2676–2685 *Nucleic Acids Research, 2006, Vol. 34, No. 9*
doi:10.1093/nar/gkl359

Evolution of small nucleolar RNAs in nematodes

Anja Zemann, Anja op de Bekke, Martin Kiefmann, Jürgen Brosius and Jürgen Schmitz

Institute of Experimental Pathology (ZMBE), University of Münster, D-48149 Münster, Germany

Received March 24, 2006; Revised April 11, 2006; Accepted April 24, 2006

ABSTRACT

In contrast to mRNAs, which are templates for translating proteins, non-protein coding (npc) RNAs (also known as 'non-coding' RNA, ncRNA), exhibit various functions in different compartments and developmental stages of the cell. Small nucleolar RNAs (snoRNAs), one of the largest classes of npcRNAs, guide post-transcriptional modifications of other RNAs that are crucial for appropriate RNA folding as well as for RNA–RNA and RNA–protein interactions. Although snoRNA genes comprise a significant fraction of the eutherian genome, identifying and characterizing large numbers of them is not sufficiently accessible by classical computer searches alone. Furthermore, most previous investigations of snoRNAs yielded only limited indications of their evolution. Using data obtained by a combination of high-throughput cDNA library screening and computational search strategies based on a modified DNAMAN program, we characterized 151 npcRNAs, and in particular 121 snoRNAs, from *Caenorhabditis elegans* and extensively compared them with those in the related, *Caenorhabditis briggsae*. Detailed comparisons of paralog snoRNAs in the two nematodes revealed, in addition to *trans*-duplication, a novel, *cis*-duplication distribution strategy with insertions near to the original loci. Some snoRNAs coevolved with their modification target sites, demonstrating the close interaction of complementary regions. Some target sites modified by snoRNAs were changed, added or lost, documenting a high degree of evolutionary plasticity of npcRNAs.

INTRODUCTION

Two very surprising discoveries have arisen from the Human Genome Project. One, humans do not have significantly more protein-coding genes than other mammals; and two, sequences corresponding to protein open reading frames comprise only 1.5% of our genome (1). The unavoidable conclusion to be drawn from this is that the differences that separate humans from other species may reside in the remaining 98.5% of the genome that encode untranslated functional RNAs and regulatory regions, or constitutes non-genic regions. The present work focuses on a defined population of non-protein coding RNAs (npcRNAs), often not quite correctly termed 'non-coding' RNA (ncRNA), derived from a *Caenorhabditis elegans* cDNA library generated with size-fractionated RNA (70–600 nt). The size limitation, while excluding mature microRNAs (miRNAs), short interfering RNAs and large ribosomal RNAs (rRNAs) that are well described elsewhere (2,3), yields predominantly small nucleolar RNAs (snoRNAs) and spliceosomal RNAs. snoRNAs are 60–300 nt long and guide the post-transcriptional modifications of ribosomal and other RNAs. Such modifications are crucial for appropriate RNA folding as well as for RNA–RNA and RNA–protein interactions (4). Furthermore, snoRNAs are thought to be involved in epigenetic mechanisms regulating gene expression. In this context, deletion of certain imprinted snoRNA clusters in the cerebral cortex is thought to play a causative role in the Prader–Willi Syndrome of mental retardation (5–7).

Based on structural motifs and function the snoRNA family is divided into two subclasses: C/D-box snoRNAs (C-box consensus UGAUGA; D-box consensus CUGA) and H/ACA snoRNAs (H-box consensus ANANNA and box ACA), which interact directly by base complementarity to their target rRNA and spliceosomal RNA sequences to direct 2′-*O*-ribose methylation and pseudouridylation, respectively. The complementary regions, known as 'antisense elements', reside at the 5′ and/or 3′ ends of snoRNAs. Although snoRNA modifications were initially thought to be restricted to rRNA and to be localized strictly in the nucleolus, a growing list of npcRNAs including transfer RNAs (tRNAs) are also modified by snoRNAs (4,8), and they have also been found in Cajal bodies, nucleoplasmic substructures involved in processing npcRNAs (9). The spectrum of snoRNA targets could potentially include even mRNAs, although it cannot be excluded yet that such existing base complementarities are simply fortuitous and without biological significance (5). Most vertebrate snoRNAs are derived from introns of pre-mRNA transcripts, especially those from ribosomal protein

*To whom correspondence should be addressed. Tel: +49 251 8352133; Fax: +49 251 8352134; Email: jueschm@uni-muenster.de

genes (RPGs) and other housekeeping proteins, and are processed in a complex sequence involving endonucleases, exonucleases and helicases (10,11). Interestingly, a growing number of host genes do not yield translatable mRNAs, and it appears that the main function of the corresponding genes and primary transcripts is the expression of snoRNAs (12–14). Many miRNAs are also hosted by npcRNAs (15).

Systematic searches using experimental RNomics, an EST-like approach tailored for small RNAs, have successfully identified large numbers of npcRNAs in Mouse (16), *Drosophila* (17), *Arabidopsis* (18) and Archaea (8). To better elucidate the evolutionary pathways of snoRNAs, we have now extended this search to the nematode, *C.elegans*, an extremely interesting model eukaryote with a simple body plan but complicated genomics including, for example, *cis*, *trans* and alternative splicing systems. As intermediates between single-celled organisms and 'higher' metazoan animals, they offer an excellent system for studies on metazoan genome function and evolution. To provide a large enough dataset for exhaustive analysis of snoRNAs in *C.elegans* we have now combined high-throughput, experimental RNomic screening with computational methods focused on RPGs and other introns of genes that harbor snoRNAs identified in our experimental approach. Furthermore, we have analyzed the phylogeny of snoRNAs by comparing the above results with those of *Caenorhabditis briggsae*, a nematode that shared a common ancestor with *C.elegans* some 100 million years ago (mya). Our three-pronged approach revealed possible mechanisms of how novel snoRNAs arose, spread in the genome, changed targets or were lost over the course of evolution.

MATERIALS AND METHODS

The experimental procedures concerning construction and analysis of libraries are described in Hüttenhofer *et al.* (16). Detailed methods for constructing the *C.elegans* library are given in Supplementary Data.

Computational strategies

The commercial software package DNAMAN was modified, in collaboration with the Lynnon Corporation, to computationally screen defined databases of intronic sequences for snoRNAs and to identify snoRNA modification target sites from compiled RNA databases. The modified DNAMAN version is available from http://www.lynnon.com (Mac OS X version 6018 or later). Note that additional freeware is available at http://lowelab.ucsc.edu/snoscan to analyze C/D-box snoRNAs (19), at http://lowelab.ucsc.edu/snoGPS for H/ACA-box snoRNAs (20) and at http://rna.tbi.univie.ac.at/cgi-bin/alifold.cgi to analyze secondary structural prediction (21).

Computational search for snoRNAs in *C.elegans*

The modified DNAMAN software allowed us to apply complex search profiles to find potential snoRNAs in a compilation of *C.elegans* introns of RPGs and genes that harbor experimentally identified snoRNAs. The following search was applied for C/D-box snoRNAs:

TGATGA(N9-35)CTGA(N4-35)TGATGA(N9-35)CTGA <mismatch=3 2 3 1>

A maximum of three mismatches were allowed in the first, two in the second, three in the third and one in the last sequence motif. N9-35 and N4-35 denote variable sequence stretches of at least 9 or 4, respectively, and a maximum of 35 nt. The search motif for H/ACA-box snoRNAs was ANA(NN)A(N50-100)ACA. No mismatches were allowed. Both searches were accompanied by intensive structural evaluations of the computationally predicted snoRNAs (Supplementary Figure 1).

