J. Mol. Evol. manuscript No.
(will be inserted by the editor)

Ilenia Boria*, **Andreas R. Gruber***, **Andrea Tanzer***, **Stephan H. Bernhart**, **Ronny Lorenz**, **Michael M. Mueller**, **Ivo L. Hofacker**, **Peter F. Stadler**

Nematode sbRNAs: homologs of vertebrate Y RNAs

March 10, 2010

Abstract Stem-bulge RNAs (sbRNAs) are a group of small, functionally yet uncharacterized noncoding RNAs first described in *C. elegans*, with a few homologous sequences postulated in *C. briggsae*. In this study we report on a comprehensive survey of this ncRNA family in the phylum Nematoda. Employing homology search strategies based on both sequence and secondary structure models and a computational promoter screen we identified a total of 240 new sbRNA homologs. For the majority of these loci we identified both promoter regions and transcription termination signals characteristic for pol-III transcripts. 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 contain 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.

Keywords
sbRNA, nematodes, Y RNA, homology search, noncoding RNA

I. Boria
Department of Medical Sciences and Interdisciplinary Research Centre for Autoimmune Diseases, Università del Piemonte Orientale, via Solaroli 17, I-28100 Novara, Italy.

I. Boria, A.R. Gruber, A. Tanzer, S.H. Bernhart, R. Lorenz, I.L. Hofacker, P.F. Stadler
Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria.
E-mail: {ilenia,agruber,at,berni,ronny,ivo}@tbi.univie.ac.at

A.R. Gruber, A. Tanzer, P.F. Stadler
Bioinformatics Group, Department of Computer Science; and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany

M.M. Mueller
Department of Chromosome Biology, Max F. Perutz Laboratories, University of Vienna, A-1030 Vienna, Austria.

P.F. Stadler
Max-Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany.
Fraunhofer Institut für Zelltherapie und Immunologie (IZI), Perlickstraße 1, D-04103 Leipzig, Germany.
Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA.
E-mail: stadler@bioinf.uni-leipzig.de
*These authors contributed equally.

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 the family. In a subsequent contribution, Aftab et al. (2008) annotated three additional experimentally verified ncRNAs as sbRNAs. These seed sequences are listed in Tab. 1. They share 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. The term "stem-bulge RNA" was coined because of this feature (Deng et al. 2006). A BLAST-based comparison with the *C. briggsae* genome revealed eleven putative homologs (Deng et al. 2006), providing further support for the stem-structure and indicating that the loop regions evolve rapidly.

The sbRNAs in *C. elegans* as well as their *C. briggsae* homologs show a common 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

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 by Perreault et al. (2007) is marked by **f**. Column L denotes the length.

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	C15H11.12	AM286259	95	**b**,**d**, **e**
CeN134	F35E12.11	AM286260	119	**b**,**f**

50% – was detectable 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 are transcribed by RNA polymerase III.

Most sbRNAs are differentially expressed in developmental stages. The highest levels of expression have been found in mature adult worms, dauer larvae and especially worms after heat shock (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, knock-down experiments were performed (Kamath et al. 2003). No phenotype was reported for these two knock-downs. CeN133 was also knocked down in a study by Sönnichsen et al. (2005), again with no visible phenotype. Considering latest results on the efficiency of RNAi on ncRNAs (Ploner et al. 2009) it has to be questioned if sbRNA expression levels were sufficiently decreased to see an effect. Furthermore, functionally required motifs may reside in the highly structurally conserved stem common to all sbRNAs. It is plausible, therefore, that other sbRNAs may functionally compensate for the reduced levels of a particular paralog.

A first attempt to gain insight into the putative biological functions of 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 functions and processes the sbRNAs are involved in remain to be uncovered.

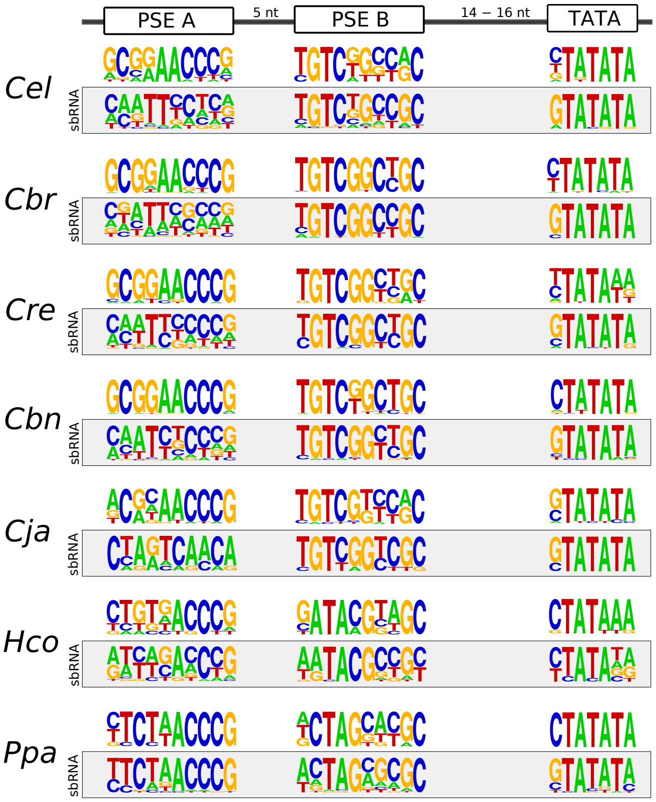


Fig. 1 Comparison of promoter elements of sbRNAs to other pol-III transcripts. The upper row for each species shows sequence logos (Crooks et al. 2004) of the promoter motifs for other pol-III transcripts (U6 snRNA, RNase P, RNase MRP, tRNA-SeC, Y RNAs), while the lower row denotes the corresponding elements for sbRNAs. High similarity is observed for the PSE B and the TATA-box for all species, while high similarity for PSE A is only observed for *H. contortus* and *P. pacificus*. Similarity was measured using the averaged Kullback-Leibler divergence of position frequency matrices of the corresponding motifs, see e.g. Aerts et al. (2003). A value of 0.20 and below can be considered to indicate high similarity. Abbreviations: Cel - *C. elegans*, Cbr - *C. briggsae*, Cre - *C. remanei*, Cbn - *C. brenneri*, Cja - *C. japonica*, Hco - *H. contortus*, Ppa - *P. pacificus*.

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. We show that, unexpectedly, sbRNAs are homologs of Y RNAs.

Materials and Methods

Sequence Data

Genomic sequences of nematode species 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), the *M. hapla* Genome Sequencing Group (www.hapla.org). Details on the assemblies used here are listed in the Electronic Supplement. The phylogenetic relations of the investigated species are depicted in Fig. 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. Due to the high sequence variation in the central loop region, this initial step recovered only a few full length sbRNAs in other species. Blastn hits that showed a query coverage of at least 50% were extended by flanking sequence and manually compared to known sbRNAs in a structural alignment. 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.

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 Selenocysteine tRNAs. These noncoding RNAs are known to utilize very similar PSE B and TATA-Box promoter elements. For *C. elegans* the sequences for RNase P, RNase MRP, and Selenocysteine tRNA could easily be retrieved from annotated Wormbase entries (rpr-1, mrpr-1, K11H12.t1) or, in case of U6, snRNAs from the literature (Dávila López et al. 2008; Marz et al. 2008). Simple blastn searches were sufficient to identify their orthologs in other nematode 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. 2007b) search pattern. The FRAGREP search resulted in approx. 1,200 hits in *C. remanei* and more moderate numbers for the other nematodes. For each hit we searched the 300 nt 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 putative terminator was extracted for further analysis.

