ilenia Boria*[*,a,b], *Andreas R. Gruber*[*,b,+], *Andrea Tanzer*[*,b,c], *Stephan Bernhart*[b], *Ronny Lorenz*[b], *Michael M. Mueller*[d], *Ivo L. Hofacker*[b], *Peter F. Stadler*[c,e,f,b,g]

[a]Department of Medical Sciences and Interdisciplinary Research Centre for Autoimmune Diseases, Università del Piemonte Orientale, via Solaroli 17, I-28100 Novara, Italy. [b]Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria. [c]Bioinformatics Group, Department of Computer Science; and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany [d]Department of Chromosome Biology, Max F. Perutz Laboratories, University of Vienna, A-1030 Vienna, Austria. [e]Max-Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany. [f]Fraunhofer Institut für Zelltherapie und Immunologie (IZI), Perlickstraße 1, D-04103 Leipzig, Germany. [g]Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA.

[*]These authors contributed equally [+]Corresponding author: Tel.: +43 5277 52731 Fax: +43 4277 52793.

Stem bulge RNAs (sbRNAs) are a group of small, functionally yet uncharacterized noncoding RNAs found initially in *C. elegans* with a few homologous sequences postulated in *C. briggsae*. In this study we report on a comprehensive homology-based survey of this ncRNA family in the phylum Nematoda, resulting in a total of 233 new sbRNA homologs. For the majority of these hits promoter regions and transcription termination signals characteristic for pol-III transcripts were identified. Surprisingly, sequence and structure comparison with known RNA families revealed that sbRNAs are homologs of vertebrate Y RNAs. Most of the sbRNAs show the characteristic Ro protein binding motif, and have in addition a region highly similar to a functionally required motif for DNA replication previously thought to be unique to vertebrate Y RNAs. The single Y RNA that was previously described in *C. elegans*, however, does not show this motif, and in general bears the hallmarks of a highly derived family member.

## 1. Introduction

Stem-bulge RNAs (sbRNAs) were discovered in the nematode *C. elegans* three years ago in a systematic screen of a ncRNA-specific full-length cDNA library by Deng et al. (2006). This initial study identified 9 distinct members of this family. In a subsequent contribution, Aftab et al. (2008) listed three additional experimentally verified ncRNAs were annotated as sbRNAs. These seed sequences are listed in Tab. 1. These sequences share two two conserved internal motifs at the 5'- and 3'-end of the molecules, respectively. Computational predictions showed that these regions are able to form a long stem interrupted by a small bulge, accounting for the name of this ncRNA family. A `blast`-based comparison with the *C. briggsae* genome revealed eleven putative homologs (Deng et al. 2006), providing support for the stem-structure and indicating that loops region evolves rapidly.

The sbRNAs in *C. elegans* as well as their *C. briggsae* homologs show a promoter structure consisting of a proximal sequence element B (PSE B) and a TATA-box (Deng et al. 2006). This type of pol-III promoter is closely related to that of snRNAs (Hernandez 2001), from which it differs by the lack of the conserved PSE A box in the proximal element, see Fig. 1 top. In a subsequent, detailed analysis of the sbRNA promoter, Li et al. (2008) showed that in contrast to the other promoters analyzed, transcription, albeit reduced by 30 to 50%, could also be seen when only one of the two parts of the promoter (either PSE B or TATA-box) was present. Taken together with the fact that sbRNAs are uncapped and terminate with a poly-U stretch, these observations leave little doubt that sbRNAs transcribed by RNA polymerase III.

Key words: sbRNA, nematodes, Y RNA, homology search, noncoding RNA

E-mail: agruber@tbi.univie.ac.at

*Preprint* 1–10. 2009

May 29, 2009



Table 1 Seed set of sbRNAs.
All twelve sbRNAs are found in the ncRNA set identified by Deng et al. (2006). Ref. **b** indicates that they were first annotated as sbRNA by Aftab et al. (2008). The sequences marked **c** were also reported in Zemann et al. (2006). RNAi experiments were conducted for sequences marked **d** (Kamath et al. 2003) and **e** (Sönnichsen et al. 2005). A Y RNA homolog computationally predicted in Perreault et al. (2007) is marked by **f**.

| **Name** | **Wormbase** | **Acc.No.** | *L* | **Refs.** |
|---|---|---|---|---|
| CeN71 | F08G2.13 | AY948635 | 74 | **c** |
| CeN72 | – | AY948636 | 98 | |
| CeN73-1 | – | AY948637 | 133 | |
| CeN73-2 | – | AY948638 | 131 | |
| CeN74-1 | M163.13 | AY948639 | 79 | **c** |
| CeN74-2 | M163.12 | AY948640 | 77 | **c** |
| CeN75 | – | AY948593 | 70 | |
| CeN76 | W01D2.8 | AY948641 | 77 | |
| CeN77 | fragmented | AY948602 | 69 | |
| CeN135 | F08G2.12 | AM286261 | 67 | **b**,**d** |
| CeN133 | W01D2.7 | AM286259 | 95 | **b**,**d**, **e** |
| CeN134 | F35E12.11 | AM286260 | 119 | **b**,**f** |

Most sbRNAs are differentially expressed in developmental stages, where mature adult worms, *dauer larvae* and especially worms after heat shock have the highest levels of expression (Deng et al. 2006). In an unrelated study focusing on the snoRNAs complement of *C. elegans*, Zemann et al. (2006) confirmed two of Deng's sbRNAs. For two sbRNAs (CeN135 and CeN133), along with almost 20,000 other genes, a knock-down experiment was performed Kamath et al. (2003). No phenotype was reported for these two knock downs. One sbRNA was also knocked down in a study by Sönnichsen et al. (2005), again with no visible phenotype. The negative outcome of these knock down experiments is not surprising, however, given that the many paralogous sbRNAs in the *C. elegans* genome can be expected to functionally compensate for the lost molecule.

A first attempt to gain insight into the putative biolog-

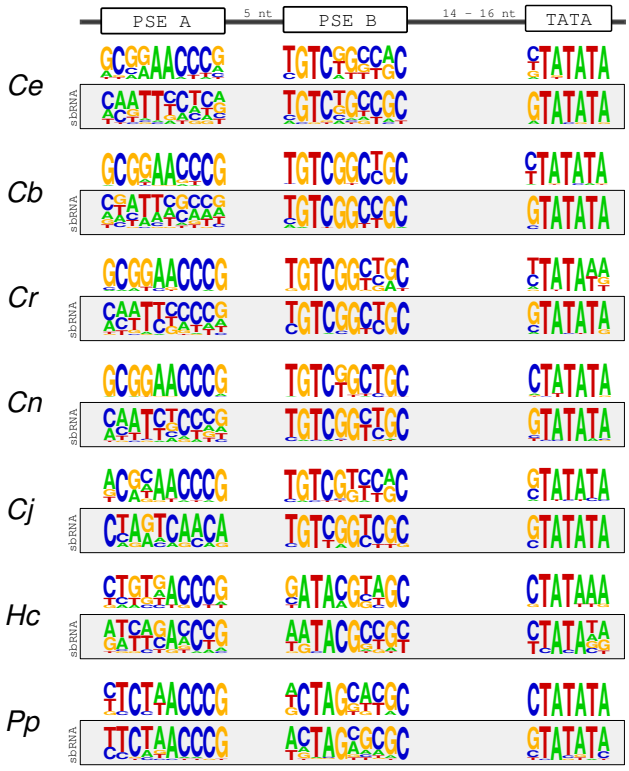



FIG. 1.—Comparison of promoter elements of sbRNAs to other pol-III transcripts. The upper row for each species shows sequence logos (Crooks et al. 2004) of the promoter motifs for other pol-III transcripts, while the lower row denotes the corresponding elements for sbRNAs. High correlation is observed for the PSE B and the TATA-box for all species, while high correlation for PSE A is only observed for *H. contortus* and *P. pacificus*. Abbreviations: Ce - *C. elegans*, Cb - *C. briggsae*, Cr - *C. remanei*, Cn - *C. brenneri*, Cj - *C. japonica*, Hc - *H. contortus*, Pp - *P. pacificus*.

ical functions to sbRNAs is reported by Aftab et al. (2008). Some sbRNAs showed increased levels of expression after depletion of the protein components of the snoRNPs. A detailed understanding of these findings is still missing and up to now biological function and processes sbRNAs are involved in remain to be uncovered.