The pattern search of DNAMAN is implemented in C language. Details of the search procedure are provided by the Lynnon Corporation (Supplementary Data).

5′ Extension of experimentally found snoRNAs

BLAST searches of the cDNA sequences were made against the *C.elegans* non-redundant (nr) NCBI database (http://www.ncbi.nlm.nih.gov/blast), the Santa Cruz server (http://genome.ucsc.edu/cgi-bin/hgBlat), the Sanger database (http://www.ensembl.org/Caenorhabditis_elegans/blastview) or the RPG databank (http://ribosome.miyazaki-med.ac.jp). Thus, the sequences absent in the truncated cDNAs (usually some 10 nt) were extended with the aid of genomic sequences. The mature 5′ ends were estimated by structural requirements of mature snoRNAs (Supplementary Figure 1).

Target site search

A compiled library of all *C.elegans* rRNA, spliceosomal and tRNA genes was searched with the modified DNAMAN software for potential snoRNA target motifs. For C/D-box snoRNA target sites we allowed a maximum of three G–U pairs and a minimum length of 9 nt. For H/ACA-box snoRNAs we used a similar search profile but allowed a split of target sites in four or five contiguous nucleotides. The detailed search process is shown in Supplementary Figure 1.

Comparison with *C.briggsae* sequences

The same database sources as mentioned above were used to computationally detect orthologous snoRNAs in *C.briggsae*.

Secondary structures of snoRNAs

The secondary structures of all experimentally and computationally identified snoRNAs were derived using the M-fold program (22); http://www.bioinfo.rpi.edu/applications/mfold/old/rna.

RESULTS

Analysis of the size-fractionated cDNA library of *C.elegans*

Following high-density array hybridization of 38 400 cDNA sequence clones to exclude known small npcRNAs or fragments of degraded large rRNAs (Supplementary Figure 2), we selected 4673 clones for sequencing. Exclusion of unreadable or very short sequences, empty vectors, *E.coli* contaminations and other ambiguities yielded 3294 clones; among these we identified 15 known spliceosomal RNAs (294 sequences), 41 known tRNAs (322 sequences), 3 isoforms of SRP (signal recognition particle) RNA (736 sequences), 29 different parts of known rRNAs that escaped prior

exclusion (1180 sequences), 22 known mRNAs (31 sequences), 7 splice leader RNA sequences (SL; 64 sequences) and two histone hairpin RNAs (2 sequences) all of which were excluded from a more detailed analysis (SL, SRP, histone hairpin and spliceosomal RNAs are listed in Supplementary Data). The remaining 665 sequences contained 120 npcRNAs including 91 snoRNAs (Figure 1).

Computational screening for additional npcRNA candidates

In addition to those npcRNAs identified experimentally, computer searches based on the following arguments yielded another 23 snoRNAs (Figure 1). Yoshihama *et al*. (11) estimated that RPGs harbor about one-third of all snoRNAs in the human genome. In our experimentally identified snoRNAs we also observed that genes harboring one snoRNA in an intron are likely to encode additional snoRNAs in the same intron or in neighboring introns of the same gene. Consequently, we extracted and analyzed snoRNA candidates from introns of all known *C.elegans* RPGs and from other intronic sequences that were found in the proximity of our experimentally detected snoRNAs. We used the following stringent criteria to validate all computationally detected snoRNA candidates (Supplementary Figure 1): (i) presence of all snoRNA structural requirements and box motifs; (ii) identification of potential modification target site complementarities, (iii) sequence conservation in *C.briggsae*, and/or signals in northern blots. From >100 potential candidates (Supplementary Data) an additional 23 novel snoRNA candidates met these stringent criteria (Figure 1; comCe).

The reliability of the computational algorithm was confirmed in that we were also able to identify all but 17 of the experimentally found or previously predicted snoRNAs with these search criteria. Those snoRNAs not confirmed in the computer search were structurally modified and therefore did not match our search profile (data not shown). An additional BLAST search of Genbank genomic sequences revealed seven snoRNA paralogs (Figure 1, blCe, blCb) and one additional spliceosomal RNA (blCe378).

snoRNAs

Of all 154 experimentally or computationally identified sequences, 59 are novel snoRNAs candidates (Figure 1, I), while 65 of the recovered snoRNA candidates were recently

