Nucleic Acids Research Advance Access published November 11, 2006

Nucleic Acids Research, 2006, Vol. 00, Database issue ***D1–D5***
doi:10.1093/nar/gkl873

Sno/scaRNAbase: a curated database for small nucleolar RNAs and cajal body-specific RNAs

Jun Xie*, Ming Zhang[1], Tao Zhou, Xia Hua, LiSha Tang and Weilin Wu[2,*]

Institute of Genetics, Fudan University, 220 Handan Road, 200433 Shanghai, China, [1]Deptartment of Bioinformatics, Institute of Microbiology and Genetics, University of Goettingen, 37077 Goettingen, Germany and [2]Center for Genomics and Bioinformatics, Karolinska Institute, Berzelius Väg 35, 17177 Stockholm, Sweden

Received August 15, 2006; Revised October 4, 2006; Accepted October 5, 2006

ABSTRACT

Small nucleolar RNAs (snoRNAs) and Cajal body-specific RNAs (scaRNAs) are named for their subcellular localization within nucleoli and Cajal bodies (conserved subnuclear organelles present in the nucleoplasm), respectively. They have been found to play important roles in rRNA, tRNA, snRNAs, and even mRNA modification and processing. All snoRNAs fall in two categories, box C/D snoRNAs and box H/ACA snoRNAs, according to their distinct sequence and secondary structure features. Box C/D snoRNAs and box H/ACA snoRNAs mainly function in guiding 2′-*O*-ribose methylation and pseudouridilation, respectively. ScaRNAs possess both box C/D snoRNA and box H/ACA snoRNA sequence motif features, but guide snRNA modifications that are transcribed by RNA polymerase II. Here we present a Web-based sno/scaRNA database, called sno/scaRNAbase, to facilitate the sno/scaRNA research in terms of providing a more comprehensive knowledge base. Covering 1979 records derived from 85 organisms for the first time, sno/scaRNAbase is not only dedicated to filling gaps between existing organism-specific sno/scaRNA databases that are focused on different sno/scaRNA aspects, but also provides sno/scaRNA scientists with an opportunity to adopt a unified nomenclature for sno/scaRNAs. Derived from a systematic literature curation and annotation effort, the sno/scaRNAbase provides an easy-to-use gateway to important sno/scaRNA features such as sequence motifs, possible functions, homologues, secondary structures, genomics organization, sno/scaRNA gene's chromosome location, and more. Approximate searches, in addition to accurate and straightforward searches, make the database search more flexible. A BLAST search engine is implemented to enable blast of query sequences against all sno/scaRNAbase sequences. Thus our sno/scaRNAbase serves as a more uniform and friendly platform for sno/scaRNA research. The database is free available at http://gene.fudan.sh.cn/snoRNAbase.nsf.

INTRODUCTION

Small nucleolar RNAs (snoRNAs) and small Cajal body-specific RNAs (scaRNAs) have been found to play vital roles in rRNA, tRNA, snRNA and even mRNA biogenesis. Since the 1990s, a vast collection of snoRNAs in eukaryotic cell have been found to be involved in rRNA methylation and pseudouridilation (1–3). Later in 2000, snoRNA homologues in archaea have been reported to function in tRNA modification (4). In humans, brain-specific snoRNAs are responsible for guiding modification of mRNAs (5). In 2001, a new type of modification guiding small RNAs, Cajal body-specific RNAs, was discovered and they guide the modification of snRNAs (6). Besides the functions in modification of different RNAs, a small number of snoRNAs, such as snoRNAs U3, U8, U14, E1, E2 and E3, are involved in the cleavage of pre-rRNAs (7,8).