We then applied sequence-structure based clustering using the LocARNA-RNAclust pipeline (Will et al. 2007; Kaczkowski et al. 2009) to these putative transcripts. Default parameters were used for both LocARNA and RNAclust. Clusters were visually examined for sequence-structure similarity to already identified sbRNAs using the RNAsoupViewer (www.bioinf.uni-leipzig.de/pages/40/software.html).

This approach offers two major advantages over purely sequence-based or (structure) model-based searches, where only the ncRNA itself is used as query: (i) since promoter elements that are shared with other ncRNA classes are used for initial filtering of the genomic data, knowledge on the variability of the sequence and/or structure of the query ncRNA is irrelevant at this stage. Instead, a search using the query ncRNA is only performed on the small set of putative transcripts. Thus, more sensitive but also computationally much more expensive tools can be used in this second step; (ii) the canonical promoter structure lends additional credibility to the candidates. The feasibility of this strategy was recently demonstrated for identifying 7SK snRNAs of arthropods (Gruber et al. 2008).

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. In a first analysis, sbRNAs were manually grouped into clusters based on length and sequence identity and aligned. RNAalifold (Hofacker et al. 2002; Bernhart et al. 2008) predictions for each group were then used as starting point for deriving a consensus structure for the well-conserved parts. These initial alignments were then refined manually and combined to a global alignment in the emacs text editor, making use of the RALEE mode (Griffiths-Jones 2005), which explicitly handles secondary structure annotation.

These structure-annotated alignments were then 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) with default parameters. The resulting initial candidates were filtered using a modified position weight matrix scoring in which base-pairs are treated like individual letters:

Let $\mathcal{A} = \{A, C, G, T\}$ be the nucleotide alphabet. Then $\mathcal{B} = \{AA, AC, AG, AT, ..., TT\}$ 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

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

where i is now either an unpaired nucleotide or a base pair, and $k(b) = 4$ if $b \in \mathcal{A}$ and $k(b) = 16$ if $b \in \mathcal{B}$. We implemented a Perl script 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 > 0.65 were then compared manually to previously identified sbRNAs. Recognizable homologs were incorporated into the sequence-structure alignment.

Identification of Promoter Elements

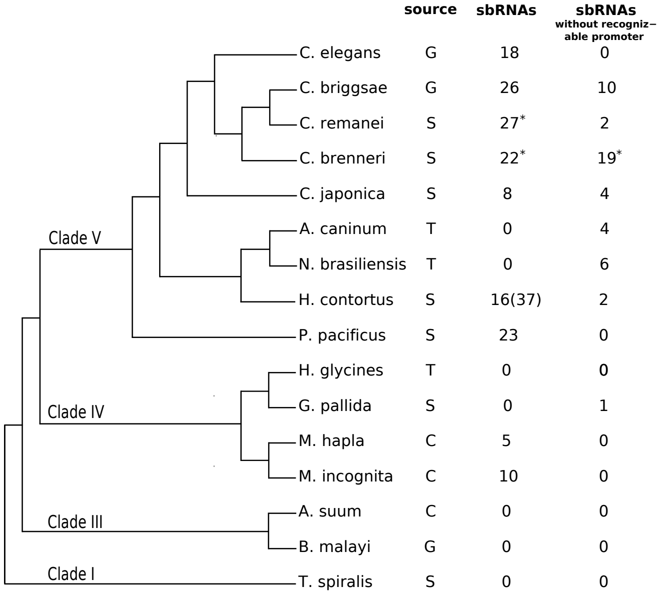
For the five species of the genus *Caenorhabditis*, *P. pacificus*, and *H. contortus* we were able to collect a sufficient number of upstream regions of ncRNAs that share at least partially the same promoter elements as sbRNAs. We created separate position weight matrices (PWMs) for the PSE A and the PSE B for each species as well as a general TATA-box PWM and used the approach by Kel et al. (2003) to score corresponding elements in the upstream sequences of our sbRNA candidates. 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.

Identification of Syntenic Regions

The UCSC Genome Browser provides gene annotations for all *Caenorhabditis* genomes used in this study. The advantage of this resource is that *C. elegans* genes were mapped using `tblastn` to other *Caenorhabditis* proteins so that the gene identifiers are available across all genomes. Wormbase, on the other hand, uses different gene identifiers for the individual species and does not supply a read-to-use homology table. In order to construct local synteny maps between *Caenorhabditis* genomes, we first used a simple `blastn` search to map our sbRNA sequence to the genomes version provided by the sequence repository at UCSC, which are older than the genome assemblies for the other analyses used here. We then extracted gene annotations within ±40 kb of each sbRNA location. In the next step, sbRNAs and adjacent genes were compared between all genomes. If sbRNAs in different genomes are located in the vicinity of genes with identical annotation, we consider these locations syntenic.

All genomes used here, except for *C. elegans* and *C. briggsae*, have not been assembled to the level of chromosomes. Thus, sbRNAs and adjacent protein coding genes might resided on different contigs making it difficult to identify both upstream and downstream markers. As a consequence, our strategy for detecting sbRNAs in syntenic regions requires at least one homologous protein within ±40 kb flanking a sbRNA.

Using this approach, we found that only two *C. elegans* sbRNA clusters, namely those on chromosome III and chromosome X have syntenically conserved locations in other *Caenorhabditis* species. These two clusters where then in detail examined using the synteny resources available at wormbase.org.



	source	sbRNAs	sbRNAs without recognizable promoter
C. elegans	G	18	0
C. briggsae	G	26	10
C. remanei	S	27*	2
C. brenneri	S	22*	19*
C. japonica	S	8	4
A. caninum	T	0	4
N. brasiliensis	T	0	6
H. contortus	S	16(37)	2
P. pacificus	S	23	0
H. glycines	T	0	0
G. pallida	S	0	1
M. hapla	C	5	0
M. incognita	C	10	0
A. suum	C	0	0
B. malayi	G	0	0
T. spiralis	S	0	0

Fig. 2 Phylogenetic distribution of the 240 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 species phylogeny is represented as a cladogram with arbitrary branch lengths, combining the Caenorhabditis species phylogeny by Sudhaus and Kiontke (2007) with the phylogeny of phylum Nematoda by Blaxter et al. (1998) and Mitreva et al. (2005) and the work from Chilton et al. (2006). *Accounting for allelic variants (Barrière et al. 2009), the number of sbRNAs in *C. remanei* is reduced to 26, while in *C. brenneri* 19 copies with intact promoter and 15 without are genomically distinct. The column "source" denotes the assembly status of the genomic DNA 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.

Results

Homology Searches

Starting from the seed sequences, both the analysis of the multiz alignments and an iterative BLAST search resulted only in a moderate number of additional homologs in the *Caenorhabditis* species and a few hits in *P. pacificus*, and 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. sbRNAs found to that point were used to construct a promiscuous search pattern for RNABOB, whose results were 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 240 sbRNAs distributed over the nematode clade V (Strongylida, Diplogasterida, and Rhabditida) and clade IV (Tylenchida, Cephalobina, and Panagrolaimida), summarized in Fig. 2. It was recently shown that a considerable fraction of the genome assemblies of *C. remanei* and *C. brenneri* represents two alleles rather than distinct genomic loci (Barrière et al. 2009). In *C. brenneri* 14 sbRNAs that assembled to separate contigs show extensive sequence similarities (> 80% identity) within 1,000 nt examined flanking regions. Six out of these 14 show nearly perfect sequence conservation in the 3' flanking region, while the 5' flanking region does not. For these cases it is likely that we see an assembly artifact instead of an allelic variant. In *C. remanei* we find two sbRNAs that are located on separate contigs and show extensive sequence identity in the flanking regions. High identity is, however, only observed in the 5' flanking region suggesting that it might again be an assembly artifact. We conclude that 8 of our 240 sbRNA sequences are duplicates.

In particular, we report a total of 18 sbRNAs genes in the *C. elegans* genome, all having confirmed promoter elements. In the other species we also list a significant number of sbRNAs that do not show significant matches to known ncRNA promoter elements. One of the hits we identified in *H. contortus* has 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 sbRNAs 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*.