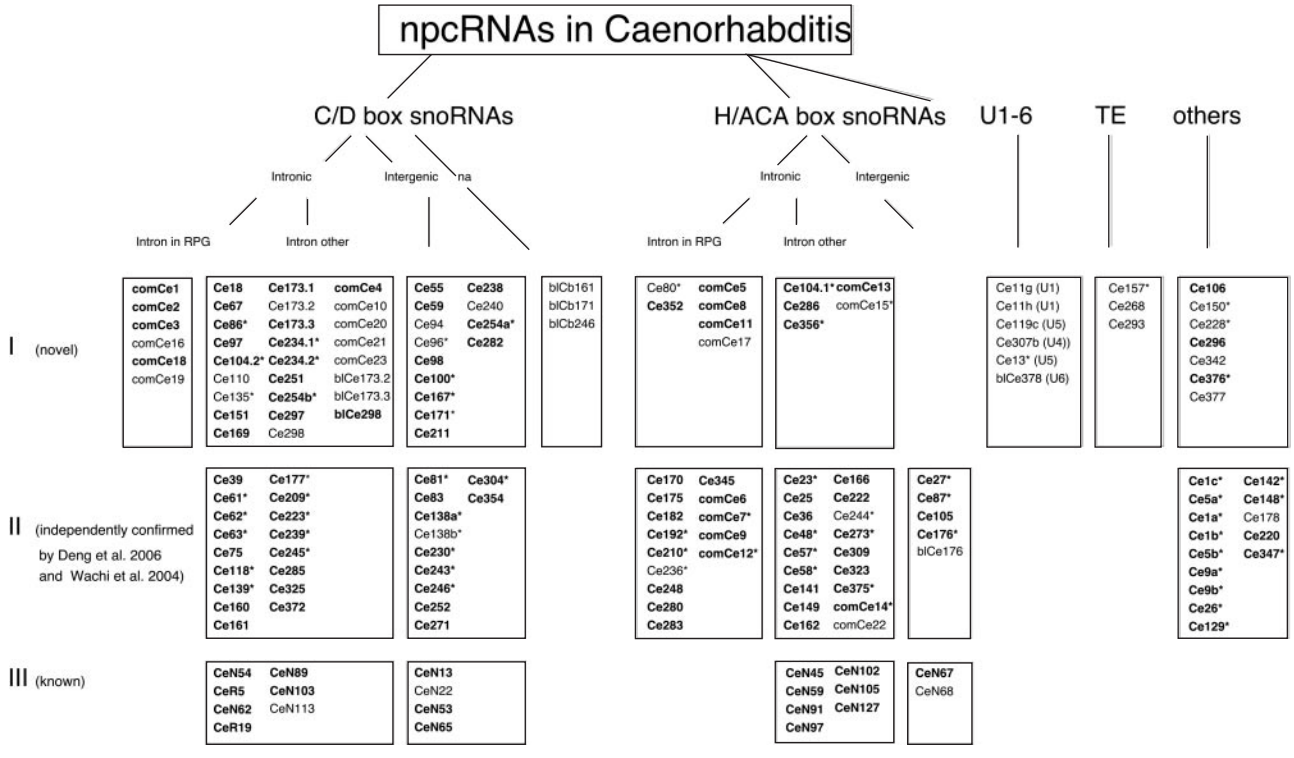

Figure 1. Grouping of experimentally and computationally identified npcRNAs. (I) Novel npcRNAs and (II) npcRNAs also confirmed recently (23,24). (III) Known snoRNAs (additional snoRNAs experimentally identified by Deng *et al*. (23) (CeN) or Wachi *et al*. (24) (CeR), but not revealed in our screen; included for completeness). Our computationally identified snoRNAs were derived from specific search profiles (com) or retrieved via BLAST searches (bl). snoRNAs were subdivided into C/D-box and H/ACA-box snoRNAs and were found in intronic, intergenic or unidentified (not analyzed = na) regions. npcRNAs found in *C.elegans* (Ce) as well as at orthologous positions in *C.briggsae* (Cb) are shown in boldface. Asterisks denote verification by northern blot analysis. Owing to the large number of spliceosomal RNA paralogs, we were not able to identify the true orthologs of the respective U1–U6 snRNAs in *C.briggsae*. RPG, ribosomal protein genes; U1–U6, spliceosomal RNAs; TE, npcRNAs homologous to transposed elements; Others, uncategorized npcRNAs. Additional, known npcRNAs that were experimentally verified and then excluded from further analysis are listed in Supplementary Data. *Note*: The group of experimentally identified spliceosomal RNAs were also detected (23), but incorrectly classified as already known. Careful examination, however, shows that all of them are novel spliceosomal RNA isoforms.

confirmed experimentally (23) (59 candidates) or (24) (6 candidates) (Figure 1, II; Supplementary Data). For completeness, Figure 1 (III) also lists 20 other snoRNA candidates that were not recovered by our screen, but were identified previously by either Deng *et al*. (23) (18 candidates) or Wachi *et al*. (24) (2 candidates), and is thus now a compilation of all presently known *C.elegans* snoRNAs.

Altogether, we found 76 unambiguous snoRNA candidates with motifs, secondary structure elements and recognizable target modification complementarities characteristic of C/D-box snoRNAs. Based on their chromosomal locations, individual candidates could be described as either intronic or intergenic snoRNAs (Figure 1 and Supplementary Data). All but 16 are also potentially functional in *C.briggsae* and are located at orthologous loci; 10 of these 16 were recognizable, but diverged, at orthologous positions in *C.briggsae* (Supplementary Figure 3). Presumably, they became inactive pseudogenes that lack motifs and structures to function as bona fide snoRNAs. Interestingly, in all C/D-box snoRNA candidates (as well as the H/ACA-box snoRNAs) we identified a characteristic uridine-rich region adjacent to the mature 3′ ends. This sequence has previously been implicated in maturation of H/ACA-box snoRNAs only (25). We also found a C/D-box homodimer (Ce234) and a chimeric C/D-H/ACA-box snoRNA (Ce104). Northern blot analysis of both resulted in hybridization to only dimeric forms, indicating that the respective dimers are the mature forms of these snoRNAs. Interestingly, we detected only six C/D-box snoRNA candidates in RPG introns compared with 20 H/ACA-box snoRNAs (Figure 1).

We also identified 48 H/ACA-box snoRNAs that were localized to intronic and intergenic regions (Figure 1 and Supplementary Data). Only seven of those are probably not functional in *C.briggsae* (Figure 1). The sequences of three H/ACA-box snoRNA orthologs are apparently diverged pseudogenes in *C.briggsae* (Supplementary Figure 3; comCb17, comCb22, blCb176).

snoRNA modification target sites and distribution patterns

In keeping with their function, snoRNAs have dual binding capacity for both small RNA modifying proteins and, via specific sequence complementarity, for their target RNAs. We identified complementarities for potential modification targets in 5S, 5.8S, 18S and 26S rRNAs; in U1, U2, U4, U5 and U6 spliceosomal RNAs and in tRNAs. Twelve 26S rRNA target sites are supported by the presence of nucleotide modifications (26) (Supplementary Figure 3, black dots). This is the first time that C/D-box snoRNAs in eukaryotes have been identified with potential target sites in various tRNAs (Ce62-tRNA^{Ile}, Ce63-tRNA^{Ser}, Ce94-tRNA^{Asn}, Ce246-tRNA^{Ile}, comCe3-tRNA^{Thr}, comCe18-tRNA^{Arg}) (Figure 2). tRNA modifications guided by snoRNAs have been reported thus far only in Archaea (8,27). Another interesting observation was the presence of two antisense elements in some of our snoRNAs (e.g. Ce173.3, Ce251, Ce298, Ce23) with complimentary regions suggesting the potential to modify target RNAs located in two different subcellular compartments. These snoRNAs are predicted to modify rRNAs that occur in the nucleolus, as well as U1, U4, U5 spliceosomal RNAs that are present in Cajal bodies. Even an individual antisense element has the potential to be complementary to more than one hypothetical target site (Supplementary Figure 3).

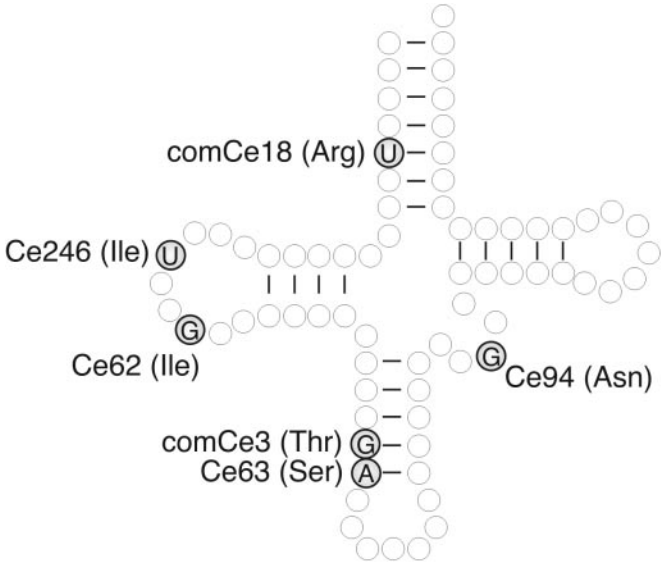

Figure 2. Possible C/D-box snoRNA modification targets in tRNAs. The modified nucleotides are circled. Computationally (com) and experimentally identified *C.elegans* snoRNAs (Ce) correspond to those listed in Figure 1.