Based on distinct sequence motifs and subcellular locations, sno/scaRNAs fall into three major groups: box C/D snoRNA, box H/ACA snoRNAs and scaRNAs (6,9–11). Box C/D snoRNAs share two short sequence motifs, box C (AUGAUGA) at the 5′ ends and box D (CUGA) at the 3′ ends, respectively. Two imperfect copies of these boxes, namely box C′ and box D′, have also been found in some box C/D snoRNAs. Immediately upstream of box D and/or D′ is a 10–21 nt antisense element complementary to targeted RNAs (10,12–14). Both the AUGAUGA and CUGA box motifs and the antisense element play essential roles in RNA methylation or processing (9). Each methylation site exclusively pairs with the fifth nucleotide upstream of box D or box D′ in the complementary region between a box C/D snoRNA and targeted RNA (15,16). Box H/ACA snoRNAs contain two conserved sequence motifs: a box H (ANANNA, where N stands for any nucleotide) and a

*To whom correspondence should be addressed. Tel: +86 21 5566 4556; Fax: +86 21 5566 4556; Email: xiejun@fudan.edu.cn
*Correspondence may also be addressed to Weilin Wu. Tel: +46 8 5248 7396; Fax: +46 8 323672; Email: Weilin.Wu@cgb.ki.se

box ACA (ACANNN), and two stem–loops near molecule 5′ and 3′ end, respectively. In the internal loop of the one or two stems is an appropriate bipartite guide sequence of 4–10 nt that forms a short snoRNA–rRNA duplex flanking the target site (10,17,18). The pseudouridylation site also obeys a spacing rule and it always appears at 14–16 nt upstream of box H or ACA within the bipartite guide sequence of a box H/ACA snoRNA (17,19). Different from the location of box C/D and box H/ACA snoRNAs in the nucleoli, scaRNAs accumulate within the Cajal bodies (conserved subnuclear organelles that are present in the nucleoplasm) (20). Moreover, a scaRNA molecule, such as U92, ACA47, ACA11, U109 and ACA57, can possess both box C/D and box H/ACA sequence motifs (e.g. U85), guiding both the methylation and pseudouridylation of snRNAs (6).

Sno/scaRNAs show high diversities in sequences, genomic organizations and processing pathways in varied organisms (8,12,21–25). A central and comprehensive knowledge base of sno/scaRNAs will undoubtedly speed up the current discovery process of sno/scaRNAs and deepen our understanding of their roles. Current databases exist of sno/scaRNAs (26–28), but their focuses on only one or two organisms featuring different sno/scaRNA characteristics made it very inconvenient in exploring sno/scaRNAs features/functions from the comparative genomics point of view. In this paper, we describe a more comprehensive, uniform, and curated sno/scaRNA database, sno/scaRNAbase. It contains 1979 sno/scaRNAs derived from 85 organisms and characterized in terms of sequence motifs, homologues, secondary structures, genomics organization, function that is experimentally verified or predicted, sno/scaRNA gene's chromosome location, and more. With its unified data form and a use-case-oriented user interface, sno/scaRNAbase allows users to browse and compare major features of known sno/scaRNAs from different organisms. It also provides the scientific community a platform to find a unified nomenclature for sno/scaRNAs that currently does not follow a general logic.

DATABASE CONTENT

Sno/scaRNAbase is a publicly available database of sno/scaRNAs obtained from 85 organisms that at least one sno/scaRNA has been reported. It has been developed using Lotus Domino Designer 6.0. One thousand nine hundred and seventy-nine sno/scaRNAs have been collected from various sources (literatures, GenBank searches, research group contacts, etc.). More than half of the data are not listed in the currently existing snoRNA databases. All data were further curated by trained biologists to ensure annotation quality. A sortable and searchable bibliography database of 1074 sno/scaRNA references (almost all sno/scaRNA reference publications) was extensively used during this curation