In this contribution we report on a comprehensive homology search for sbRNAs in the phylum Nematoda, and on an in depth analysis of the large gene family uncovered by this survey.

## 2. Materials and Methods

### 2.1 *Sequence Data*

Nematode genomic sequences were downloaded from WormBase (WS198, `www.wormbase.org`), the Sanger Institute (`www.sanger.ac.uk`), TraceDB (`www.ncbi.nlm.nih.gov/pub/TraceDB`), the Sophia-Antipolis Institute (`meloidogyne.toulouse.inra.fr`) (Abad et al. 2008), and the *M. hapla* Genome Sequencing Group (`www.hapla.org`). Details on the assemblies used here are listed in the Electronic Supplement. Phylogenetic relationships of the investigated species are depicted in Fig. 2.

### 2.2 *Sequence-Based Homology Search*

Starting from an initial set of experimentally verified sbRNAs, listed in Tab. 1, we performed a `blastn` search with default parameters against the available genome assemblies of nematode species. In addition, we extracted putative sbRNA sequences from the `multiz` 6-way alignments of nematode species available at the UCSC Genome browser (`genome.ucsc.edu`) for known *C. elegans* sbRNA loci.

### 2.3 *Homology Search with Promoter Elements*

We applied a computational promoter search using the characteristic promoter elements of sbRNAs (PSE B and TATA-box) in species of the genus *Caenorhabditis*, in *P. pacificus* and in *H. contortus*. In the first step we extracted regions 200 nt upstream of RNase P, RNase MRP, U6 snRNAs, and tRNA-Sec. These noncoding RNAs are known to also have PSE B and TATA-Box promoter elements. For *C. elegans* the corresponding sequences for RNase P, RNase MRP, and tRNA-Sec could easily be retrieved from annotated Wormbase entries (rpr-1, mrpr-1, K11H12.t1) or in case of U6 snRNAs from the published literature (Dávila López et al. 2008; Marz et al. 2008). Simple `blast` searches were sufficient to identify orthologs in other species. We then created multiple sequence alignments of the upstream regions using `Jalview` (Waterhouse et al. 2009) for each species, marked blocks corresponding to the PSE B and the TATA-box and generated a `fragrep` (Mosig et al. 2007a) search pattern. The `fragrep` search resulted in 1,200 hits in *C. remanei* and even more moderate numbers for the other nematodes. For each hit we searched the 300nt of genomic DNA downstream of the putative promoter regions for a possible terminator consisting of a consecutive run of at least four T residues. The region ranging from 20 nt downstream of the TATA-box to the terminator was extracted.

This approach offers two major advantages over purely sequence based or model based searches: (i) the initial filtering of the genomic data is not restricted by limited knowledge on the variability of the sequence and/or structure of the ncRNA itself, and (ii) the canonical promoter structure lends additional credibility to the candidates. The feasibility of this strategy we recently demonstrated for the 7SK RNAs of arthropods (Gruber et al. 2008).

Sequence-structure based clustering using the `locarna-RNAclust` pipeline (Will et al. 2007; Kaczkowski et al. 2009) was then applied to all these sequences. Clusters were then visually examined for sequence-structure similarity to already identified sbRNAs using the `RNAsoupViewer` (`www.bioinf.uni-leipzig.de/pages/40/software.html`).

### 2.4 *Model-Based Homology Search*

Multiple sequence alignments of the seed sequences and the hits of both the sequence-based homology search and the promoter screen were constructed manually. We used the `RALEE` mode (Griffiths-Jones 2005) in `emacs` which explicitly handles secondary structure annotation. `RNAalifold` (Hofacker et al. 2002) predictions for

closely related sequences were used as starting point for deriving a consensus structure for the well-conserved parts.

These structure-annotated alignments were used to deduce a non-stringent sequence/structure model (available in the Electronic Supplement), which was then employed to screen the nematode genomes with `rnabob` (`selab.janelia.org/software.html`). The resulting initial candidates were filtered using a modified position weight matrix scoring in which base-pairs are treated like individual letters. Let $\mathscr{A} = \{A,C,G,T\}$ be the nucleotide alphabet. Then $\mathscr{B} = \{AA,AC,AG,AT,...,TT\}$ is the alphabet of all standard and non-standard base pairs. The modified equation for the information vector $I$ at position $i$ in the approach of Kel et al. (2003) is

$$I(i) = \sum_{b \in \mathscr{A} \text{ or } \mathscr{B}} f_{i,b} \ln(k(b)\, f_{i,b}) \qquad (1)$$

where $i$ is now either an unpaired nucleotide or a base pair, and $k(b) = 4$ if $b \in \mathscr{A}$ and $k(b) = 16$ if $b \in \mathscr{B}$. We implemented a `Perl` that takes the `rnabob` output and position weight matrices derived from the structural alignment as input and outputs `RNABOB` hits augmented by a *matrix similarity score* (mSS) as defined by Kel et al. (2003). Hits with a mSS higher than 0.65 were then compared manually to previously identified sbRNAs.

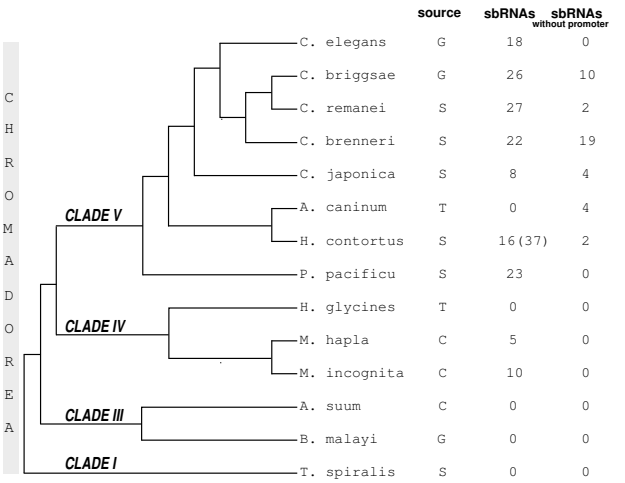



FIG. 2.—Phylogenetic distribution of the 233 identified sbRNA homologs. Hits are divided into sbRNAs with confirmed promoter regions, and those hits that did not yield any significant homology to known ncRNA promoters. The column "source" denotes the assembly status of the sequences (T: Traces, C: contigs, S: supercontigs, G: chromosomal level). For *H. contortus* we found a hit with 37 adjacent copies. For the list of sbRNA with verified promoter regions this hit was just counted once.