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 previously described 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). In other species of the phylum nematoda, studies of snRNA promoters of this type (pol-III type 3) have not been conducted so far. For all species except *C. elegans*, 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 *A. caninum*, *N. brasiliensis*, *G. pallida*, *M. hapla*, and *M. incognita* we were not able to find a sufficient number of high-confidence homologs of other pol-III transcripts to build reliable species-specific position weight matrices (PWMs) or to determine the exact position of PSEs and the TATA-box. In these cases upstream regions were visually compared for stretches of homologous regions. Results of promoter analysis are summarized in Fig. 2.

Secondary Structure

In order to derive a consensus secondary structure, we used the subset of those 155 (out of 240) 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.

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

S1 Stem S1 consists of at least four conserved base-pairs. It is extended at the outer end in most of the sequences. The closing inner AU pair of stem S1 is absolutely conserved in all sequences.

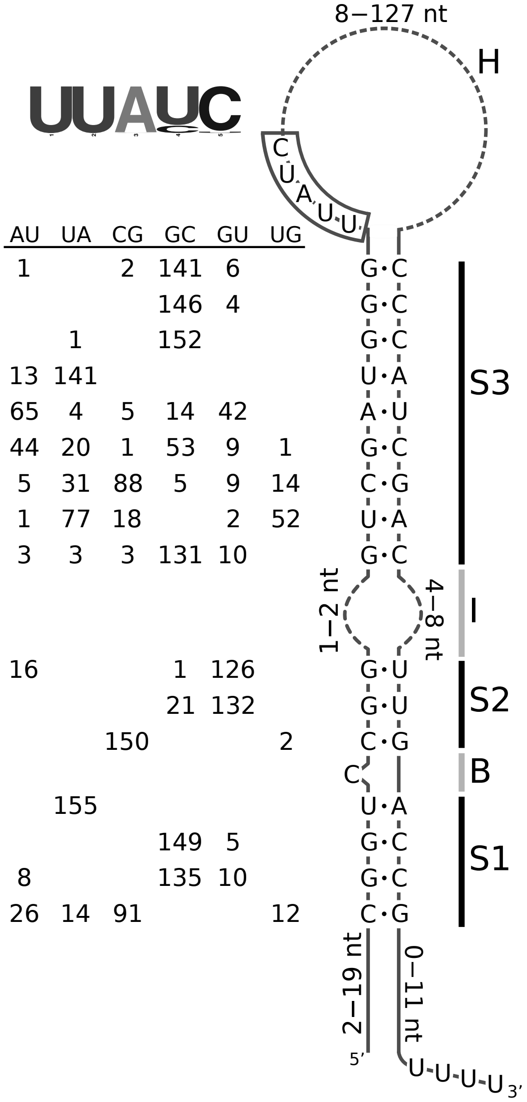
S2 Stem S2 is composed of three base-pairs only, and the majority of sequences shows two GU wobble-pairs. From a thermodynamic point of view this is a rather weak stem, but supporting evidence is given by compensatory mutations.

S3 Stem S3 is composed of nine base-pairs. The outer part of S3 shows many compensatory mutations, suggesting that the ability to form this double stranded 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 separated 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, it does not seem to be a general sequence motif for the entire set of sbRNAs there.

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 generally unstructured, forming a single hairpin loop. The longer sbRNA homologs tend to form short structural elements that appear conserved within subgroups.



AU	UA	CG	GC	GU	UG
1		2	141	6	
			146	4	
	1		152		
13	141				
65	4	5	14	42	
44	20	1	53	9	1
5	31	88	5	9	14
1	77	18		2	52
3	3	3	131	10	
16			1	126	
			21	132	
		150			2
	155				
			149	5	
8			135	10	
26	14	91			12

Fig. 3 Secondary structure model of sbRNAs derived from 155 sbRNAs with verified promoter regions. The table on the left shows the absolute counts of canonical and wobble base-pairs observed at a given position. The schematic drawing of the structure displays the most frequent base-pair. The sequence logo shows the frequencies of nucleotides for the UUAUC motif, which immediately follows the conserved stem. In only two out of 240 sbRNAs we observed one or two additional G residues inserted between the stem and this motif.

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, which may serve as alternative termination signals (Gunnery et al. 1999; Guffanti et al. 2004) can be observed downstream of their genomic location.

sbRNAs are Y RNAs

Comparison with other RNA families revealed that nematode sbRNAs show substantial similarities in both sequence and secondary structure to vertebrate Y RNAs (see Mosig et al. (2007a) and Perreault et al. (2007) for Y RNA structure). The sbRNA CeN134 was reported as a possible Y RNA in the kingdom-wide survey for Y RNA homologs by Perreault et al. (2007). The connection of Y RNAs and sbRNAs was not commented on, and other sbRNA family members in *C. elegans* were not recognized, however. Fig. 4 summarizes a detailed comparison of the Nematode sbRNA consensus with the analysis of vertebrate Y RNAs by Mosig et al. (2007a) and the orthologs of the previously reported *C. elegans* 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).

All three structures share not only the overall organization but also several sequence features. In particular the inner part of stem S3, the two outer pairs of stem S2, the conserved cytidine bulge B, and the inner pairs of stem S1 are the same. These regions largely coincide with the most conserved ones within each of the three groups.

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 consists 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 13 *H. contortus* sequences 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 paralogous vertebrate subfamilies Y1 and Y3.

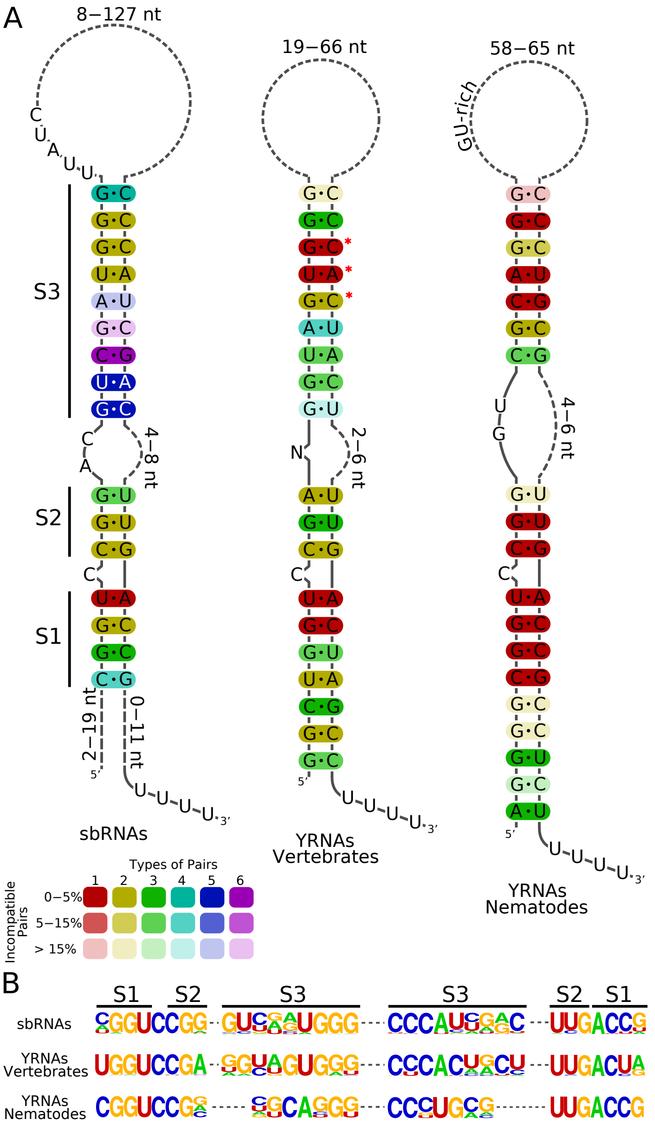


Fig. 4 **A** Comparison of secondary structures for nematode sbRNAs, vertebrate Y RNAs and the previously described Y RNA family in the genus *Caenorhabditis*. Red stars denote the region identified by Gardiner et al. (2009) to be crucial for the function of Y RNA in DNA replication. **B** Sequence logos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

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*
Cel1	intergenic	W01D2.7, Ce150*	81
Cel2	intergenic	–	85
Cel3	intronic	–	155
Cel5	intergenic	–	121
Cel6	intergenic	M163.15	83
Cel7	intergenic	M163.14, Ce94*	98

Evolutionary History of sbRNAs

In *C. elegans* we uncovered six new sbRNA homologs (Tab. 2) in addition to the twelve previously described sbRNAs. All six are supported by promoter elements. 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 Cel7 as a C/D box snoRNA. This sequence yields a negative snoRNA classification by snoReport (Hertel et al. 2008) and can be unambiguously recognized as a sbRNA homolog based on both sequence and secondary structure.