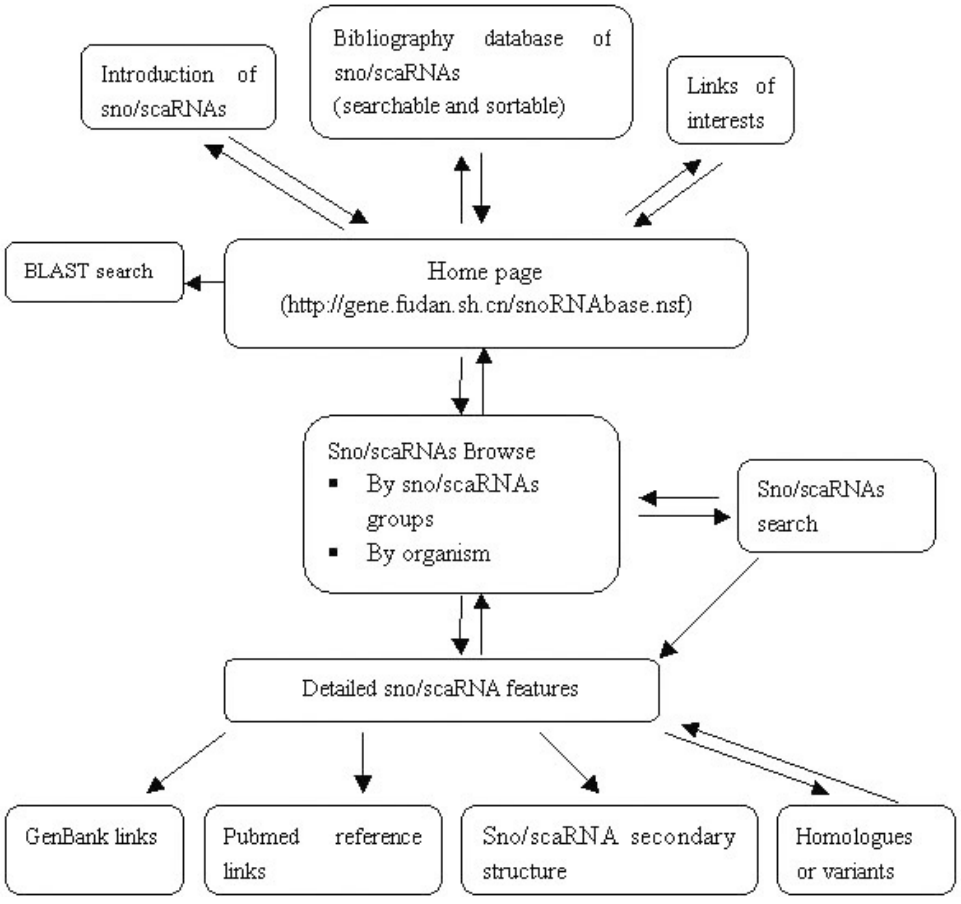

Figure 1. The schematic illustration of the sno/scaRNAbase.

process, and now becomes one part of the sno/scaRNAbase (Figure 1).

For each sno/scaRNA, we strive to extract as much sno/scaRNA information as possible from the sources mentioned above. Besides the regular sequence information (the sequence, the GenBank accession no., alias names, references, etc), the following important sno/scaRNA features have also been taken into account in the database design: (i) conserved sequence motifs and antisense elements of sno/scaRNA families, (ii) methylation or pseudouridylation sites that a sno/scaRNA guides, (iii) sno/scaRNA gene's chromosome location, (iv) genomic organization, (v) function that is experimentally verified or predicted (vi) other highly similar sequences in sno/scaRNAbase, and (vii) predicted secondary structure.

DATABASE OUTLINE

The sno/scaRNAbase browse/search page

A comprehensive interface was designed to explore and search different types of sno/scaRNAs (Figures 1 and 2). Each record is linked to a detailed sno/scaRNA feature page (see blow *the sno/scaRNA record page*). We provide the following sorting pages.

(i) *All by Organism* is an overview of all sno/scaRNAs grouped by organism. This view collapses in default according to organisms, and expands when users click a triangle next to an organism name. This allows users to see all available sno/scaRNAs in a certain organism.

(ii) *BoxC/D snoRNAs, box H/ACA snoRNAs* and *scaRNAs* are specific for three types of sno/scaRNAs. These browsing pages provide general information, such as sno/scaRNA name, references, GenBank entry, sequence length and the organism that a sno/scaRNA was isolated from. Detailed information, including possible functions, sequence motifs, and organization, is available by clicking the link associated with a sno/scaRNA name.