From our 121 experimentally and computationally identified snoRNAs in *C.elegans*, 98 potentially functional orthologs were identified in *C.briggsae* (Figure 1). Forty of these orthologous pairs contain matching sequence complementarities to the same RNA modification targets in *C.briggsae* and *C.elegans* (Supplementary Figure 3a and b). Surprisingly, the potential target sites for the majority of them changed over a period of 100 million years (Supplementary Figure 3c and d).

snoRNA paralogs, generated perhaps by gene duplication, have been observed frequently and are a potential source for the creation of novel snoRNAs (28). We identified 20 snoRNAs and their corresponding paralogs (11 pairs are orthologs in both *C.elegans* and *C.briggsae*, 6 pairs in *C.elegans* and 3 pairs in *C.briggsae* only; Figure 3a). To help determine whether the computationally identified H/ACA-box snoRNA paralogs are functional, we analyzed the compensatory nucleotide substitution patterns in their double-stranded stem structures. Compensatory changes tend to maintain the secondary structure of stem regions and indicate selection pressure for functionality. Characteristic compensatory changes could be found for all identified H/ACA-box snoRNA paralogs suggesting that they retained their functionality, at least for a sufficient period to form individual compensations after duplication (data not shown). Compensatory substitution pattern analyses in C/D-box snoRNAs are not of much help in determining their functionality because they do not possess sufficient amounts of double-stranded structures; thus, C/D-box snoRNAs were omitted from this analysis.

The chromosomal localization of snoRNA paralogs could be categorized based on two distinguishing events: those in which snoRNA paralogs inserted into different positions in the same gene (*cis*-duplication), and those in which snoRNA

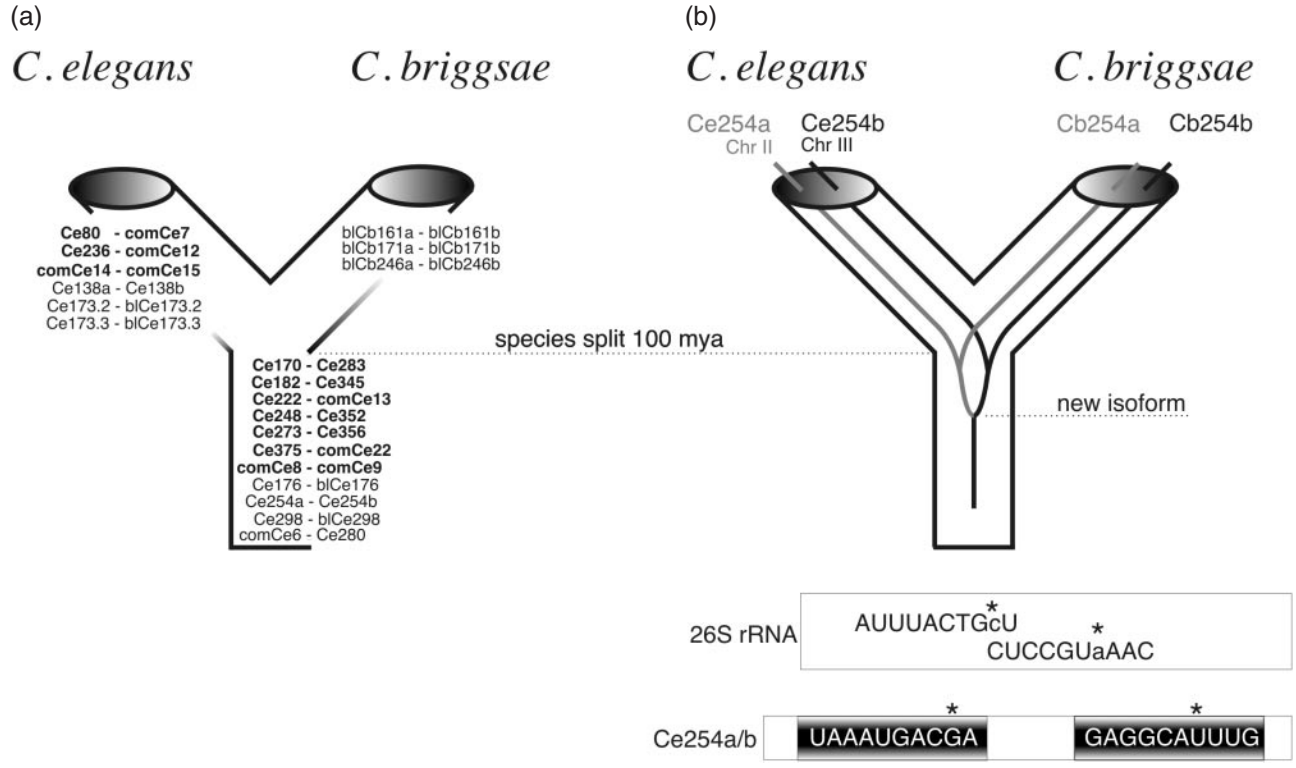

Figure 3. snoRNAs and their paralogs in *C.elegans* and *C.briggsae*. (**a**) Compilation of snoRNAs and their paralogs in *C.elegans* and *C.briggsae* prior to (base of the tree), and after the two species split (branches). *Cis*-duplication is shown in boldface, *trans*-duplication in regular letters. (**b**) Evolutionary scenario for Ce254 in *C.elegans* and *C.briggsae*. This snoRNA *trans*-duplicated to yield a novel paralog located on a different chromosome in the common ancestor of *C.elegans* and *C.briggsae*. After the two species split, both paralogs retained their functionality in both species. Asterisks denote the modified nucleotide of 26S rRNA (top) or the corresponding complementary sequence positions in the snoRNA antisense regions (bottom). Chr II and Chr III indicate the snoRNA location on chromosomes II and III.

paralogs inserted into target genes (or intergenic regions) other than the original host gene or, in the event of host gene duplication, moved to a different chromosomal location along with the host gene (*trans*-duplication).

Cis-duplication of snoRNA paralogs. The presence and/or absence of 20 snoRNA paralogs were analyzed in *C.elegans* and *C.briggsae*. Figure 4a shows a *C.elegans* snoRNA (comCe12) that is conserved at the orthologous locus (intron 2 of the rps-29 gene) in *C.briggsae*. A paralog of this snoRNA (Ce236) was also present in our cDNA library. In *C.elegans* Ce236 is located in intron 1 of the same rps-29 host gene, while the orthologous position in the rps-29 gene of *C.briggsae* is empty. Since the probability of a clean excision of the snoRNA without parts of the flanking sequences at this position in *C.briggsae* is negligible, this indicates a duplication process involving integration into the adjacent intron (*cis*-duplication) after *C.elegans* split from a common ancestor with *C.briggsae*. Interestingly, the function of Ce236 in *C.elegans* may have been replaced in *C.briggsae* by another non-orthologous snoRNA (Cb309) that has the potential to modify the identical nucleotide in 26S rRNA, while the Cb309 ortholog in *C.elegans* (Ce309) modifies a target sequence in 18S rRNA. A similar scenario is shown in Figure 4b. We found a snoRNA (comCe7) present at orthologous positions of the rpl-24 gene in *C.elegans*, *C.briggsae* and *Caenorhabditis remanei*. A paralog (Ce80) could be detected in intron 1 of the same gene in *C.elegans* only. Figure 4c shows a snoRNA (comCe14) present in orthologous positions of the hypothetical protein gene K07C5 in *C.elegans* and *C.briggsae*. A corresponding paralog is found in intron 7 of the same gene in *C.elegans* (comCe15) but not in *C.briggsae*. In all presence/absence cases examined, the intronic sequences flanking the duplicated snoRNAs were recognizable at the corresponding, empty loci.