### 2.5 *Identification of Promoter Elements*

For species of the genus *Caenorhabditis*, *P. pacificus*, and *H. contortus* we were able to collect a sufficient number of upstream regions of ncRNAs that at least partially share the same promoter elements as sbRNAs. We created position weight matrices (PWMs) for the PSE A, PSE B (each species separately) and a general TATA-box PWM and used the approach by Kel et al. (2003) to score corresponding elements in sbRNA upstream sequences. Sequence motifs corresponding to PSE A were only classified as reliable if their score was higher than 0.75 and if they were exactly located 5 nt upstream of a PSE B. Alignments and PWMs are available in the Electronic supplement.

### 2.6 *Mapping of Syntenic Regions*

For *C. elegans* we retrieved WormBase gene entries. For the other *Caenorhabditis* species, the coordinates of the mapped *C. elegans* were open reading frames downloaded as `bed` files from the UCSC Genome Browser. As the sequence repository at UCSC features different genome assemblies than those used in our analysis, we first mapped our sbRNA hits to those assemblies. This was done by simple `blast` searches. For each sbRNA we then extracted the name of proteins that reside in a window of 40kb upstream and downstream of that hit. Finally, files were merged to list for each protein sbRNAs that are located in its vicinity. Syntenic regions between different species were identified as overlaps of the corresponding gene list.

## 3. Results

### 3.1 *Homology Search*

Starting from the seed sequences, both the analysis of the `multiz` alignments and an iterative `blastn` search resulted only in a moderate number of additional homologs in the *Caenorhabditis* species and a few hits in *P. pacificus*, but failed to give any plausible candidate in other nematodes. In a second approach to identify new sbRNAs, we took advantage of the well characterised promoter elements of known sbRNAs (Li et al. 2008) and performed a computational promoter screen, a strategy that was recently employed successfully for another ncRNA family (Gruber et al. 2008). Initial candidates where used to construct a promiscuous pattern for an `rnabob` search, whose results where then filtered further using a PWM-based method to detect faint sequence similarities as described in detail in the Methods section.

After manual inspection of the search results, we retained a list of 233 sbRNAs distributed over the nematodes clade V (Strongylida, Diplogasterida, and Rhabditida) and clade IV (Tylenchida, Cephalobina, and Panagrolaimida), summarized in Fig. 2. In particular, we report a total of 18 sbRNAs for *C. elegans*, all with confirmed promoter elements. In the other species we also a significant number of sbRNAs that did not show significant matches to known ncRNA promoter elements. In *H. contortus* we identified one hit with several (37) adjacent copies on one contig. We cannot exclude that this might be an assembly artifact and therefore we count this hit just once in the list of sbRNA with promoter elements. Our survey failed to retrieve homologs in the genomes of *A. suum*, *B. malayi* and *T. spiralis* and in the shotgun trace sequences of *Heterodera glycines*.

### 3.2 *Analysis of Upstream Regions*

For *C. elegans* the core promoter of sbRNAs has been shown to consist only of a PSE B and a TATA-box (Li et al. 2008), while other polymerase III transcripts including the known Y RNA (Van Horn et al. 1995) have an additional, conserved element located 5 nt upstream of the PSE B called PSE A (Thomas et al. 1990; Missal et al. 2006). Comprehensive studies of snRNA promoters of this kind (pol-III type 3) have only been conducted for *C. elegans* in the phylum Nematoda (Li et al. 2008). For all other species we identified corresponding promoter elements by sequence and positional conservation.

A detailed analysis of the upstream regions of sbRNAs with position weight matrices used in the computational promoter screen revealed that the shortened core promoter characteristic for sbRNAs in *C. elegans* can only be found in the genus *Caenorhabditis*. Upstream sequences of sbRNAs in *P. pacificus* and *H. contortus* show the presence of both a PSE A and a PSE B. A detailed representation of the core promoter for these species is shown in Fig. 1 together with corresponding elements of other putative pol-III transcripts. For *M. hapla*, *M. incognita* and *A. caninum* we were not able to find a significant number of high confidence homologs of other pol-III transcripts to build reliable position weight matrices (PWMs) or to determine the exact position of PSEs and the TATA-box. In these cases upstream regions were just visually compared for stretches of homologous regions. Results of promoter analysis are summarized in Fig. 2.

### 3.3 *Secondary Structure*

In order to derive a consensus secondary structure, we used the subset of those 155 (out of 233) sbRNA homologs that exhibit clearly recognizable pol-III promoters to avoid contamination by possible pseudogenes. The structural alignment was constructed manually. Due to high sequence variation in the central loop this region remained unaligned and was investigated separately.

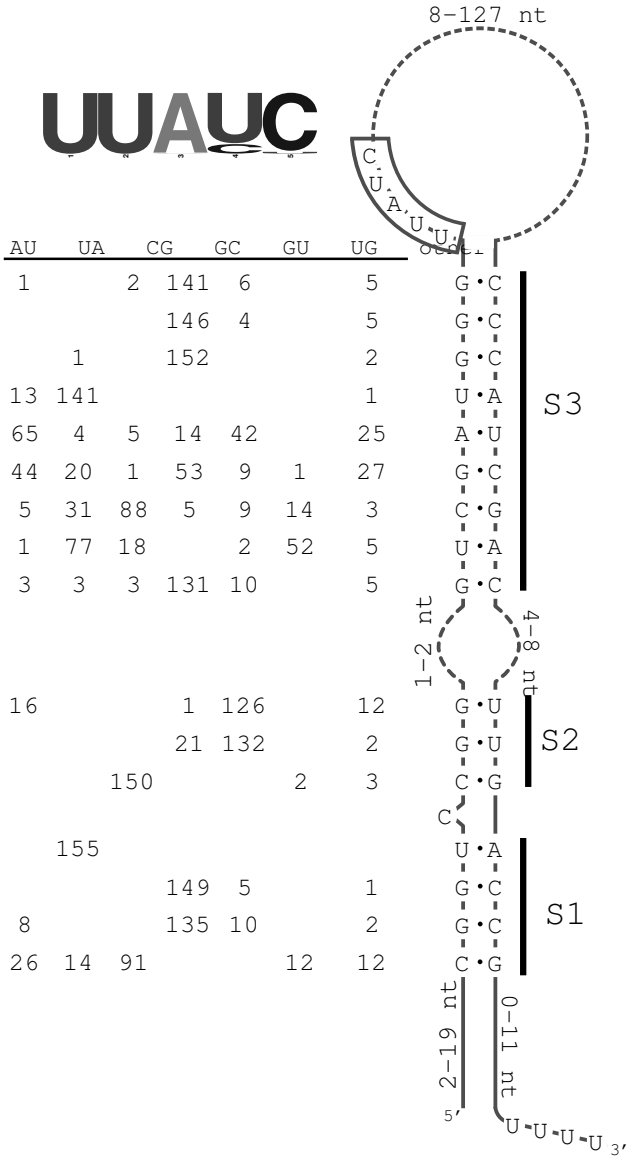



FIG. 3.—Secondary structure model of sbRNAs derived from 155 sbRNAs with verified promoter regions. The table on the right gives the absolute counts of base-pairs observed at a given position. The structure drawing displays just the most frequent base-pair. The sequence logo shows the frequencies of nucleotides for the motif UUAUC, which immediately follows the conserved stem. Just in two out of 233 sbRNAs we observed one or two additional G residues inserted between the stem and this motif.