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; 2006). In brief, the significance of pairwise alignments is evaluated by comparing the score with the score distribution of of pairwise alignments of shuffled input sequences. The resulting z-scores serve as similarity measure that can be used to construct hierarchical clustering. While this approach of course does not reconstruct an accurate phylogeny, it is capable of identifying clusters with statistically significant mutual similarities (Tanzer and Stadler 2006). The clustering not only identifies sbRNAs as unambiguous homologs of Y RNAs, it also confirms the observation that nematode sbRNAs are more similar to vertebrate Y RNAs than to the previously described nematode Y RNAs.

In vertebrates, Y RNAs show features required for both 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, supporting the hypothesis of a late divergence of CeY and sbRNAs.

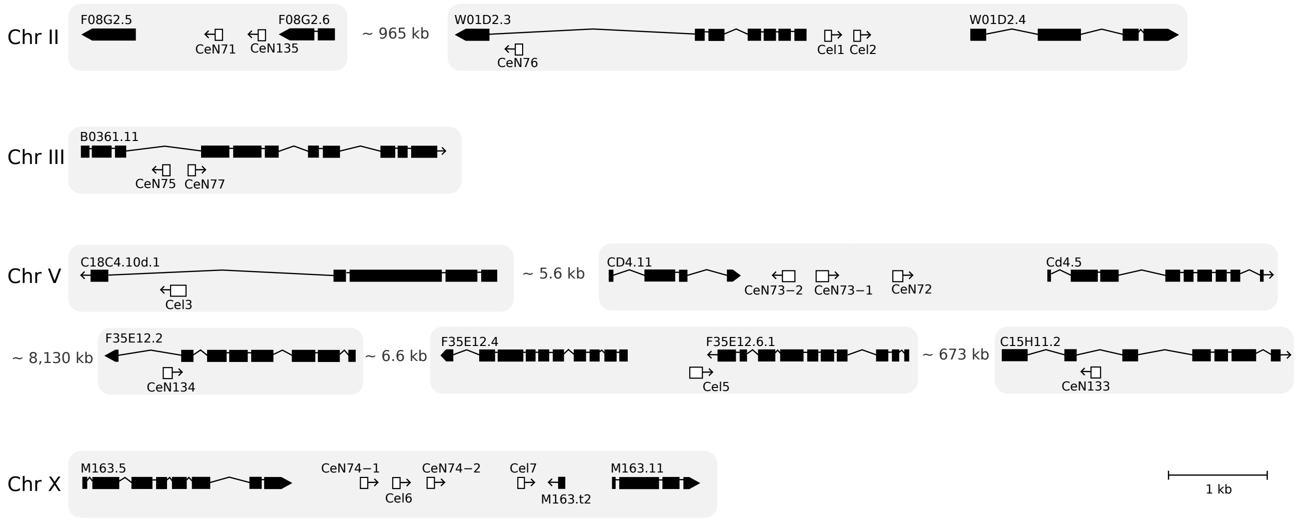


Fig. 5 Schematic drawing of the organization of the five sbRNA clusters in *C. elegans*. Each line represents a sbRNA cluster. White boxes denote sbRNAs, annotated Wormbase genes (release WS205) flanking sbRNA loci are shown in black.

The 18 *C. elegans* sbRNAs identified to-date are organized in five clusters, Fig. 5. Each cluster consists of multiple copies of one sbRNA subfamily. Thus, clusters seem to have arisen by local tandem duplications of one ancestral sbRNA. The mechanism by which sbRNAs were multiplied remains unknown. Nevertheless, we find evidence that not only single genes, but also groups of several sbRNAs might be affected by a single duplication event.

The sbRNA cluster on chromosome X. The chr. X cluster can be found with syntenic regions in all five *Caenorhabditis* species, Fig. 6 and Supplemental Fig. S2A. 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). In *C. elegans*, A was lost and B was copied 2 times. After the divergence of *C. elegans* and *C. briggsae*, B was duplicated leading to a cluster comprising three sbRNAs: A, B1 and B2, as found in *C. brenneri*. Clusters in both *C. briggsae* and *C. remanei* contain two copies of sbRNAs of family B2 suggesting a duplication prior to the speciation event. However, phylogenetic analysis rather suggest individual duplications in both species.

The sbRNA cluster on chromosome III. Both sequence similarity and cluster organisation indicate that the chr. III cluster has undergone different complex fates in each species, comprising multiple local duplication and deletion events (Fig. 6 and Supplemental Fig. S2A). Unlike the cluster on chromosome X, were single genes were effected, here two genes in tail-to-tail orientation seem to form a unit which is propagated. The two genes both contain their own PSEB and PSEA elements and thus do not seem to rely on promoter sharing.

In *C. elegans* the cluster is composed of one such unit (CeN75/CeN77) reflecting the ancestral state. Duplication of the ancestral 75/77 pair resulted in tandem copies 75A/77A and 75B/77B after the speciation event leading to *C. elegans*.

In *C. brenneri*, one of the two copies (75B/77B) was deleted and the other one (75A/77A) duplicated leading to Cbn29/Cbn30 and Cbn25/Cbn26. Thus, the cluster in *C. brenneri* consist only of members of families 75A and 77A. In addition, we find two more copies of 75A (Cbn28 and Cbn27), which most likely result from duplications of the adjacent Cbn29 (family 75A). In an alternative scenario, the whole unit of Cbn29/Cbn30 (77A/75A) was duplicated and each copy of 77B was subsequently lost. Such a scenario, however would be more costly than individual duplications and thus appears less probable. Cbn31, which is also present at this locus, shows some homology to the other members of the cluster. However, neither phylogenetic analysis nor sequence motifs in the loop regions allowed an unambiguous assignment to any of the the two families.

Members of both the 75A/77A and 75B/77B families are present in *C. remanei*. As in *C. brenneri*, we find an individual duplication of 75A (Cre12). The unit of 75B/77B was duplicated once such that in *C. remanei* there is one copy of 75A/77A (Cre10/Cre11), two copies of 75B/77B (Cre8/Cre9 and Cre14/Cre13) and another cape of 75A (Cre12). Interestingly, in *C. remanei* this locus seem to have undergone extensive genomic rearrangement. The exon structure of the surrounding gene (B0361.11) was altered, such that in *C. remanei* the sbRNA cluster resides in intron 2 instead of intron 3 (see location in *C. elegans*, Fig. 5).