(iii) The *Search, Home, and Help* buttons link to search form, home page, and help page on this interface, respectively. The search page, as described below, uses either accurate or approximate searches to enable more flexible database search. This search helps identify inconsistency in the current nomenclature.

The sno/scaRNAbase search engine

Sno/scaRNAbase search is straightforward. A full-text search is capable of searching any fields of all sno/scaRNAbase records with user-defined keywords. For example, a full-text search of 'ctga' will return sno/scaRNAs with the 'ctga' in box D or D′ fields. It is necessary since different sno/scaRNAs demonstrate different features and these features sometimes are documented differently in references, thus it is unpractical to provide a specific search on all records fields through a uniform search form.

To enable the database search flexible, and to better track those sno/scaRNAs that are inconsistently documented in original publications, we provide three options for getting search results. One is using approximate searches that either consider a keyword as a root word and retrieve all sno/scaRNAs containing any words derived from this root, or take into account all words with spelling similar to the keyword and return any sno/scaRNAs containing these words. The former search, which is called a 'word variants search', is necessary because of the presence of multiple copies of sno/scaRNAs and the inconsistency of sno/scaRNA records. In this way, when searching a snoRNA name, different copies of the snoRNA usually distinguished by adding a

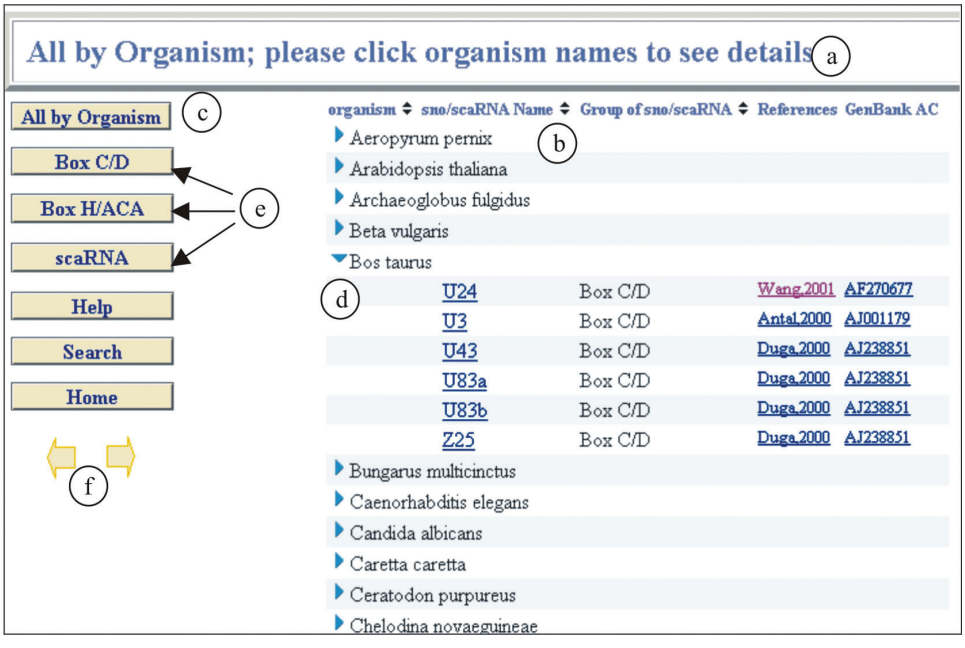

Figure 2. The sno/scaRNAbase browse page. **a**. Browsing result ordered by organism. **b**. Sorting buttons. **c**. Browsing selection I: browsing sno/scaRNAs by organism. **d**. A collapsing or expanding button. **e**. Browsing selection II: browsing sno/scaRNAs by three categories. **f**. Buttons for viewing previous or next pages.