The combination of thermodynamic structure predictions and phylogenetic analysis revealed several conserved structural elements, summarized in Figure 3. Nematode sbRNAs exhibit three conserved stem structures:

S1 Stem S1 is generally composed of four conserved base-pairs, but can be extended at the 5' end for most of the sequences. The closing 3' AU pair of stem S1 is the most conserved base-pair, all sequences can form that pair and no compensatory mutations are observed.

S2 Stem S2 is composed of three base-pairs only, where the majority of sequences shows two GU wobble-pairs. From a thermodynamic point of view it is a rather weak stem, supporting evidence for this stem is given by compensatory mutations.

S3 Stem S3 is composed of nine base-pairs. The 5' part of S3 shows a lot of compensatory mutations, which suggests that the ability to form this base-paired region is more important than the actual sequence. Stem S3 closes with three conserved GC pairs, preceded by a conserved UA pair. Only 13 sequences, all from *H. contortus*, show an AU pair at this position.

B Stems S1 and S2 are interspersed by a conserved single bulged cytosine.

I Stems S2 and S3 are separated by a small internal loop. Although some related sbRNAs show conservation of some nucleotide positions, this does not seem to be a general feature observed for the total set of sbRNAs.

H The central loop enclosed by the stem starts with the conserved sequence motif UUAUC. Detailed analysis of this motif showed that it is in general not involved in a structural context. For short sbRNAs, the entire central region is unstructured in general forming a single hairpin loop. In contrast, the longer sbRNAs homologs tend to form short, conserved structural elements.

T At the 3' end we generally observe a stretch of at least four U residues, which are believed to function as transcription termination signals. For most sbRNAs further poly U/T stretches can be observed downstream of their genomic location, which may serve as alternative termination signals (Gunnery et al. 1999; Guffanti et al. 2004).

### 3.4 *sbRNAs are Y RNAs*

Comparison to other RNA families revealed that nematode sbRNAs show high sequence/structure similarity to vertebrate Y RNAs (Mosig et al. 2007b; Perreault et al. 2007). One of the hits in the homology search conducted by Perreault et al. (2007) even matches CeN134. Figure 4 summarizes a detailed comparison of the sbRNA consensus with the analysis of vertebrate Y RNAs by Mosig et al. (2007b) and the Y RNAs from the genus *Caenorhabditis*. The latter were found using `GotohScan` (Hertel et al. 2009) starting from the experimentally known *C. elegans* CeY sequence (Van Horn et al. 1995).

In mammals, stem S1, the bulged cytidine (B) and stem S2 have been shown to be required for Ro binding (Green et al. 1998; Stein et al. 2005), and thus for the formation of the Ro RNP particles, which are involved in RNA quality control. These features are well conserved between Y RNAs (vertebrates and nematodes) and sbRNAs (Fig. 4B). This strongly suggests that sbRNAs contain a functional Ro binding site.

Recently, it has been shown that Y RNAs are also required for chromosomal DNA replication in human cell nuclei (Christov et al. 2006, 2008). The primary motif for this function resides at the 3' end of stem S3 and consist of a stretch of three base-pairs (denoted by red stars in Fig. 4A) (Gardiner et al. 2009). In particular the UA base-pair turned out to be crucial for Y RNA functionality in DNA replication. Indeed, *C. elegans* CeY and a Y RNA homolog from *D. radiodurans* (Chen et al. 2007), both lacking this feature, were not able to compensate for vertebrate Y RNAs in DNA replication. All sbRNAs with the exception of the 13 sequences of *H. contortus* also show the conserved UA base-pair at this position.

Overall, nematode sbRNAs show more similarities with vertebrate Y RNAs than the previously reported *Caenorhabditis* Y RNAs. In addition to unambiguous structure homology in the helical regions, the conserved loop motif UUAUC is also present in the two paralogous vertebrate subfamilies Y1 and Y3.

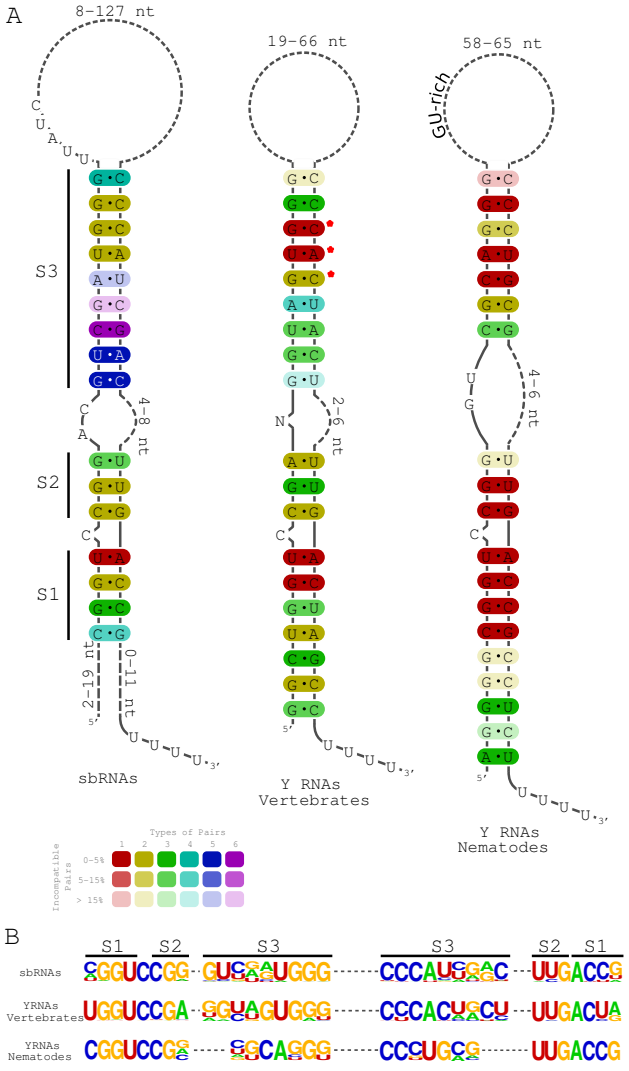



FIG. 4.—**A** Comparison of secondary structures for nematode sbRNAs, vertebrate Y RNAs and the previously described Y RNA family in the genus *Caenorhabditis*. Red stars denote the region identified by Gardiner et al. (2009) to be crucial for the function of Y RNA in DNA replication. **B** Sequence logos for helical regions S1, S2, and S3.

### 3.5 *Evolutionary History of sbRNAs*

In *C. elegans* we uncovered six new sbRNA homologs in addition to the twelve previously known sbRNAs. All six are supported by promoter elements (Tab. 3.5). Three hits have already been assigned a Wormbase ID, and for two of these there is evidence of transcription from a previously conducted study by Zemann et al. (2006). The same study annotated Ce7 as a C/D box snoRNA. These sequence yield a negative snoRNA classification by `snoReport` (Hertel et al. 2008) and can be unambiguous recognized as homologs to sbRNAs based on both sequence and secondary structure.