In an alternative scenario, the duplication of the ancestral 75/77 pair took place after the speciation of *C.*

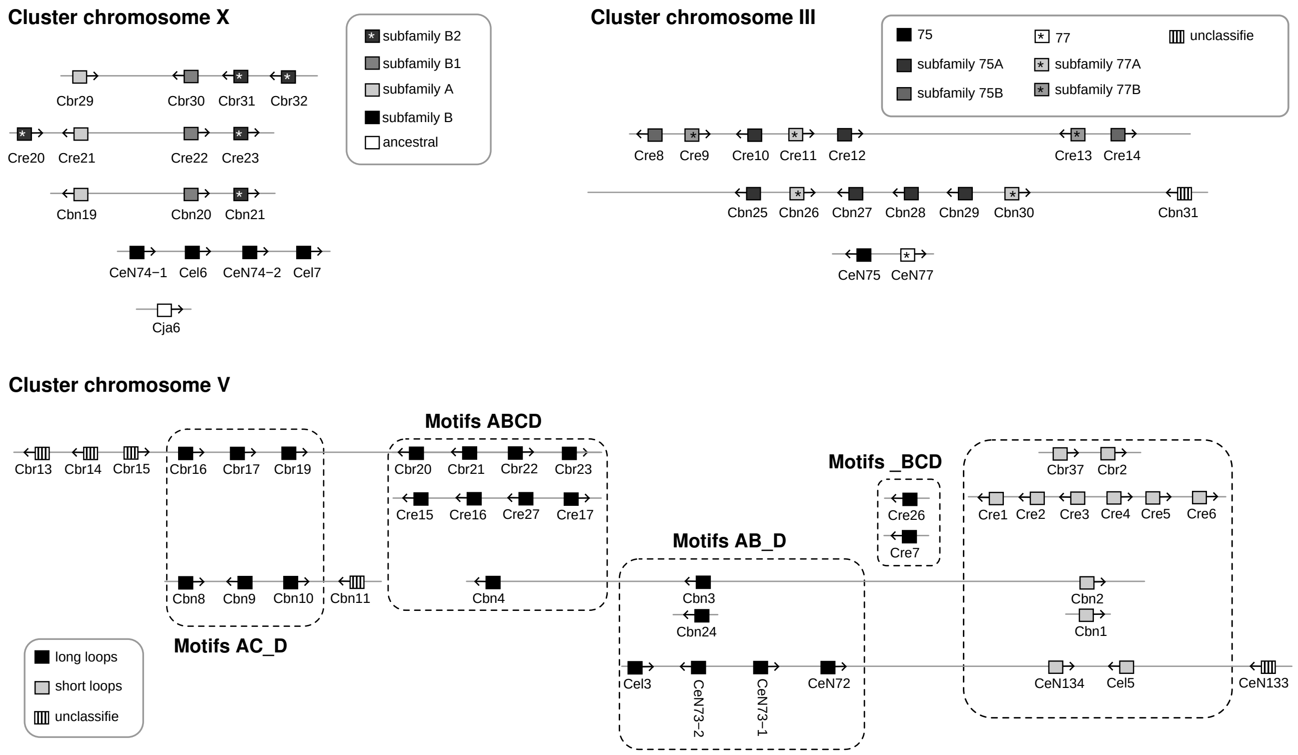


Fig. 6 Schematic drawing of the genomic organization of the sbRNA clusters on chromosome III, chromosome X, and chromosome V of *C. elegans* and their homologs. Based on phylogenetic analysis, conserved motifs and position within a cluster, individual genes were group into different subfamilies (shown as different shades of gray). Clusters are shaped by duplications of single genes as well as units of sbRNAs followed by deletions of individual genes. The cluster on chromosome V 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. contortus* (data not shown), whereas the longer ones appear in *Caenorhabditis*. Besides the structure and sequence motifs common to all sbRNAs, both families of this cluster reveal no homology in the heavily structured loops and therefore do not seem to have arisen by gene duplication. Gene duplications of the "long" sbRNAs coincided with duplications of substructures in the multiloop. A, B, C, D refers to these substructures. The loop region of CeN72 is too degenerated to assign this gene to either of the two groups based on sequence similarity. Due to its close vicinity to CeN73-1, Fig. 5, we grouped it with the long ones. For details see text and Supplemental Fig. S1 and Fig. S2. The figure shows the organisation of sbRNA clusters only and does not reflect genomic distances. Arrows indicate sbRNA orientation: plus strand (→) and minus strand (←). sbRNA which could not be assigned unambiguously to a subfamily are labelled as "unclassified". Abbreviations: Ce – *C. elegans*, Cbr – *C. briggsae*, Cre – *C. remanei*, Cbn – *C. brenneri*, Cja – *C. japonica*

brenneri. However, motifs in the loop region of all 75A family members in both *C. brenneri* and *C. remanei* are highly conservation and thus support the scenario outlined above.

Corresponding sbRNAs 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.

Two sbRNA clusters on chromosome V. The clusters on *C. elegans* chromosome V, Fig. 6, are distinct from all other sbRNAs discussed so far because their loop regions are both much longer than those of other sbRNAs and heavily structured. The clusters belong to two distinct sbRNA subfamilies of different length. Members of the shorter ones, white boxes in Fig. 6, are present in *C. japonica*, *C. elegans*, *C. brenneri*, *C. remanei*, and *C. briggsae* were also found in *H. contortus* (Electronic Supplement). The longer paralogs, indicated by filled boxes in Fig. 6, appear in *Caenorhabditis* only. Both families represented here seem to be ancestral to (or at least as old as) the family comprising the majority of sbRNAs. Further support for their evolutionary age comes from the presence of at least one of these families in *H. contortus*. As in the chr. III and chr. X clusters, there are multiple duplications and deletions of individual genes.

Taking a closer look at the loop regions of the individual genes showed that several gene duplications coincided with changes of the organization of the loop regions, i.e., regional duplications and deletions of substructures (see Supplemental Fig. S1). Thus, we grouped members of the cluster into four subfamilies based on their loop motifs (Fig. 6C). Sequence/structure alignments revealed that each of these subfamilies contains at least three stems in the loop region with hairpin A being the best conserved one. In addition, each *Caenorhabditis* species seem to have undergone individual duplications of a subfamily. Based on both phylogenetic analysis and structure information, we deduced the following evolutionary scenario:

The ancestral copy of the long sbRNAs probably consisted of three hairpins (A_CD). This first round of gene duplications results in subfamilies AB_D and ABCD, where hairpin B seems to have arisen by a duplication of the upstream hairpin A (Fig. 6D). In particular, the loop motifs are almost identical, suggesting that they have arisen in the ancestral sbRNA by the duplication of an already existing secondary structure element. In a subsequent duplication of subfamily ABCD hairpin C was deleted leading to subfamily AB_D.

In *C. elegans* A_CD was lost again, AB_D was copied 2 times and hairpin C of ABCD degraded (Ce3). *C. brenneri* still shows all three ancestral subfamilies, but again underwent individual gene duplications. After the speciation of *C. brenneri*, subfamily AB_D was lost, such that in *C. briggsae* we find only members of ABCD and A_CD. In *C. remanei*, the whole cluster was heavily remodelled. A_CD was deleted and a duplicate of ABCD lost hairpin A.

Our analysis suggests that at least loop regions of these sbRNAs contain functional motifs, possibly establishing interactions with binding partners such as proteins or RNAs. In particular, the high conservation of motifs in hairpin A and B (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 may be responsible for gene specific functions.

Reconstructing such complex patterns of gene duplications strongly depends on the genome information available. Data from additional *Caenorhabditis* as well as fully assembled genomes would be required to disentangle the apparently complex history of this cluster with any certainty. Thus, additional data and improved assemblies of the *Caenorhabditis* genomes will help to resolve the ambiguities in the scenario described above and may favour a slightly different reconstruction of the details evolutionary history in particular of these "non-syntenic" sbRNA clusters.

The sbRNA cluster on chromosome II. The cluster on *C. elegans* chromosome II consists of very short sbRNAs. The loop motif does not exceed 20 nt in length and seems to be unstructured. Due to these short loop motifs the evolutionary history of this sbRNA cluster could not be resolved unambiguously.