Trans-duplication of snoRNA paralogs. We could distinguish two forms of *trans*-duplications, both of which are exemplified in Figure 5. In some instances of segmental duplications of entire genes that harbor snoRNAs in one or more introns, the snoRNA did not move to another part of the host gene but hitchhiked with the host to a new location after duplication. In other cases, snoRNAs inserted into introns of a new host gene without traces of the original host gene, or into a new intergenic location. Figure 5a describes an example of segmental duplication of a hypothetical protein gene (C06A1.3) yielding a duplicated pseudogene (with respect to the protein-coding capacity) including the paralogous snoRNAs Ce173.1-3. Figure 5b shows two experimentally identified snoRNA paralogs; Ce254b is located in intron 1 of a hypothetical protein gene (Y53F4B.12). The paralog Ce254b duplicated, along with the 5′ (~100 nt) and 3′ (~50 nt) sequences of its original flanking intron, but

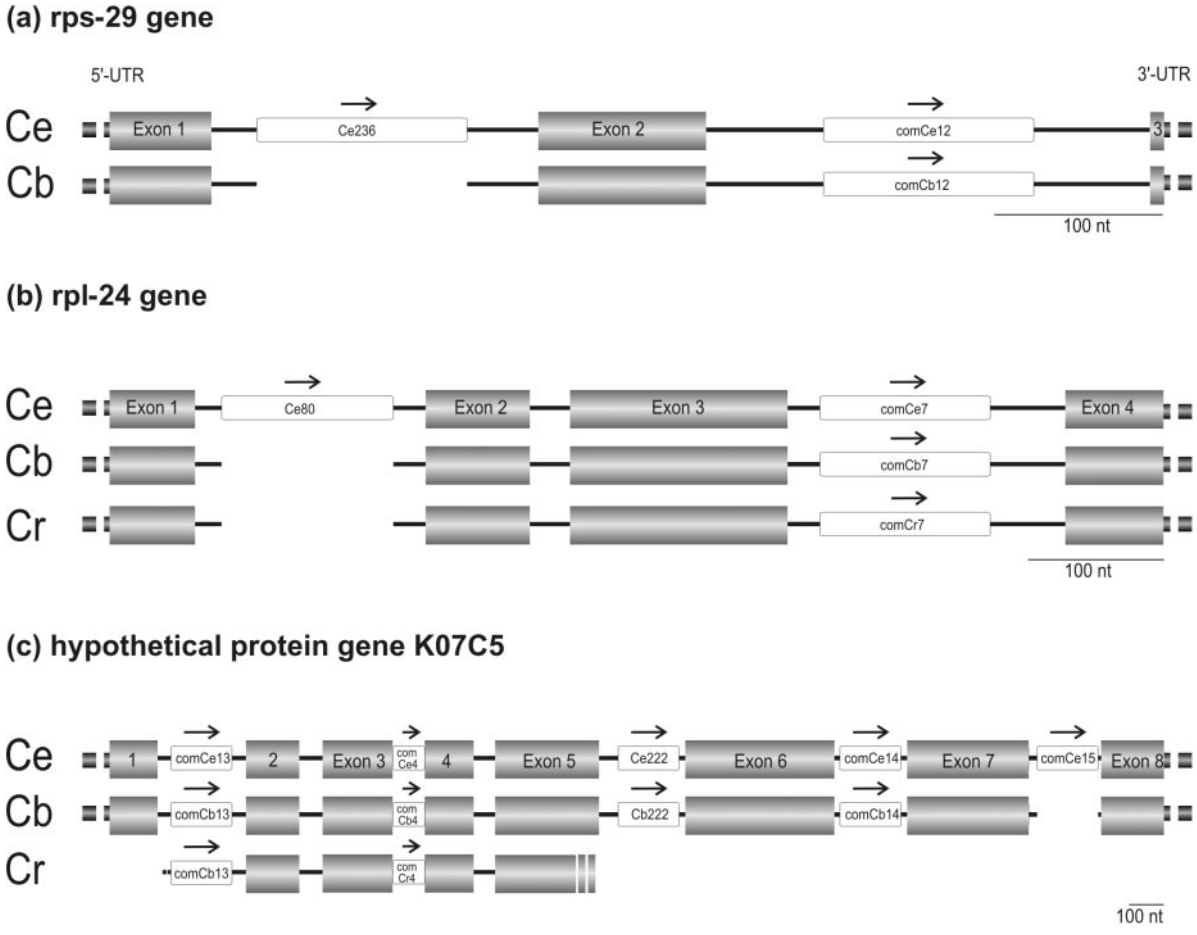

Figure 4. Examples of H/ACA-box snoRNAs. *Cis*-duplication characterized by presence/absence analyses in two or three nematode species (Ce, *C.elegans*; Cb, *C.briggsae*; Cr, *C.remanei*). Protein-coding regions are indicated by thick gray boxes, 5′- and 3′-UTRs are shown as intermediate-sized gray bars and introns as black lines. The framed white areas represent the snoRNAs with their orientation indicated by arrows. (**a**) *Cis*-duplication of comCe12 in *C.elegans* (Ce236) and not in *C.briggsae*. (**b**) *Cis*-duplication of comCe7 in *C.elegans* (Ce80) but not in *C.briggsae or C.remanei*. (**c**) Duplication of comCe14 in *C.elegans* (comCe15) but not in *C.briggsae* or *C.remanei*. Only a partial sequence of *C.remanei* was available. Several additional snoRNAs exist in neighboring introns of K07C5. Note that in *C.elegans*, *C.briggsae* and *C.remanei* the C/D-box snoRNA comCe4 located in intron 3 occupies the entire intron with the exception of two additional guanosine residues that are part of the functional splice sites.

without detectable surrounding exons, and migrated into a new location on a different chromosome. Interestingly, the separate left and right antisense elements of Ce254 (Figure 3b, bottom) modify bases in 26S rRNA that are shifted by only 6 nt. Hence, the sequences on 26S rRNA that are complementary to the two snoRNA antisense elements overlap by 2 nt (Figure 3b, top). This indicates that modification of the two methylation targets is not likely to occur at the same time. We found both paralogs at orthologous positions in *C.briggsae*, indicating that the duplication event took place in a common ancestor of both worms. Both forms retained their modification targets over 100 my demonstrating strong functional constraints. The fact that two conserved snoRNA paralogs modify the same targets indicates that one may not be enough to perform modification of all rRNA molecules, and that quantitative aspects play an important role in snoRNA function. Figure 5c and d describe snoRNA paralogs that are located in totally different surroundings following duplication. In the latter case it is noteworthy that the comCe6 paralog moved from one RPG (rpl-7) to another (rps-13) as the Ce280 paralog or vice versa.

Functional plasticity of snoRNA paralogs. Data provided by both the experimental and computational searches, as well as comparisons of paralogous snoRNAs in both *C.elegans* and *C.briggsae* enabled us to analyze target sites and hence function of duplicated snoRNA genes. We observed three different fates of the snoRNAs following duplication: (i) one of the paralogs apparently became inactive and decayed during the course of evolution; (ii) the new paralog maintained the same function as the original snoRNA and (iii) the new paralog either partially (one antisense element maintained the same target and the other acquired a new one) or fully diverged with respect to the complementary targets. Of the 20 pairs of paralogs, we found 4, 16 and 10 examples for the above three scenarios, respectively. One example of target site plasticity is illustrated in C/D-box snoRNA Ce246, which was detected experimentally in the *C.elegans* cDNA library and computationally in *C.briggsae*. In *C.briggsae* one paralog differs from the other mainly by a 2 nt deletion 5′ adjacent to the D′-box, shifting the methylation site by 2 nt (blCb246a-blCb246b). In 26S rRNA G_{860} is modified by one paralog and A_{862} by the other.