The 18 *C. elegans* sbRNAs identified to-date are organized in five clusters, Fig. 5. Each cluster consists of multiple copies of one sbRNA family. Thus, clusters seem to have arisen by local tandem duplications of one ancestral

Table 2 Newly identified sbRNA homologs in *C. elegans*. Hits marked with * are also reported by Zemann et al. (2006).

| Name | Location | Other names | *L* |
|---|---|---|---|
| Ce1 | intergenic | W01D2.7, Ce150* | 81 |
| Ce2 | intergenic | – | 85 |
| Ce3 | intronic | – | 155 |
| Ce5 | intergenic | – | 121 |
| Ce6 | intergenic | M163.15 | 83 |
| Ce7 | intergenic | M163.14, Ce94* | 98 |

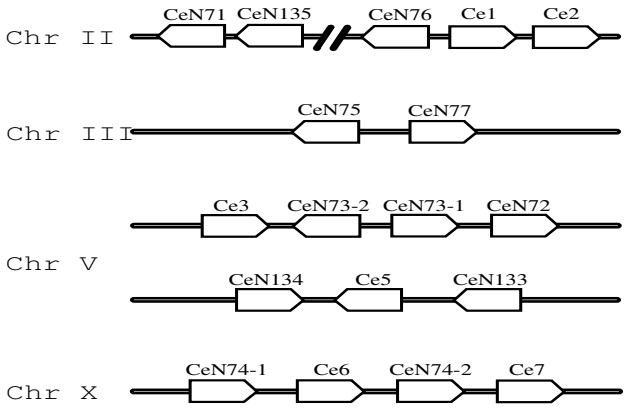



FIG. 5.—Schematic drawing of the organization of the five sbRNA clusters in *C. elegans*.

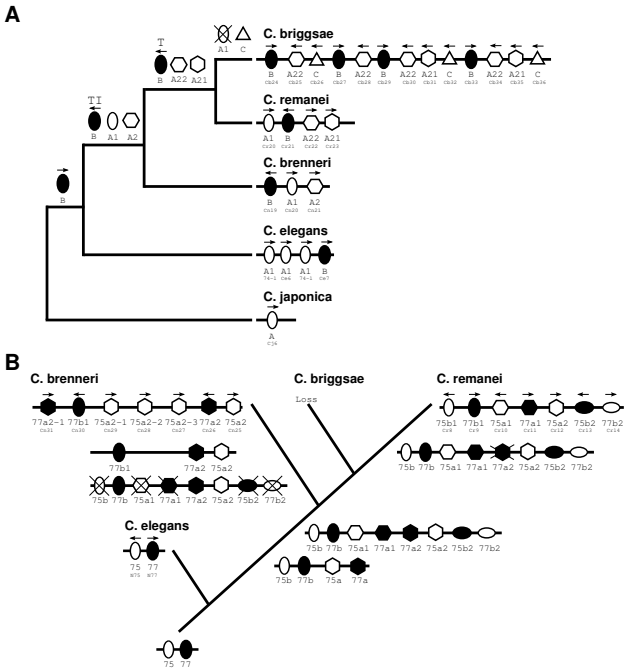



FIG. 6.—Evolutionary history of sbRNA clusters on (**A**) chrX an **B** chrIII of *C. elegans*. In both cases clusters are shaped by duplications of single genes as well as units of sbRNAs followed by deletions of individual genes. For details see text. Arrows indicate sbRNA orientation: plus strand (→) and minus strand (←); deleted genes are crossed out; T: transposition, I: inversion.

sbRNA. The mechanism by which the sbRNAs were multiplied remains unknown. Nevertheless, we find evidence, that not only single genes, but also groups of several sbRNAs might be effected by a single duplication event.

Due to the rapid evolution of the relatively short sbRNA sequences it is impossible to derive a reliable gene phylogeny based on sequence information alone. We therefore follow the strategy introduced for microRNA clusters by Tanzer and Stadler (2004). Furthermore, we systematically included synteny information. Syntenic clusters were identified in the genus Caenorhabditis based on their flanking protein coding genes (see Methods for details). Surprisingly, syntenic conservation can be established only for two of the five clusters: those located on *C. elegans* Chr.III and Chr.X. For the other clusters, only the sequence information could be used.

Standard phylogenetic methods are not applicable because the loop-part of the sbRNAs cannot be reliably aligned, while at the same time the better conserved stems barely contain phylogenetic information. We therefore used a *z*-score approach (Tanzer and Stadler 2004). In brief, the significance of pairwise alignments is assessed against the score distribution of alignments of the shuffled sequences. The *z*-scores are then used as similarity measure in a hierarchical clustering. The sbRNAs are homologs of Y RNAs and appear to be slightly closer related to vertebrate Y RNAs than to the previously described nematode Y RNAs.

In vertebrates, Y RNAs show features required for their known functions in DNA replication and binding to Ro. Their nematode homologs apparently underwent subfunctionalization so that sbRNAs and Y RNAs contain different features, Fig. 4. The exact time point of the divergence of sbRNAs and the CeY lineage cannot be determined with any certainty. While the *z*-score clustering points at an early divergence, CeY homologs were detectable within the genus Caenorhabditis only, suggesting a late duplication. Within Caenorhabditis, we observe a rapid radiation of divergent sbRNAs.

THE SBRNA CLUSTER ON CHROMOSOME III. Both sequence similarity and cluster organization indicates that the Chr.III cluster has undergone different complex fates in each species, comprising multiple local duplication and deletion events. Duplication of the ancestral sb-75/sb-77 pair resulted in tandem copies sb-75b/sb-77b and sb-75a/sb-77a. These four sbRNAs were then duplicated and inverted leading to a total of eight sbRNAs. Only three of the eight sbRNAs were retained and subsequently duplicated in *C. elegans*. In *C. brenneri*, one of the two sb-77a copies was deleted. Thus, the cluster we find in recent *C. brenneri* consist of four members of sbRNA family sb75 and three copies of sbRNA family sb77. In *C. remanei*, however, numerous genes were lost, possibly due to extensive genomic rearrangement at this locus. The exon structure of the surrounding gene (B0361.11) was altered as well, such that in *C. remanei* the sbRNA cluster resides in intron 2 instead of intron 3. Corresponding sbRNA in *C.*

*briggsae* seem to have been lost, since the corresponding intron is just 60 nt in size. *C. japonica* has a normal sized intron of 2,000 nt as seen in other species, but no sbRNA signatures have been detected there.

THE SBRNA CLUSTER ON CHROMOSOME X. The Chr.X cluster can be found with syntenic regions in all five *Caenorhabditis* species, Fig. 6B. The number of sbRNA copies at these loci, however, varies dramatically. The cluster apparently derives from a single sbRNA, with *C. japonica* representing the ancestral state. The first duplication gave rise to two distinctive sbRNA families (A and B). Members of family A were duplicated several times such that we find slightly varying gene numbers and cluster arrangements in *C. elegans*, *C. brenneri*, and *C. remanei*. In *C. briggsae*, however, we find three local tandem duplications of the entire cluster comprising 4 sbRNAs, as observed for sb-75/sb-77. Notably, the promoter of the 3' most gene of the multiplied unit was lost. Since the sequences of this copy termed C still show both characteristic secondary structure and sequence motifs common to sbRNAs, these genes might still be functional.