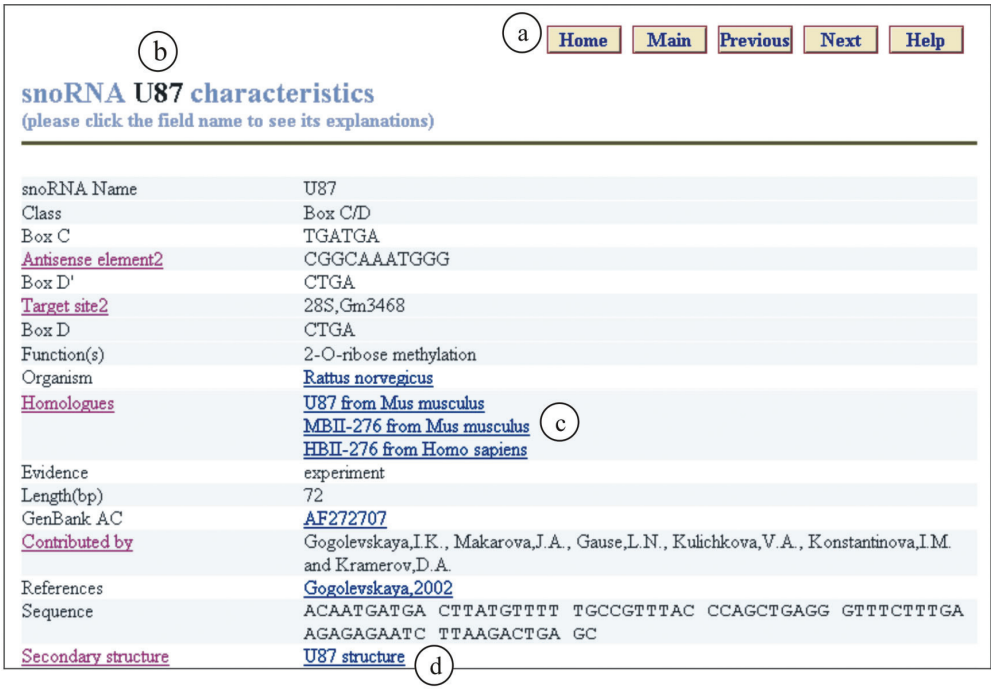

Figure 3. An example of a sno/scaRNA record page. **a**. Links to home/main pages, the previous/next sno/scaRNA record page, and the help page. **b**. A selected sno/scaRNA name. **c**. Possible homologues found in the sno/scaRNAbase. **d**. A link to the predicted secondary structure.

suffix to the end of a snoRNA will be returned. For example, entering 'U14' in the sno/scaRNA name field will return U14.1, U14.2, U14.3, and U14.4, etc. The latter search, which is defined as a 'fuzzy search', searches words that are spelled similar to a keyword. For instance, a full-text search of 'tgatga' in *Arabidopsis thaliana* with options of *using word variants* and *showing 100 as the maximum number records to return* returns 87 hits. While there are 98 hits if *fuzzy search* option is selected. Those 98 hits include snoRNAs with keyword 'tgatga', 'tgacga' (e.g. snoR4-2), 'tgatgg' (e.g. snoR101), and 'cgatga' (e.g. snoR27), etc. This is especially useful when searching sno/scaRNAs with a certain sequence motif. The other two options for controlling search results are: *Max Number of documents to return* and *Show results in order of relevance, newest first* (listing the latest record added in the sno/scaRNAbase first), or *oldest first* (listing the oldest record added in the sno/scaRNAbase first).

The search result page returns not only a list of sno/scaRNAs that are further linked to detailed sno/scaRNA record pages, but also the search string that was used, which is useful for users to refine searches.