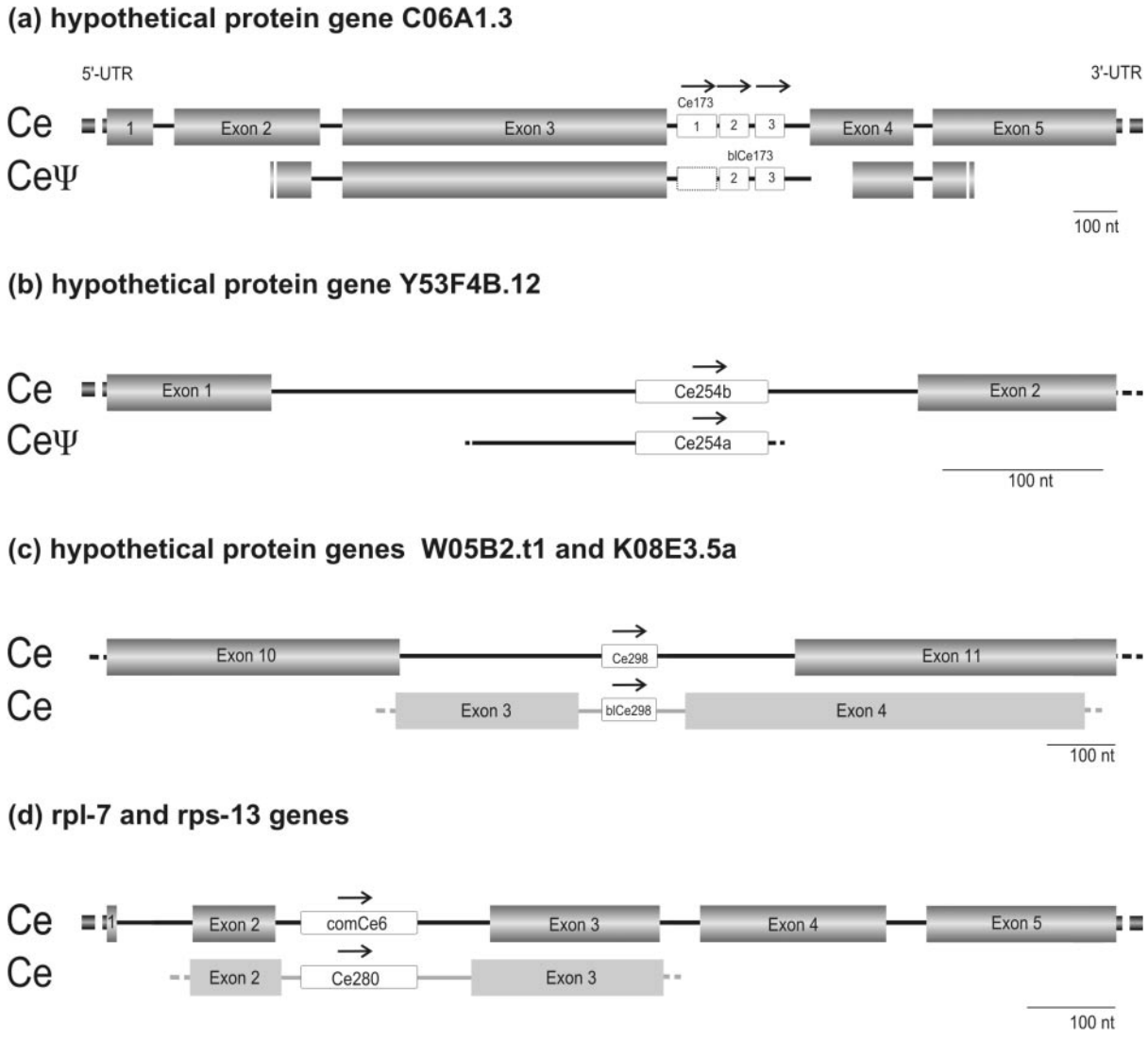

Figure 5. *Trans*-duplication of snoRNAs. (**a** and **b**) Represent *trans*-duplication of snoRNAs along with the flanking parts of their host genes. Note, that the outer flanks of the pseudogene sequences are highly diverged or deleted, and thus not alignable to the hypothetical protein genes C06A1.3 or Y53F4B.12. (a) Ce173-1 is diverged in the Ce pseudogene. (**c** and **d**) Represent *trans*-duplication of snoRNAs lacking their original flanking regions. Integration took place in new host genes (shown as light gray boxes and lines). Symbols are analogous to those in Figure 4.

Coevolution of snoRNA and rRNA modification target sites

Coevolution is defined as a change in the genetic composition of one species in response to a genetic change in another (29,30). This definition can be adapted to molecular interactions within organisms. Biologically significant interactions within macromolecules [e.g. RNA secondary structure; (31)] or between macromolecules, [e.g. RNA and proteins], can be demonstrated by compensatory changes in one or the other (32). Two of the C/D-box snoRNAs (Ce138, Ce234.2) exemplify adaptive evolution of the snoRNA complementary region to their 26S rRNA target sequence (Figure 6). In the lineage leading to *C.elegans*, an A→U substitution occurred in the 26S rRNA target site of the Ce138 snoRNA. This base change is not present in *C.briggsae* or *C.japonica* 26S rRNA sequences (data not shown). Accordingly, we found a compensatory U→A substitution in the antisense element of the snoRNA ortholog in *C.elegans* (Figure 6a), but not in *C.briggsae* or *C.japonica*. At another 26S rRNA position we found an A→G substitution in *C.briggsae* but not in *C.elegans* or *C.japonica* (data not shown). The corresponding *C.briggsae* snoRNA Cb234.2 shows a compensatory change from U→C (Figure 6b).

DISCUSSION

The combined impact of experimental and computational npcRNA screening

Our goal was to obtain as comprehensive a view as possible of cellular snoRNA expression in *C.elegans*. Creating a cDNA library based on size-fractionated, expressed RNAs,

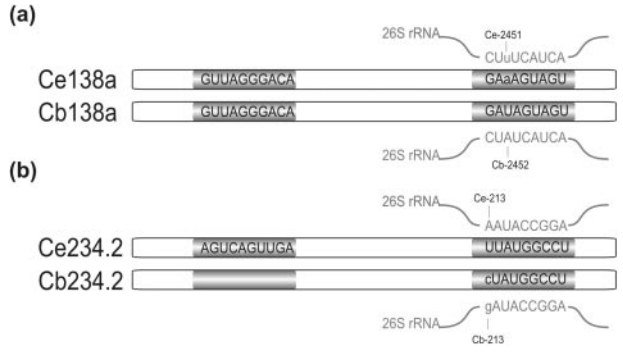

Figure 6. Coevolution of snoRNAs and their target sites. (**a**) The antisense region of snoRNA Ce138 differs by one nucleotide compared with Cb138. This change is a response to a base change in 26S rRNA of *C.elegans* (A→U) compared with 26S rRNA of *C.japonica* and *C.briggsae*. (**b**) A specific change in the 26S rRNA sequence of *C.briggsae* (A→G) compared with *C.japonica* and *C.elegans* is followed by a corresponding complementary change in snoRNA Cb234.2. Numbers above and below sequences denote the nucleotide positions in the 26S rRNAs.