The entire region on chromosome X has undergone frequent major rearrangements. As a result, the order of genes within the cluster has changed several times. The same holds for the neighbouring protein coding genes. Since we tried to identify sbRNA clusters based on homology of features of the genomic locus, as intronic position or the order of protein coding genes in the immediate vicinity of the sbRNAs in question, we might as well have lost track of several sbRNAs. In the case of the Chr.X cluster, the order of the protein coding genes was altered and new genes appeared at the locus. Thus, some of the sbRNAs found here might not be innovations, but have been relocated by large scale genomic rearrangements.

TWO SBRNA CLUSTERS ON CHROMOSOME V. The clusters on *C. elegans* chromosome V, Fig. 7B, are distinct from all other sbRNAs discussed so far because their loop regions are both much longer and heavily structured. The clusters belong to two distinct sbRNA subfamilies with different length. Members of the shorter ones, open symbols in Fig. 7B, were found in *H. contortus*, *C. japonica*, *C. elegans*, *C. brenneri*, *C. remanei*, and *C. briggsae*. The longer paralogs, indicated by filled symbols in Fig. 7B, appear in *C. elegans*. Both families represented here seem to be ancestral to or at least as old as the family comprising the majority of sbRNAs. Further support is given by the presence of at least one family in *H. contortus*. As in the Chr.III and Chr.X clusters, there are duplications and deletions of individual genes.

Taking a closer look at the loop regions showed that the substructures of the loop-region H also evolved by regional duplications and deletions of substructures (see Supplemental Figure S1). Sequence/structure alignments revealed that three stems in the loop region of the larger sbRNAs are conserved, as shown in the consensus structure, see Fig S1-A and D. Hairpin 2 shows high similarity to the adjacent stem on its 5'-flank. In particular, the loop motifs are almost identical, suggesting that they have arisen in the

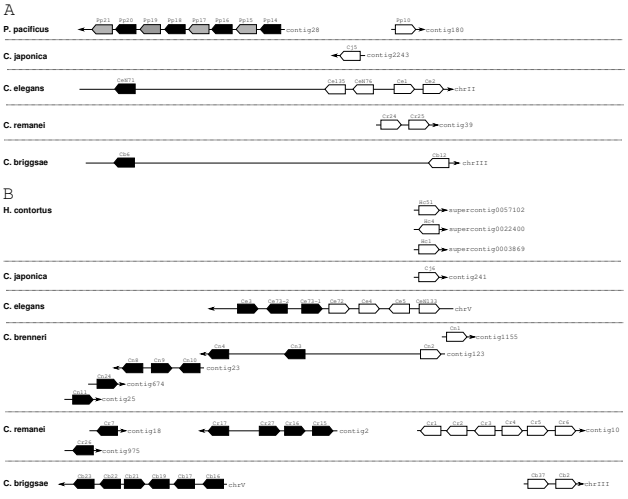



FIG. 7.—**A** Genomic organization of the *C. elegans* sbRNA cluster on chromosome II and its homologs. The cluster consist of two sbRNA families, both with very short loop motifs ($\leqslant$ 20 nt). **B** Genomic organization of the *C. elegans* sbRNA cluster on chromosome V and its homologs. The clusters consist of two sbRNA families of different loop sizes (white boxes mark shorter ones, black the longer ones). The shorter ones date back to *H.c.*, whereas the longer ones appear in *c. elegans*. Besides the structure and sequence motifs common to all sbRNAs, both families reveal no homology in the heavily structured loops and therefore do not seem to have arisen by gene duplication. Positions represent organisation and phylogenetic relations of sbRNAs and do not reflect physical distances on chromosomes and contigs.

ancestral sbRNA by the duplication of an already existing secondary structure element.

Our results suggest that at least loop regions of these sbRNAs contain functional motifs, possibly establishing interactions with binding partners such as proteins or RNAs. Especially the high conservation of motifs in hairpin 1 and 2 (CTTG) is striking. Most sbRNAs here have at least one stem in the loop region of this type. Hairpins 3 and 4 in contrast seem to be more flexible and probably carry out gene specific functions.

THE SBRNA CLUSTER ON CHROMOSOME II. The cluster on *C. elegans* chromosome II, Fig. 7A, consists of very short sbRNAs. The loop motif does not exceed 20 nt in length and seems to unstructured. Due to these short loop motifs the evolutionary history of this sbRNA cluster could not be resolved unambiguously. Nevertheless, there is evidence that the cluster comprises at least two distinct subfamilies. Units consisting of two sbRNAs, black and gray symbols in Fig. 7A, were duplicated resulting in a total of eight sbRNA copies on *P. pacificus* contig28. Both, *C. elegans* and *C. briggsae* retained one copy of the that family. The other family can be found with varying copy numbers in all *Caenorhabditis* species and in *P. pacificus*.

## 4. Discussion

In this study we identified sbRNA homologs in species of nematode clades IV and V by a combination of several search strategies. While sequence only based homology search failed to retrieve homologs in distantly

related species, the computational promoter screen and the model based search were successfull in a broader range of species. Finding RNA homologs by their characteristic promoter elements is a promising strategy, however it requires prior knowledge of promoters and regulatory elements. Creating an appropriate data set to deduce a search pattern in turn often requires RNA homology search. Secondly, the types of promoters, namely pol-III type 3, used for this kind of studies (Gruber et al. 2008; Pagano et al. 2007) so far are limited to a small number of RNA families. In a recent contribution some of us reported on a similar approach using promoter elements to identify 7SK snRNA homologs in arthropods (Gruber et al. 2008). In that case the low number of hits allowed manual comparison. Here, the large number of initial candidates could be mastered only by computational methods such as sequence/structure-based clustering (Will et al. 2007). This approach is computationally expensive, but has the benefit that one is not limited to structure or sequence constraints that have been known from the beginning. The deviant promoter structure described by Deng et al. (2006) is restricted to the genus *Caenorhabditis*. In contrast, the sbRNAs of other nematodes conform very well to the canonical type-3 pol-III structure.

As a third strategy we applied model-based RNA homology search combining sequence and structure information gathered in two previous steps. Instead of focusing on specificity, we opted for a non-stringent `rnabob` model and used a PWM-based approach for subsequent filtering. In total we end up with 233 loci across Chromadorea that we identified as sbRNAs with very high confidence.

Deng et al. (2006) annotated sbRNAs as a novel RNA family, because of their unique promoter structure and no obvious sequence homology to other known RNA families. Our structural and phylogenetic analysis revealed, to our own surprise, that sbRNAs are homologs of Y RNAs. Although sbRNAs are a large and fairly diverse family of ncRNAs, only a single representative, to most derived CeY RNA (encoded by the *yrn-1* gene) was experimentally found to be bound to the *C. elegans* Ro60 ortholog ROP-1 in vivo (Van Horn et al. 1995). The same study also reported that human Y RNAs are not bound by the ceROP-1 protein *in vitro*, whereas the CeY RNA is bound by human Ro60 even more efficiently than the human Y3 and Y4 RNAs. Van Horn et al. (1995) also noted that the human Ro60 protein significantly differs from its *C. elegans* ortholog. In particular, there are a 6- and a 19-amino acid insert in the ROP-1 RNA recognition motif. Stein et al. (2005) solved the structure of *Xenopus laevis* Ro60 and determined its RNA binding residues. Interestingly, of the 28 residues that were shown to contact Y RNA only 11 are conserved in from and worm, while 27 are shared between human and frog. On the 14 residues in contact with misfolded RNAs, most (11) are conserved between from and worm, only two less than the 12 shared by human and frog.