The sno/scaRNA record page

An example of a sno/scaRNA record page is shown in Figure 3. Unless a record is not available, almost all sno/scaRNAs have the following information: *sno/scaRNA name*, *other name*, *class*, *nucleotide sequence, sequence length*, *GenBank accession number*, *Pubmed references*, *the organism that a sno/scaRNA was isolated from, possible homologues,* and *predicted secondary structure*. The secondary structures were calculated by RNAfold (29), which has been proved remarkably effective in predicting RNA structures (30). Regarding the possible homologues, they are determined by Blastn. Those hits with high similarity (currently *E*-value < 2e−0.5 and bit score >40 are used as thresholds), to a certain degree, indicate they were possibly derived from a common ancestor. To better understand their relationships, those highly similar sno/scaRNAs, together with the organisms they were isolated from, are summarized in order of descending similarity. In this way, users can analyze different copies of a sno/scaRNA in one organism and its homologues in other organisms. The record page is also linked to GenBank sequence records, Pubmed references, and the GenBank taxonomy site.

FUTURE DEVELOPMENTS

Sno/scaRNAbase is a periodically updated database dedicated to understanding sno/scaRNAs. More updated records, as well as more useful links (e.g. GeneCards and Genelinx), will be added to make the sno/scaRNAbase a more comprehensive knowledge base. In addition, we will merge duplicated entries reported from different sources, and plan to add experimentally verified sno/scaRNA secondary structure data. Further, we hope this database helps our explorations of sno/scaRNA functions and facilitates the genetic characterization of novel sno/scaRNAs, especially from the evolutionary point of view.

ACKNOWLEDGEMENTS

We thank Shanghai R&D Public Service Platform and Natural Science Foundations of China (No. 30300059, 30470356) for financial support. We also thank anonymous referees for their helpful comments and suggestions. Funding

to pay the Open Access publication charges for this article was provided by Natural Science Foundations of China (No. 30470356).

Conflict of interest statement. None declared.