and enriched to remove large numbers of known RNAs, yielded highly efficient experimental search results. We present here a detailed analysis of 120 different npcRNA species (Figure 1, groups I, II) from 665 informative sequences selected from an initial 38 400 clones. Moreover, to complement and validate the results of this experimental approach, we customized commercially available computer software to generate a search tool for identifying snoRNA candidates and their modification target sites according to a set of stringent criteria. From >100 potential intronic snoRNA candidates, 23 additional candidates fulfilled these conditions; another 8 npcRNAs were found by BLAST search. Thus, it is obvious that while computational approaches are not capable of supplanting experimental work, they do constitute a very useful complementation. This was particularly exemplified by our ability to analyze experimentally identified snoRNAs that, although apparently still functional, had diverged from the canonical motifs used for the computer search. In fact, the pitfalls of not complementing experimental results with such careful computational analyses can be clearly seen in a recent experimental screen (23). Of the *C.elegans* 56 novel snoRNAs shown in Figure 1, 14 were also reported recently but were analyzed either incorrectly or not at all (23). As examples, Ce96 (CeN25-2) or Ce135 (CeN25-1) were described as members of a novel class of small nuclear-like RNAs (23) (see their Figure 3D). Nevertheless, we could discern clear characteristics of C/D-box snoRNAs for both of these npcRNAs using computational analyses. Ce173.1-3 (CeN128) was described as one single H/ACA-box snoRNA species. By comparative analyses of *C.elegans* and *C.briggsae* we could distinguish them as three independent C/D-box snoRNAs. The same was true for Ce234.1-2 (CeN47) that they defined as one single snoRNA species. Ce110 (CeN42) is clearly a C/D-box snoRNA but they defined it as an H/ACA-box snoRNA even though part of the predicted H/ACA-box snoRNA would clearly overlap exonic sequences. They also identified six other unclassified npcRNAs [Ce86 (CeN35), Ce151 (CeN129), Ce254a (CeN23-1), Ce254b (CeN23-2), Ce282(CeN52), Ce105 (CeN66)] which we could clearly assign to specific snoRNA categories.

Target site plasticity

Our *in silico* target site complementarity search provided evidence of a high degree of plasticity in target site modification. In some cases we found evidence to suggest that the two complementary regions of particular snoRNAs modify targets in different compartments of the nucleoplasm, namely rRNAs in the nucleolus and spliceosomal RNAs in the Cajal body. In addition to the classical modification targets, we also found snoRNA complementarities for target sites in five different tRNAs. Modification of tRNAs by snoRNAs has been demonstrated so far only for Archaea and not for Eukarya (8,27). There is evidence that, following duplication, several snoRNA paralogs evolved new target site complementarities. Comparing *C.elegans* and *C.briggsae*, we observed that several specific modification sites of rRNAs are targeted by otherwise unrelated snoRNAs in both species. Losing, gaining or changing target sites are frequent phenomena that document the plasticity of modification interactions. Another source of plasticity is the compensatory changes of snoRNA target site complementary sequences that arose following base substitutions in their targets as illustrated in the case of Ce138 and Ce234.2 (Figure 6). Although several of our predicted target sites were confirmed by experimental approaches (26), a more conclusive verification of other target sites is necessary.

Birth and evolution of snoRNAs

Little is known about the origin and distribution of snoRNAs. Polycistronic clusters of snoRNAs are frequent in plants, and propagation due to cluster duplication is generated by polyploidization (33). However, polycistronic clusters of snoRNAs are the exception in vertebrates, as snoRNAs in those organisms are mainly singular and intron-encoded. To elucidate the process of snoRNA propagation in a 'model' eukaryote, we analyzed presence/absence patterns of snoRNA paralogs in *C.elegans* and compared them with those in *C.briggsae*. We identified three snoRNA paralogs with clear presence/absence patterns (Figure 4). These patterns suggest a copy/paste mechanism in the duplication of certain singular snoRNAs into neighboring introns of the same gene (*cis*-duplication). *Cis*-duplication seems to be a dominant process for H/ACA-box snoRNA propagation, but thus far, we did not identify any C/D-box snoRNA paralogs generated by *cis*-duplication. We found most genes harboring predominantly one type of intronic snoRNAs, (e.g. H/ACA-box snoRNAs in RPGs; Figure 1 and Supplementary Data), one notable exception being the C/D-box snoRNA comCe4 which is present in the midst of several other H/ACA-box snoRNAs in the hypothetical protein gene K07C5.4 (Figure 4).

Our data also suggest that snoRNAs can be propagated by complete or partial gene duplication that includes the embedded snoRNAs, an event that has been purported to precede evolutionary novelties (Figure 5) (34,35). Brosius (36) suggested that snoRNAs could be propagated by retroposition, a mechanism that might be responsible for *trans*-duplicated snoRNAs, but, because insertions of retroposed sequences are virtually random and should not lead to accumulations in neighboring introns, seems not to be involved in *cis*-duplication. Local, unequal recombination is a more

probable mechanism for *cis*-duplications, especially in *C.elegans*, because of the A/T-rich surroundings of snoRNA sequences.

In summary, the gain, loss and change of targets of snoRNAs over relatively short evolutionary times, possibly similar to the evolution of miRNAs (37–39), indicate that npcRNAs are not merely fossils from the long gone RNA/RNP world but continue to contribute to the changing needs of cells and genomes. This constitutes an astounding and unexpected level of plasticity for a primordial macromolecule such as RNA.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Yue Huang for implementing modifications of the DNAMAN software and Marsha Bundman for editorial assistance. This work was supported by the German Human Genome Project through the BMBF (#01KW9966), and grants from the Fonds der Chemischen Industrie from the European Union (EU; LSHG-CT-2003-503022) to J.B., and the Nationales Genomforschungsnetz (NGFN; 0313358A) to J.B. and J.S. Funding to pay the Open Access publication charges for this article was provided by NGFN.

Conflict of interest statement. None declared.