We found here that nematode sbRNAs are more similar to human Y RNAs than to the published *C. elegans* ROP-1 binding Y RNA, in particular in terms of their secondary structure. As sbRNA resemble human Y RNAs, and human Y RNAs are not bound by ROP-1, we suggest that nematode sbRNAs likely are not incorporated in RoRNPs despite their homology with the *C. elegans* RoRNP component CeY. There is still the possibility that sbRNAs actually are CeROP-1 binding partners, however under conditions or in developmental stages different from those analyzed in the study of Van Horn et al. (1995), who exclusively used *C. elegans* embryos for co-immunoprecipitation of ROP-1 bound Y RNAs. In this respect an ill-defined role of *rop-1* in *C. elegans dauer larvae* formation turns out to be quite interesting (Labbé et al. 2000), as it allows speculations about alternative binding partners of ROP-1 during or after the process of *dauer* formation. Additionally, Labbé et al. (2000) showed proteolytic processing of the ROP-1 protein during L2/L3 larval transition. As Van Horn et al. (1995) only analyzed ROP-1 binding RNAs in *C. elegans* embryos, it might be speculated that RNA binding affinities of the RO60 ortholog might be changed after proteolytic cleavage.

Therefore, further research will be necessary to understand whether sbRNAs are actual Ro60 binding partners *in vivo* and accordingly can be identified as Y RNAs. The ROP-1 binding partner might vary in different larval stages and especially during and after the process of *dauer larvae* formation or under stress conditions respectively, as sbRNAs were shown to be over expressed after heat shock (Deng et al. 2006). If sbRNAs cannot be established as Ro60 binding partners *in vivo*, from an evolutionary perspective it still might be interesting to know if nematode sbRNAs are bound by human Ro60 protein.

As sbRNAs conserve a motif that was recently shown by Gardiner et al. (2009) to be essential for the function of vertebrate Y RNAs in DNA replication, it is very tempting to speculate about an involvement of sbRNAs in nematode chromosomal DNA replication. Our unpublished data of a *C. elegans yrn-1* deletion indicate that the ceY RNA — in contrast to human Y RNAs — is not essential for chromosomal DNA replication. As RNAi depletion of some sbRNAs do not show any phenotype (Kamath et al. 2003; Sönnichsen et al. 2005), it is plausible to speculate that either not all sbRNAs might be involved in a hypothetical function in nematode DNA replication or, alternatively, that different sbRNAs might substitute for each other similar to vertebrate Y RNAs (Gardiner et al. 2009; Christov et al. 2006). If this is the case, research by reverse genetics will not be easy given that the sbRNA family comprises at least 18 paralogs in *C. elegans*.

## Supplemental Information

An Electronic Supplement located at `http:www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-???/` compiles sequence data, alignments in machine-readable form, and `RNABOB` search patterns.

## Acknowledgments

This research originated from an RNA Bioinformatics course at the University of Vienna in the fall semester 2008. Subsequently, it was then funded in part by the Austrian GEN-AU projects “bioinformatics integration network III” and “noncoding RNA”, and the DFG under the auspices of the SPPs 1174 “Deep Metazoan Phylogeny” and 1258

"Sensory and Regulatory RNAs".

## Literature Cited


Abad, P., J. Gouzy, J. M. Aury, et al. 2008. Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. Nat Biotechnol **26**:909–915.

Aftab, M. N., H. He, G. Skogerbo, and R. Chen. 2008. Microarray analysis of ncRNA expression patterns in Caenorhabditis elegans after RNAi against snoRNA associated proteins. BMC Genomics **9**:278–278.

Chen, X., E. J. Wurtmann, J. Van Batavia, B. Zybailov, M. P. Washburn, and S. L. Wolin. 2007. An ortholog of the Ro autoantigen functions in 23S rRNA maturation in D. radiodurans. Genes Dev **21**:1328–1339.

Christov, C. P., T. J. Gardiner, D. Szüts, and T. Krude. 2006. Functional requirement of noncoding Y RNAs for human chromosomal DNA replication. Mol Cell Biol **26**:6993–7004.

Christov, C. P., E. Trivier, and T. Krude. 2008. Noncoding human Y RNAs are overexpressed in tumours and required for cell proliferation. Br J Cancer **98**:981–988.

Crooks, G. E., G. Hon, J. M. Chandonia, and S. E. Brenner. 2004. WebLogo: a sequence logo generator. Genome Res **14**:1188–1190.

Dávila López, M., M. A. Rosenblad, and T. Samuelsson. 2008. Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components. Nucleic Acids Res **36**:3001–3010.

Deng, W., X. Zhu, G. Skogerbo, et al. 2006. Organization of the Caenorhabditis elegans small non-coding transcriptome: genomic features, biogenesis, and expression. Genome Res **16**:20–29.

Gardiner, T. J., C. P. Christov, A. R. Langley, and T. Krude. 2009. An evolutionarily conserved motif of vertebrate Y RNAs is essential and sufficient for chromosomal DNA replication in human cell extracts. RNA .

Green, C. D., K. S. Long, H. Shi, and S. L. Wolin. 1998. Binding of the 60-kDa Ro autoantigen to Y RNAs: evidence for recognition in the major groove of a conserved helix. RNA **4**:750–765.

Griffiths-Jones, S. 2005. RALEE–RNA ALignment editor in Emacs. Bioinformatics **21**:257–259.

Gruber, A. R., C. Kilgus, A. Mosig, I. L. Hofacker, W. Hennig, and P. F. Stadler. 2008. Arthropod 7SK RNA. Mol Biol Evol **25**:1923–1930.

Guffanti, E., R. Corradini, S. Ottonello, and G. Dieci. 2004. Functional Dissection of RNA Polymerase III Termination Using a Peptide Nucleic Acid as a Transcriptional Roadblock. J. Biol. Chem. **279**:20708–20716.

Gunnery, S., Y. Ma, and M. B. Mathews. 1999. Termination sequence requirements vary among genes transcribed by RNA polymerase III. J Mol Biol. **286**:745–757.

Hernandez, N. 2001. Small Nuclear RNA Genes: a Model System to Study Fundamental Mechanisms of Transcription. J. Biol. Chem. **276**:26733–26736.

Hertel, J., D. de Jong, M. Marz, D. Rose, H. Tafer, A. Tanzer, B. Schierwater, and P. F. Stadler. 2009. Non-Coding RNA Annotation of the Genome of *Trichoplax adhaerens*. Nucleic Acids Res. **37**:1602–1615.

Hertel, J., I. L. Hofacker, and P. F. Stadler. 2008. SnoReport: computational identification of snoRNAs with unknown targets. Bioinformatics **24**:158–164.

Hofacker, I. L., M. Fekete, and P. F. Stadler. 2002. Secondary Structure Prediction for Aligned RNA Sequences. J. Mol. Biol. **319**:1059–1066.

Kaczkowski, B., E. Torarinsson, K. Reiche, J. H. Havgaard, P. F. Stadler, and J. Gorodkin. 2009. Structural profiles of miRNA families from pairwise clustering. Bioinformatics **25**:291–294.

Kamath, R. S., A. G. Fraser, Y. Dong, et al. 2003. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature **421**:231–237.

Kel, A. E., E. Gössling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender. 2003. MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res **31**:3576–3579.