Discussion

Deng et al. (2006) annotated sbRNAs as a novel RNA family because of their unique promoter structure and the lack of obvious sequence homology with other known RNA families. Our analysis of the patterns of sequence and structure conservation established that sbRNAs are homologs of Y RNAs. We identified sbRNA homologs in species of nematode clades IV and V by a combination of several search strategies. While homology search based solely on sequence failed to identify many of the sbRNAs, the computational promoter screen and the searches with secondary structure models were successful in a broader range of species. We show here that a screen for characteristic promoter elements can substantially improve both sensitivity and specificity of RNA homology searches. This strategy, however, requires prior knowledge of promoter or other regulatory DNA elements. The construction of the promoter search patterns itself requires a collection of known RNA genes that are under the control of similar promoters. Due to the lack of a comprehensive ncRNA annotation for most invertebrate genomes, this amounts again to a homology search problem – although for better conserved ncRNAs. So far, promoter-based approaches have been employed systematically only for pol-III type 3 promoters (Gruber et al. 2008; Pagano et al. 2007), which are associated with a quite limited set of ncRNA families. In a recent contribution, some of us reported on the identification of the 7SK snRNA homologs in arthropods (Gruber et al. 2008) using a similar approach. In that study, the small number of initial hits allowed a manual analysis. Here, we had to use a a less stringent search because of the variability in the promoter structure itself. The deviant pol-III promoter structure of the sbRNAs described by Deng et al. (2006) turned out to be restricted to the genus *Caenorhabditis.* As a consequence, a large number initial candidates has to be a evaluated. This task 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 to been known from the beginning. As a third strategy we applied model-based RNA homology search combining sequence and structure information gathered in the 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 240 loci across the currently available genome data of Chromadorea that we identified as sbRNAs with very high confidence. Accounting for the allelic variants included in some genomes, this number reduced to 231 distinct sbRNA genes.

We have been unable to find unambiguous sbRNA/Y RNA genes in basal nematodes. This does not come as a surprise. *B. malayi*, *T. spiralis*, and *A. suum* are separated by large evolutionary distances from their closest relatives with sequenced genomes. Signals of sequence homology are therefore faint for the short sequences in question. In the case of the Chromadorea we could start from several experimentally validated sequences in *C. elegans* to retrieve a large number of homologs from closely related species. It is these data that allowed a detailed study of the sequence and structure constraints of sbRNAs. These models, in turn, were necessary to recognize the homology of sbRNAs with the previously described Y RNA of *C. elegans* and with the vertebrate Y RNAs. The information contained in these models, however, does not provide sufficient specificity to retrieve homologs from distantly related genomes with acceptable confidence. This also explains the surprising fact that the descriptor-based survey for Y RNAs by Perreault et al. (2007) hit one of the sbRNAs with borderline significance but failed to recognize most other family members.

The number of sbRNAs detected in this study varies significantly between species. For the two syntenically conserved sbRNA clusters we showed in detail that they exhibit a complex evolutionary history resulting in very different sbRNA complements even in fairly closely related species. The syntenically non-conserved clusters provide further evidence for the rapid evolution of the sbRNA complement. Strictly speaking, we cannot rule out that there are additional, highly-derived, members of the sbRNA/Y RNA family. The group of sequences identified here, however, shows coherent features and we did not detect ambiguous borderline-cases. Sampling biases, e.g. due to incomplete genome assemblies, thus might affect the exact sbRNAs counts, such technical artifacts can by no means account for the large differences between closely related species within the genus *Caenorhabditis*. Most likely, therefore, the striking differences observed in other clades, also reflects evolutionary variation rather than computational limitations.

Recent results on the function of mammalian Y RNAs suggest that they have at least two distinct modes of action. On the one hand, they are part of the Ro-RNA particle which is involved RNA quality control (Stein et al. 2005). On the other hand, they are essential for chromosomal DNA replication (Christov et al. 2006).

Despite the fact that sbRNAs form a large and diverse family of ncRNAs, only a single representative, the most derived CeY RNA (encoded by the *yrn-1* gene) was found to bind 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. Of the 28 residues of the *Xenopus laevis* Ro60 protein that are in contact with the Y RNA (Stein et al. 2005), only 11 are conserved in frog and worm, while 27 are shared between human and frog. Of the 14 amino acids in contact with mis-folded RNAs, on the other hand, almost all that are conserved between frog and human are also conserved in the worm. We found here that the other nematode sbRNAs are more similar to human Y RNAs than to ceY, in particular in terms of their secondary structure. Taken together, this suggests that sbRNAs (except ceY) in fact do not bind to ROP-1 at all. In this context, the ill-defined role of *rop-1* in *C. elegans* dauer larvae formation is of interest, which suggests alternative binding partners of ROP-1. The *Caenorhabditis* sbRNAs, however, conserve a motif that was recently demonstrated by Gardiner et al. (2009) to be essential for the function of vertebrate Y RNAs in DNA replication.

It is very tempting, therefore, to speculate about an involvement of sbRNAs in nematode chromosomal DNA replication. Our unpublished data of a *C. elegans yrn-1* deletion, furthermore, indicate that the ceY RNA — in contrast to human Y RNAs — is not essential for chromosomal DNA replication. The available information suggests that the sbRNA family has undergone subfunctionalization that separated the RNA responsible for the Ro-related function (ceY) from a much larger family of sbRNAs responsible for the replication-associated functionality. If this is true, then reports (Kamath et al. 2003; Sönnichsen et al. 2005) that depletion of some individual sbRNAs does not cause a phenotype detectable in high throughput studies are not surprising. For the hypothetical role of sbRNAs in DNA replication it is plausible to speculate that either not all sbRNAs might be involved 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 indeed the case, research by reverse genetics will not be easy given that the sbRNA family comprises at least 18 paralogs in *C. elegans*.

All sbRNAs, including the previously described ceY RNA, are subject to strong evolutionary pressure on the conserved stem structure. The central loop, on the other hand, seems to evolve rapidly since conserved motifs in the central loop are only recognizable in closely related species. This extreme variability poses the question if these loop motifs are of biological relevance at all. Hogg and Collins (2008) suggest that the loop regions of Y RNAs might specify substrate specificities, although there is not direct evidence for this hypothesis. Without a clearly defined and experimentally supported functional role for sbRNA, one could only speculate about the reasons and implications of species-specific differences.

Supplemental Information

An Electronic Supplement located at http:www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-020/ compiles a list of detected sbRNAs, sequence data and alignments in machine-readable form.

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”, the AMS Vienna and the DFG under the auspices of the SPPs 1174 “Deep Metazoan Phylogeny” and 1258 “Sensory and Regulatory RNAs” and the AMS Vienna.