REFERENCES

1. Lander,E.S. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Stein,L., Bao,Z., Blasiar,D., Blumenthal,T., Brent,M.R., Chen,N., Chinwalla,A., Clarke,L., Clee,C., Coghlan,A. *et al*. (2003) The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol*., **1**, 166–192.
3. Lim,L.P., Lau,N.C., Weinstein,E.G., Abdelhakim,A., Yekta,S., Rhoades,M.W., Burge,C.B. and Bartel,D.P. (2003) The microRNAs of *Caenorhabditis elegans*. *Genes Dev*., **17**, 991–1008.
4. Kiss,A.M., Jady,B.E., Bertrand,E. and Kiss,T. (2004) Human box H/ACA pseudouridylation guide RNA machinery. *Mol. Cell. Biol*., **24**, 5797–5807.
5. Cavaillé,J., Buiting,K., Kiefmann,M., Lalande,M., Brannan,C.I., Horsthemke,B., Bachellerie,J.-P., Brosius,J. and Hüttenhofer,A. (2000) From the cover: identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *Proc. Natl Acad. Sci. USA*, **97**, 14311–14316.
6. Mattick,J.S. and Makunin,I.V. (2005) Small regulatory RNAs in mammals. *Hum. Mol. Genet*., **14**, R121–132.
7. Kishore,S. and Stamm,S. (2006) The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science*, **311**, 230–232.
8. Tang,T.-H., Bachellerie,J.-P., Rozhdestvensky,T., Bortolin,M.-L., Huber,H., Drungowski,M., Elge,T., Brosius,J. and Hüttenhofer,A. (2002) Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. *Proc. Natl Acad. Sci. USA*, **99**, 7536–7541.
9. Darzacq,X., Jady,B.E., Verheggen,C., Kiss,A.M., Bertrand,E. and Kiss,T. (2002) Cajal body-specific small nuclear RNAs: a novel class of 2′-*O*-methylation and pseudouridylation guide RNAs. *EMBO J*., **21**, 2746–2756.
10. Kiss,T. (2002) Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell*, **109**, 145–148.
11. Yoshihama,M., Uechi,T., Asakawa,S., Kawasaki,K., Kato,S., Higa,S., Maeda,N., Minoshima,S., Tanaka,T., Shimizu,N. *et al*. (2002) The human ribosomal protein genes: sequencing and comparative analysis of 73 genes. *Genome Res*., **12**, 379–390.
12. Tycowski,K.T., Shu,M.D. and Steitz,J.A. (1996) A mammalian gene with introns instead of exons generating stable RNA products. *Nature*, **379**, 464–466.
13. Pelczar,P. and Filipowicz,W. (1998) The host gene for intronic U17 small nucleolar RNAs in mammals has no protein-coding potential and is a member of the 5′-terminal oligopyrimidine gene family. *Mol. Cell. Biol*., **18**, 4509–4518.
14. Makarova,J.A. and Kramerov,D.A. (2005) Noncoding RNA of U87 host gene is associated with ribosomes and is relatively resistant to nonsense-mediated decay. *Gene*, **363**, 51–60.
15. Rodriguez,A., Griffiths-Jones,S., Ashurst,J.L. and Bradley,A. (2004) Identification of mammalian microRNA host genes and transcription units. *Genome Res*., **14**, 1902–1910.
16. Hüttenhofer,A., Kiefmann,M., Meier-Ewert,S., O'Brien,J., Lehrach,H., Bachellerie,J.-P. and Brosius,J. (2001) RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J*., **20**, 2943–2953.
17. Yuan,G., Klambt,C., Bachellerie,J.-P., Brosius,J. and Hüttenhofer,A. (2003) RNomics in *Drosophila melanogaster*: identification of 66 candidates for novel non-messenger RNAs. *Nucleic Acids Res*., **31**, 2495–2507.
18. Marker,C., Zemann,A., Terhorst,T., Kiefmann,M., Kastenmayer,J.P., Green,P., Bachellerie,J.-P., Brosius,J. and Hüttenhofer,A. (2002) Experimental RNomics: identification of 140 candidates for small non-messenger RNAs in the plant *Arabidopsis thaliana*. *Curr. Biol*., **12**, 2002–2013.
19. Lowe,T.M. and Eddy,S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, **283**, 1168–1171.
20. Schattner,P., Brooks,A.N. and Lowe,T.M. (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res*., **33**, 686–689.
21. Hofacker,I.L., Fekete,M. and Stadler,P.F. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol*., **319**, 1059–1066.
22. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*., **31**, 3406–3415.
23. Deng,W., Zhu,X., Skogerbo,G., Zhao,Y., Fu,Z., Wang,Y., He,H., Cai,L., Sun,H., Liu,C. *et al*. (2006) Organization of the *Caenorhabditis elegans* small non-coding transcriptome: genomic features, biogenesis, and expression. *Genome Res*., **16**, 20–29.
24. Wachi,M., Ogawa,T., Yokoyama,K., Hokii,Y., Shimoyama,M., Muto,A. and Ushida,C. (2004) Isolation of eight novel *Caenorhabditis elegans* small RNAs. *Gene*, **335**, 47–56.
25. Schattner,P., Decatur,W.A., Davis,C.A., Ares,M. Jr, Fournier,M.J. and Lowe,T.M. (2004) Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res*., **32**, 4281–4296.
26. Higa,S., Maeda,N., Kenmochi,N. and Tanaka,T. (2002) Location of 2′-*O*-methyl nucleotides in 26S rRNA and methylation guide snoRNAs in *Caenorhabditis elegans*. *Biochem. Biophys. Res. Commun*., **297**, 1344–1349.
27. Singh,S.K., Gurha,P., Tran,E.J., Maxwell,E.S. and Gupta,R. (2004) Sequential 2′-*O*-methylation of archaeal pre-tRNATrp nucleotides is guided by the intron-encoded but *trans*-acting box C/D ribonucleoprotein of pre-tRNA. *J. Biol. Chem*., **279**, 47661–47671.
28. Chen,C.-L., Liang,D., Zhou,H., Zhuo,M., Chen,Y.-Q. and Qu,L.-H. (2003) The high diversity of snoRNAs in plants: identification and comparative study of 120 snoRNA genes from *Oryza sativa*. *Nucleic Acids Res*., **31**, 2601–2613.
29. Ehrlich,P.R. and Raven,P.H. (1964) Butterflies and plants: a study in coevolution. *Evolution*, **18**, 586–608.
30. Page,R.D.M. and Holmes,E.C. (1998) *Molecular Evolution: A Phylogenetic Approach*. Blackwell Science Ltd, Oxford.
31. Woese,C., Magrum,L., Gupta,R., Siegel,R.B., Stahl,D.A., Kop,J., Crawford,N., Brosius,J., Gutell,R., Hogan,J.J. *et al*. (1980) Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. *Nucleic Acids Res*., **8**, 2275–2293.
32. Metzenberg,S., Joblet,C., Verspieren,P. and Agabian,N. (1993) Ribosomal protein L25 from *Trypanosoma brucei*: phylogeny and molecular co-evolution of an rRNA-binding protein and its rRNA binding site. *Nucleic Acids Res*., **21**, 4936–4940.

33. Brown,J.W., Clark,G.P., Leader,D.J., Simpson,C.G. and Lowe,T. (2001) Multiple snoRNA gene clusters from Arabidopsis. *RNA*, **7**, 1817–1832.
34. Bridges,C.B. (1935) Salivary chromosome maps with a key to the banding of the chromosomes of *Drosopila melanogaster*. *J. Hereditas*, **26**, 60–64.
35. Muller,H. (1936) Bar duplication. *Science*, **83**, 528–530.
36. Brosius,J. (2003) The contribution of RNAs and retroposition to evolutionary novelties. *Genetica*, **118**, 99–116.
37. Pasquinelli,A.E. and Ruvkun,G. (2002) Control of developmental timing by microRNAs and their targets. *Annu. Rev. Cell and Dev. Biol.*, **18**, 495–513.
38. Grun,D., Wang,Y.-L., Langenberger,D., Gunsalus,K.C. and Rajewsky,N. (2005) microRNA target predictions across seven *Drosophila* species and comparison to mammalian targets. *PLoS Comput. Biol.*, **1**, e13.
39. Houbaviy,H.B., Dennis,L., Jaenisch,R. and Sharp,P.A. (2005) Characterization of a highly variable eutherian microRNA gene. *RNA*, **11**, 1245–1257.