REFERENCES

1. Bachellerie,J.P., Cavaille,J. and Huttenhofer,A. (2002) The expanding snoRNA world. *Biochimie*, **84**, 775–790.
2. Kiss,T. (2002) Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell*, **109**, 145–148.
3. Lafontaine,D.L., Bousquet-Antonelli,C., Henry,Y., Caizergues-Ferrer,M. and Tollervey,D. (1998) The box H + ACA snoRNAs carry Cbf5p, the putative rRNA pseudouridine synthase. *Genes Dev*., **12**, 527–537.
4. Clouet d'Orval,B., Bortolin,M.L., Gaspin,C. and Bachellerie,J.P. (2001) Box C/D RNA guides for the ribose methylation of archaeal tRNAs. The tRNATrp intron guides the formation of two ribose-methylated nucleosides in the mature tRNATrp. *Nucleic Acids Res*., **29**, 4518–4529.
5. Cavaille,J., Buiting,K., Kiefmann,M., Lalande,M., Brannan,C.I., Horsthemke,B., Bachellerie,J.P., Brosius,J. and Huttenhofer,A. (2000) Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *Proc. Natl Acad. Sci. USA*, **97**, 14311–14316.
6. Jady,B.E. and Kiss,T. (2001) A small nucleolar guide RNA functions both in 2′-O-ribose methylation and pseudouridylation of the U5 spliceosomal RNA. *EMBO J*., **20**, 541–551.
7. Venema,J., Vos,H.R., Faber,A.W., van Venrooij,W.J. and Raue,H.A. (2000) Yeast Rrp9p is an evolutionarily conserved U3 snoRNP protein essential for early pre-rRNA processing cleavages and requires box C for its association. *RNA*, **6**, 1660–1671.
8. Grandi,P., Rybin,V., Bassler,J., Petfalski,E., Strauss,D., Marzioch,M., Schafer,T., Kuster,B., Tschochner,H., Tollervey,D. *et al*. (2002) 90S pre-ribosomes include the 35S pre-rRNA, the U3 snoRNP, and 40S subunit processing factors but predominantly lack 60S synthesis factors. *Mol. Cell*, **10**, 105–115.
9. Weinstein,L.B. and Steitz,J.A. (1999) Guided tours: from precursor snoRNA to functional snoRNP. *Curr. Opin. Cell Biol*., **11**, 378–384.
10. Balakin,A.G., Smith,L. and Fournier,M.J. (1996) The RNA world of the nucleolus: two major families of small RNAs defined by different box elements with related functions. *Cell*, **86**, 823–834.
11. 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.
12. Bachellerie,J.P., Caffarelli,J. and Qu,L.H. (2000) Nucleotide modifications of eukaryotic rRNAs: the world of small nucleolar RNA guides revisited. *The Ribosome: Structure, Function, Antibiotics and Cellular Interactions*. ASM Press, Washington, DC, pp. 191–203.
13. Bachellerie,J.P. and Cavaille,J. (1997) Guiding ribose methylation of rRNA. *Trends Biochem. Sci*., **22**, 257–261.
14. Maxwell,E.S. and Fournier,M.J. (1995) The small nucleolar RNAs. *Annu. Rev. Biochem*., **64**, 897–934.
15. Kiss-Laszlo,Z., Henry,Y., Bachellerie,J.P., Caizergues-Ferrer,M. and Kiss,T. (1996) Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. *Cell*, **85**, 1077–1088.
16. Cavaille,J., Nicoloso,M. and Bachellerie,J.P. (1996) Targeted ribose methylation of RNA *in vivo* directed by tailored antisense RNA guides. *Nature*, **383**, 732–735.
17. Bortolin,M.L., Ganot,P. and Kiss,T. (1999) Elements essential for accumulation and function of small nucleolar RNAs directing site-specific pseudouridylation of ribosomal RNAs. *EMBO J*., **18**, 457–469.
18. Ganot,P., Caizergues-Ferrer,M. and Kiss,T. (1997) The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. *Genes Dev*., **11**, 941–956.
19. Ganot,P., Bortolin,M.L. and Kiss,T. (1997) Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. *Cell*, **89**, 799–809.
20. Matera,A.G. (1998) Of coiled bodies, gems, and salmon. *J. Cell Biochem*., **70**, 181–192.
21. Brown,J.W., Echeverria,M. and Qu,L.H. (2003) Plant snoRNAs: functional evolution and new modes of gene expression. *Trends Plant Sci*., **8**, 42–49.
22. Dheur,S., Vo le,T.A., Voisinet-Hakil,F., Minet,M., Schmitter,J.M., Lacroute,F., Wyers,F. and Minvielle-Sebastia,L. (2003) Pti1p and Ref2p found in association with the mRNA 3′ end formation complex direct snoRNA maturation. *EMBO J*., **22**, 2831–2840.
23. Rebane,A., Tamme,R., Laan,M., Pata,I. and Metspalu,A. (1998) A novel snoRNA (U73) is encoded within the introns of the human and mouse ribosomal protein S3a genes. *Gene*, **210**, 255–263.
24. Runte,M., Huttenhofer,A., Gross,S., Kiefmann,M., Horsthemke,B. and Buiting,K. (2001) The IC-SNURF-SNRPN transcript serves as a host for multiple small nucleolar RNA species and as an antisense RNA for UBE3A. *Hum. Mol. Genetics*, **10**, 2687–2700.
25. Omer,A.D., Lowe,T.M., Russell,A.G., Ebhardt,H., Eddy,S.R. and Dennis,P.P. (2000) Homologs of small nucleolar RNAs in Archaea. *Science*, **288**, 517–522.
26. Samarsky,D.A. and Fournier,M.J. (1999) A comprehensive database for the small nucleolar RNAs from *Saccharomyces cerevisiae*. *Nucleic Acids Res*., **27**, 161–164.
27. Lestrade,L. and Weber,M.J. (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res*., **34**, D158–D162.
28. Brown,J.W., Echeverria,M., Qu,L.H., Lowe,T.M., Bachellerie,J.P., Huttenhofer,A., Kastenmayer,J.P., Green,P.J., Shaw,P. and Marshall,D.F. (2003) Plant snoRNA database. *Nucleic Acids Res*., **31**, 432–435.
29. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte Fur. Chemie*., **125**, 167–188.
30. Zuker,M. (2000) Calculating nucleic acid secondary structure. *Curr. Opin. Struct. Biol*., **10**, 303–310.