Labbé, J. C., J. Burgess, L. A. Rokeach, and S. Hekimi. 2000. ROP-1, an RNA quality-control pathway component, affects Caenorhabditis elegans dauer formation. Proc Natl Acad Sci U S A **97**:13233–13238.

Li, T., H. He, Y. Wang, H. Zheng, G. Skogerbo, and R. Chen. 2008. In vivo analysis of Caenorhabditis elegans noncoding RNA promoter motifs. BMC Mol Biol **9**:71–71.

Marz, M., T. Kirsten, and P. F. Stadler. 2008. Evolution of Spliceosomal snRNA Genes in Metazoan Animals. J. Mol. Evol. **67**:594–607.

Missal, K., X. Zhu, D. Rose, W. Deng, G. Skogerbo, R. Chen, and P. F. Stadler. 2006. Prediction of structured non-coding RNAs in the genomes of the nematodes Caenorhabditis elegans and Caenorhabditis briggsae. J Exp Zoolog B Mol Dev Evol **306**:379–392.

Mosig, A., J. L. Chen, and P. F. Stadler. 2007a. Homology Search with Fragmented Nucleic Acid Sequence Patterns. In: Giancarlo, R., and S. Hannenhalli, editors, Algorithms in Bioinformatics (WABI 2007), volume 4645 of *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Verlag, 335–345.

Mosig, A., M. Guofeng, B. M. Stadler, and P. F. Stadler. 2007b. Evolution of the vertebrate Y RNA cluster. Theory Biosci **126**:9–14.

Pagano, A., M. Castelnuovo, F. Tortelli, R. Ferrari, G. Dieci, and R. Cancedda. 2007. New small nuclear RNA gene-like transcriptional units as sources of regulatory transcripts. PLoS Genet **3**.

Perreault, J., J. P. Perreault, and G. Boire. 2007. Ro-associated Y RNAs in metazoans: evolution and diversification. Mol Biol Evol **24**:1678–1689.

Sönnichsen, B., L. B. Koski, A. Walsh, et al. 2005. Full-genome RNAi profiling of early embryogenesis in Caenorhabditis elegans. Nature **434**:462–469.

Stein, A. J., G. Fuchs, C. Fu, S. L. Wolin, and K. M. Reinisch. 2005. Structural insights into RNA quality control: the Ro autoantigen binds misfolded RNAs via its central cavity. Cell **121**:529–539.

Tanzer, A., and P. F. Stadler. 2004. Molecular Evolution of a MicroRNA Cluster. J. Mol. Biol. **339**:327–335.

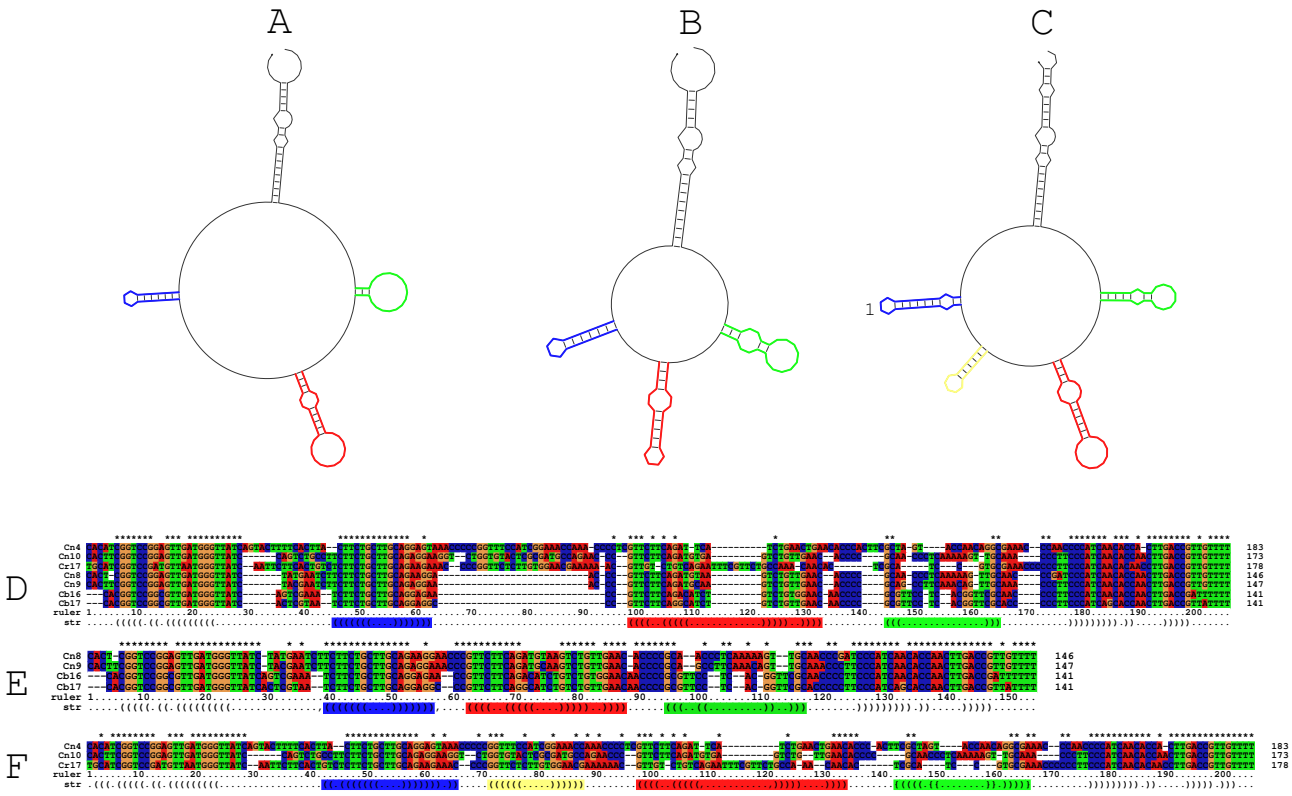
Thomas, J., K. Lea, E. Zucker-Aprison, and T. Blumenthal. 1990. The spliceosomal snRNAs of Caenorhabditis elegans. Nucleic Acids Res **18**:2633–2642.

Van Horn, D. J., D. Eisenberg, C. A. O'Brien, and S. L. Wolin. 1995. Caenorhabditis elegans embryos contain only one major species of Ro RNP. RNA **1**:293–303.

Waterhouse, A. M., J. B. Procter, D. M. Martin, M. Clamp, and G. J. Barton. 2009. Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. Bioinformatics .

Will, S., K. Reiche, I. L. Hofacker, P. F. Stadler, and R. Backofen. 2007. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comput Biol **3**.

Zemann, A., A. op de Bekke, M. Kiefmann, J. Brosius, and J. Schmitz. 2006. Evolution of small nucleolar RNAs in nematodes. Nucleic Acids Res **34**:2676–2685.




**Supplemental Figure.**
Structure evolution of sbRNA loop regions. Gene duplication coincides with duplication of substructures within the loops regions as shown here for members of the long sbRNAs residing on *C. elegans* chromosome V. The second hairpin (yellow, C and F) is only present in Cn4, Cn10 and Cr17 and shows high sequence similarity to the adjacent hairpin (blue). A,B,C: consensus structures calculated with `RNAalifold` based on hand curated `clustalw` multiple sequence alignments (D,E,F).