References

Abad P, Gouzy J, Aury JM, Castagnone-Sereno P, Danchin EG, Deleury E, Perfus-Barbeoch L, Anthouard V, Artiguenave F, Blok VC, Caillaud MC, Coutinho PM, Dasilva C, De Luca F, Deau F, Esquibet M, Flutre T, Goldstone JV, Hamamouch N, Hewezi T, Jaillon O, Jubin C, Leonetti P, Magliano M, Maier TR, Markov GV, McVeigh P, Pesole G, Poulain J, Robinson-Rechavi M, Sallet E, Ségurens B, Steinbach D, Tytgat T, Ugarte E, van Ghelder C, Veronico P, Baum TJ, Blaxter M, Bleve-Zacheo T, Davis EL, Ewbank JJ, Favery B, Grenier E, Henrissat B, Jones JT, Laudet V, Maule AG, Quesneville H, Rosso MN, Schiex T, Smant G, Weissenbach J, Wincker P (2008) Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. Nat. Biotechnol. 26:909–915

Aftab MN, He H, Skogerbø G, Chen R (2008) Microarray analysis of ncRNA expression patterns in *Caenorhabditis elegans* after RNAi against snoRNA associated proteins. BMC Genomics 9:278–278

Aerts S, Van Loo P, Thijs G, Moreau Y, De Moor B (2003) Computational detection of cis-regulatory modules. Bioinformatics 19:5–14

Barrière A, Yang SP, Pekarek E, Thomas CG, Haag ES, Ruvinsky I (2009) Detecting heterozygosity in shotgun genome assemblies: Lessons from obligately outcrossing nematodes. Genome Res. 19:470–80

Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF (2008) RNAalifold: improved consensus structure prediction for RNA alignments. BMC Bioinformatics 9:474–474

Blaxter ML, De Ley P, Garey JR, Liu LX, Scheldeman P, Vierstraete A, Vanfleteren JR, Mackey LY, Dorris M, Frisse LM, Vida JT, Thomas WK (1998) A molecular evolutionary framework for the phylum Nematoda. Nature 392:71–75

Chen X, Wurtmann EJ, Van Batavia J, Zybailov B, Washburn MP, Wolin SL (2007) An ortholog of the Ro autoantigen functions in 23s rRNA maturation in *D. radiodurans*. Genes Dev. 21:1328–1339

Chilton NB, Huby-Chilton F, Gasser RB, Beveridge I (2006) The evolutionary origins of nematodes within the order Strongylida are related to predilection sites within hosts. Mol. Phylogenet. Evol. 40:118-128

Christov CP, Gardiner TJ, Szüts D, Krude T (2006) Functional requirement of noncoding Y RNAs for human chromosomal DNA replication. Mol. Cell. Biol. 26:6993–7004

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

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res. 14:1188–1190

Dávila López M, Rosenblad MA, Samuelsson T (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, Zhu X, Skogerbø G, Zhao Y, Fu Z, Wang Y, He H, Cai L, Sun H, Liu C, Li B, Bai B, Wang J, Jia D, Sun S, He H, Cui Y, Wang Y, Bu D, Chen R (2006) Organization of the *Caenorhabditis elegans* small noncoding transcriptome: genomic features, biogenesis, and expression. Genome Res. 16:20–29

Gardiner TJ, Christov CP, Langley AR, Krude T (2009) A conserved motif of vertebrate Y RNAs essential for chromosomal DNA replication. RNA 15:1375–85

Green CD, Long KS, Shi H, Wolin SL (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 AR, Kilgus C, Mosig A, Hofacker IL, Hennig W, Stadler PF (2008) Arthropod 7SK RNA. Mol. Biol. Evol. 25:1923–1930

Guffanti E, Corradini R, Ottonello S, Dieci G (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, Ma Y, Mathews MB (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, de Jong D, Marz M, Rose D, Tafer H, Tanzer A, Schierwater B, Stadler PF (2009) Non-coding RNA annotation of the genome of *Trichoplax adhaerens*. Nucleic Acids Res. 37:1602–1615

Hertel J, Hofacker IL, Stadler PF (2008) SnoReport: computational identification of snoRNAs with unknown targets. Bioinformatics 24:158–164

Hofacker IL, Fekete M, Stadler PF (2002) Secondary structure prediction for aligned RNA sequences. J. Mol. Biol. 319:1059–1066

Hogg RJ, Collins K (2008) Structured non-coding RNAs and the RNP Renaissance. Curr. Op. Chem. Biol. 12:684–689

Kaczkowski B, Torarinsson E, Reiche K, Havgaard JH, Stadler PF, Gorodkin J (2009) Structural profiles of miRNA families from pairwise clustering. Bioinformatics 25:291–294

Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, Welchman DP, Zipperlen P, Ahringer J (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. Nature 421:231–237

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

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

Marz M, Kirsten T, Stadler PF (2008) Evolution of spliceosomal snRNA genes in metazoan animals. J. Mol. Evol. 67:594–607

Missal K, Zhu X, Rose D, Deng W, Skogerbø G, Chen R, Stadler PF (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

Mitreva M, Blaxter ML, Bird DM, McCarter JP (2005) Comparative genomics of nematodes. Trends Genet. 21:573–581

Mosig A, Guofeng M, Stadler BM, Stadler PF (2007a) Evolution of the vertebrate Y RNA cluster. Theory Biosci. 126:9–14

Mosig A, Chen JL, Stadler PF (2007b) Homology search with fragmented nucleic acid sequence patterns. In R. Giancarlo and S. Hannenhalli (Eds.), Algorithms in Bioinformatics (WABI 2007), Volume 4645 of Lecture Notes in Computer Science, Berlin, Heidelberg, pp. 335–345. Springer Verlag.

Pagano A, Castelnuovo M, Tortelli F, Ferrari R, Dieci G, Cancedda R (2007) New small nuclear RNA gene-like transcriptional units as sources of regulatory transcripts. PLoS Genet. 3:e1

Ploner A, Ploner C, Lukasser M, Niederegger H, Hüttenhofer A (2009) Methodological obstacles in knocking down small noncoding RNAs. RNA 15:1797–804

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

Sönnichsen B, Koski LB, Walsh A, Marschall P, Neumann B, Brehm M, Alleaume AM, Artelt J, Bettencourt P, Cassin E, Hewitson M, Holz C, Khan M, Lazik S, Martin C, Nitzsche B, Ruer M, Stamford J, Winzi M, Heinkel R, Röder M, Finell J, Häntsch H, Jones SJ, Jones M, Piano F, Gunsalus KC, Oegema K, Gönczy P, Coulson A, Hyman AA, Echeverri CJ (2005) Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*. Nature 434:462–469

Stein AJ, Fuchs G, Fu C, Wolin SL, Reinisch KM (2005) Structural insights into RNA quality control: the Ro autoantigen binds misfolded RNAs via its central cavity. Cell 121:529–539

Sudhaus W, Kiontke K (2007) Comparison of the cryptic nematode species *Caenorhabditis brenneri sp.n.* and *C. remanei* (Nematoda: Rhabditidae) with the stem species pattern of the *Caenorhabditis Elegans* group. Zootaxa 1456:45–62

Tanzer A, Stadler PF (2004) Molecular evolution of a microRNA cluster. J. Mol. Biol. 339:327–335

Tanzer A, Stadler PF (2006) Evolution of MicroRNAs. In: Ying, SY (ed) MicroRNA Protocols. Humana Press, Totowa, NJ, pp. 335–350

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

Van Horn DJ, Eisenberg D, O'Brien CA, Wolin SL (1995) *Caenorhabditis elegans* embryos contain only one major species of Ro RNP. RNA 1:293–303

Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview Version 2 – a multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189–1191

Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comput. Biol. 3:e65

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

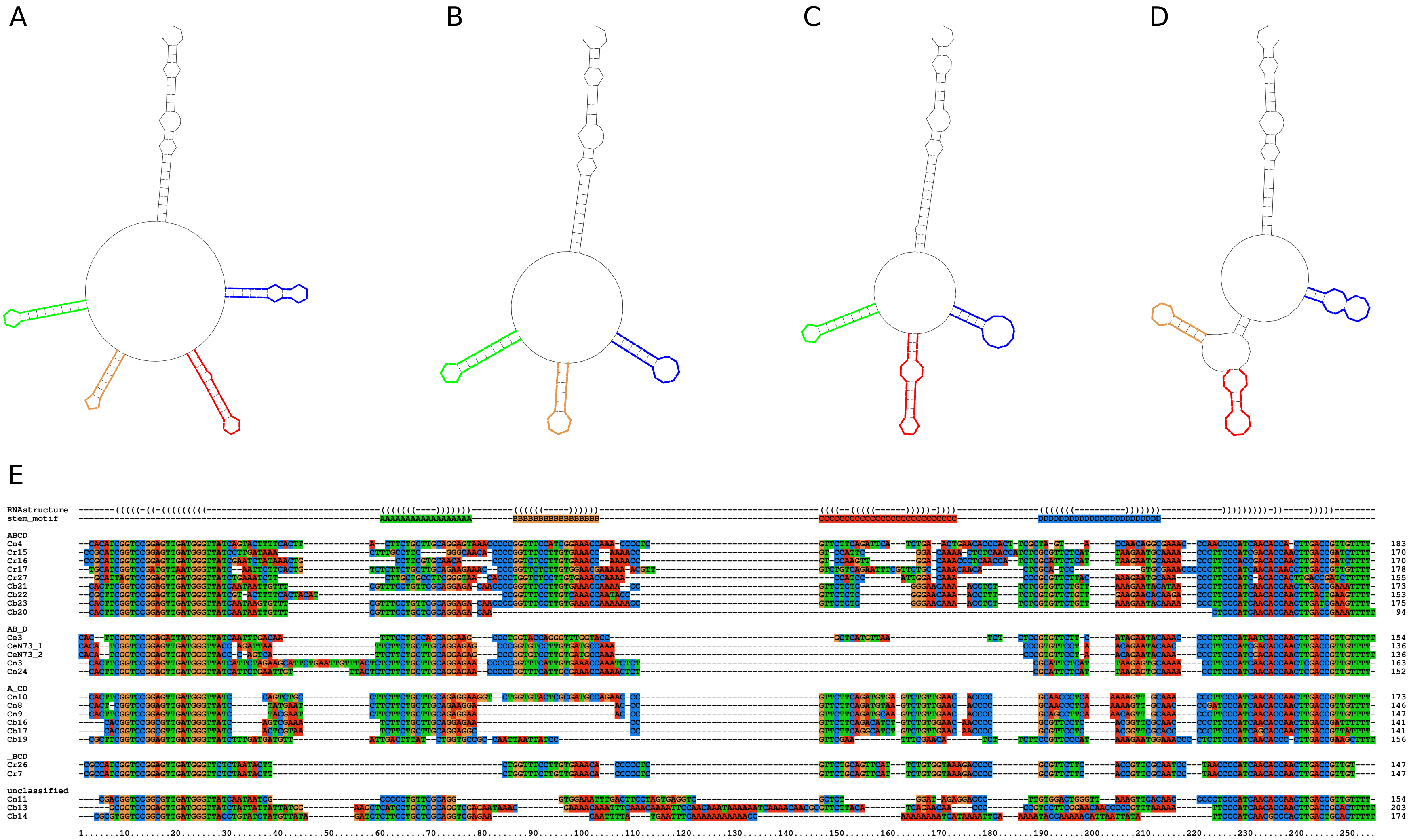


Fig. 7 Supplemental Fig.S1
Structure evolution of sbRNA loop regions. Gene duplication coincides with duplication of substructures within the loops regions as shown here for members of the long sbRNAs residing on *C. elegans* chromosome V. The ancestral gene of this family possibly consisted of four hairpins. During subsequent gene duplications, individual hairpins were lost. The hairpin B (yellow) shows high sequence similarity to the adjacent hairpin A (green) and probably arose by local duplication of a structural element. A,B,C,D: RNA secondary structures of representatives of each family. family ABCD: Cn4 (**A**); family AB_D: CeN73_1 (**B**); family A_CD: Cn8 (**C**); family _BCD: Cr26 (**D**). (**E**) hand curated CLUSTAL W multiple sequence alignments including the consensus structure and location of structural motifs.

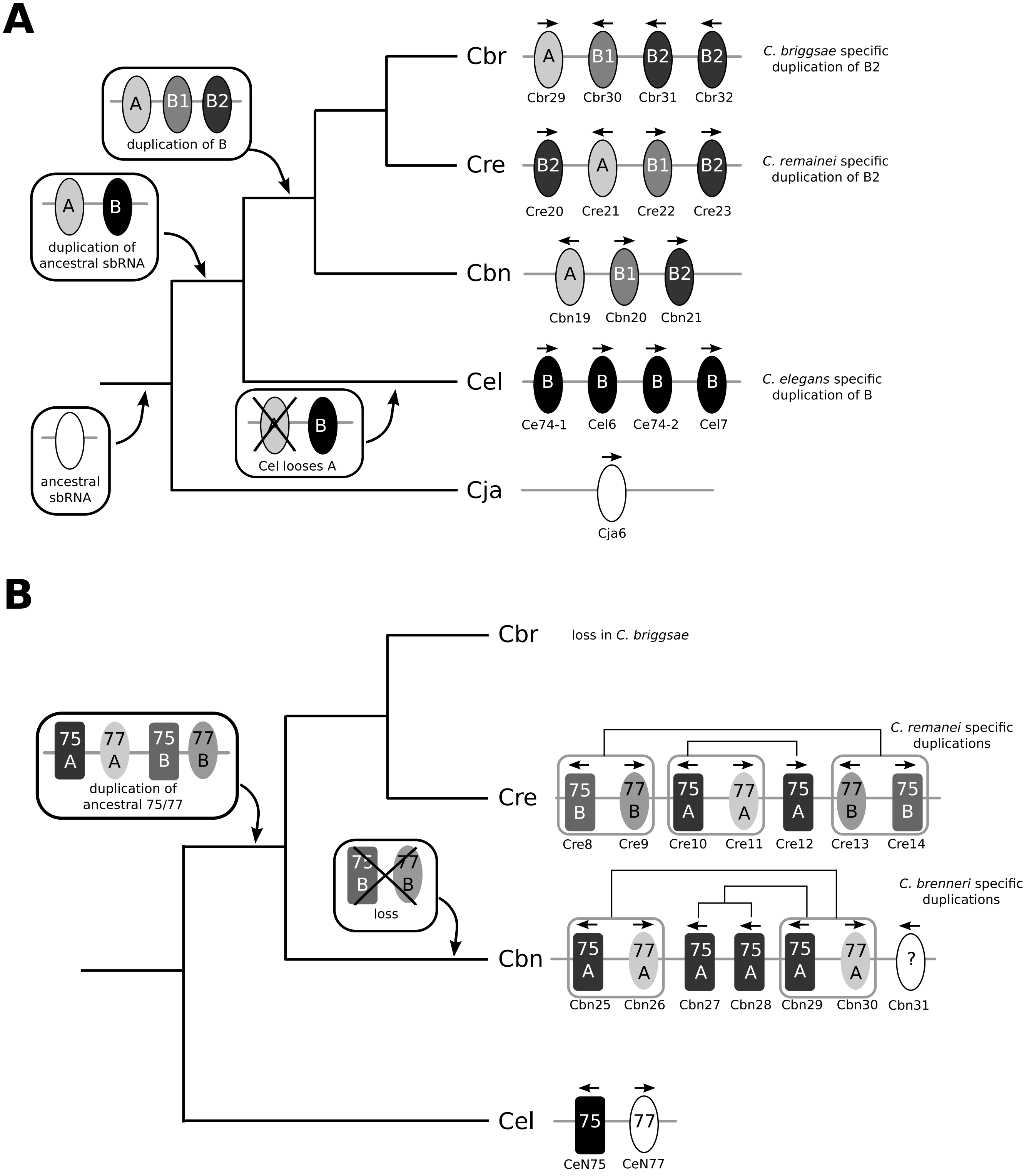


Fig. 8 Supplemental Fig.S2
Evolutionary history of the *C. elegans* sbRNA clusters on chromosome X (**A**) and chromosome III (**B**) of *C. elegans* and their homologs. Clusters are shaped by duplications and deletions before and after speciation events. On the clusters on chromosome X, genes are often duplicated as units of two genes in tail-to-tail orientation (shown in grey boxes). Arrows indicate sbRNA orientation: plus strand (→) and minus strand (←).For details see text and Fig. 6.