5112–5123 *Nucleic Acids Research, 2006, Vol. 34, No. 18* *Published online 20 September 2006*
doi:10.1093/nar/gkl672

snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome

Jian-Hua Yang, Xiao-Chen Zhang, Zhan-Peng Huang, Hui Zhou, Mian-Bo Huang[1], Shu Zhang[1], Yue-Qin Chen* and Liang-Hu Qu[1,*]

Key Laboratory of Gene Engineering of the Ministry of Education and [1]State Key Laboratory for Biocontrol, Zhongshan University, Guangzhou 510275, PR China

Received July 6, 2006; Revised and Accepted August 28, 2006

ABSTRACT

Small nucleolar RNAs (snoRNAs) represent an abundant group of non-coding RNAs in eukaryotes. They can be divided into guide and orphan snoRNAs according to the presence or absence of antisense sequence to rRNAs or snRNAs. Current snoRNA-searching programs, which are essentially based on sequence complementarity to rRNAs or snRNAs, exist only for the screening of guide snoRNAs. In this study, we have developed an advanced computational package, snoSeeker, which includes CDseeker and ACAseeker programs, for the highly efficient and specific screening of both guide and orphan snoRNA genes in mammalian genomes. By using these programs, we have systematically scanned four human–mammal whole-genome alignment (WGA) sequences and identified 54 novel candidates including 26 orphan candidates as well as 266 known snoRNA genes. Eighteen novel snoRNAs were further experimentally confirmed with four snoRNAs exhibiting a tissue-specific or restricted expression pattern. The results of this study provide the most comprehensive listing of two families of snoRNA genes in the human genome till date.

INTRODUCTION

The small nucleolar RNAs (snoRNAs) represent an abundant group of small non-coding RNAs (ncRNAs) in eukaryotes (1). With the exception of RNase MRP, all the snoRNAs fall into two major families, box C/D and box H/ACA snoRNAs, on the basis of common sequence motifs and structural features. A large number of snoRNAs characterized to date are box C/D snoRNAs that share two conserved motifs, the 5′ end box C and the 3′ end box D, and the box H/ACA snoRNAs that exhibit a common hairpin–hinge–hairpin–tail secondary structure with the box H and ACA (2). Although several snoRNAs, such as U3, snR30 and RNase MRP, are required for specific cleavage of pre-rRNAs, the majority of box C/D snoRNAs function as guides for site-specific 2′-*O*-ribose methylation, and most box H/ACA snoRNAs function as guides for pseudouridylation in the post-transcriptional processing of rRNAs (2). Studies have shown that some snoRNAs and Small Cajal body-specific RNAs (scaRNAs) participate in the modification of snRNAs (3,4). Some modifications in Archaea tRNAs are also introduced by box C/D small RNAs, which are the homologs of snoRNAs in eukaryotes (5). Notably, an increasing number of orphan snoRNAs, which lack antisense to rRNAs or snRNAs, has been experimentally identified along with the numerous guide snoRNAs from different eukaryotes (6,7). The finding of an orphan snoRNA, HBII-52, being associated with human disease, has triggered a great interest in orphan snoRNAs (8,9).

Many studies have proven that computational analysis of genomic databases is a useful way to identify snoRNAs from eukaryotes on a large-scale (7,10,11). To date, several searching programs based on pattern recognition scan algorithms have been developed, such as Snoscan (10) for box C/D snoRNA, and SnoGPS (11) and MFE (12) methods for box H/ACA snoRNA. In comparison to experimental approaches that tend to favor the discovery of the most abundant RNAs (13,14), computational analysis provides an unbiased genome-wide search for snoRNA genes. However, the current snoRNA-searching programs are essentially based on sequence complementarity to rRNAs or snRNAs and are therefore limited to the detection of guide snoRNAs and not orphan snoRNAs. Another limitation with the existing programs is that it is difficult to systematically search the

*To whom correspondence should be addressed at Biotechnology Research Center, Zhongshan University, Guangzhou 510275, PR China. Tel: +86 20 84112399; Fax: +86 20 84036551; Email: lsbrc16@zsu.edu.cn
*Correspondence may also be addressed to Liang-Hu Qu. Tel/Fax: +86 20 84112399; Email: lsbrc04@zsu.edu.cn

human genome for snoRNAs because of the vast data source. Although two recent studies have performed a computational detection of human snoRNAs, they only focused on particular segments of the human genome (15,16). It is therefore important to develop an advanced search program for genome-wide screening of all snoRNAs, including the orphan snoRNAs.

In this study, we developed a computational package which includes two novel snoRNA-searching programs, CDseeker and ACAseeker, specific to the detection of C/D snoRNAs and H/ACA snoRNAs, respectively. Based on new algorithms, our programs detected both guide and orphan snoRNA genes in a genome-wide analysis. We also systematically searched the human genome for snoRNAs with four human–mammal (human/mouse, human/rat, human/dog and human/cow) whole-genome alignment (WGA) sequences using these programs. As a result, a majority of known C/D and H/ACA snoRNA genes were detected. In addition, 54 novel genes were identified with 18 being experimentally confirmed. This study presents the most complete list of two families of snoRNA genes in the human genome till date.

MATERIALS AND METHODS

Data source

WGAs for human (May 2004), mouse (March 2005), rat (June 2003), dog (July 2004), cow (September 2004), chimp (November 2003) and monkey (January 2005) were downloaded from the UCSC Genome Bioinformatics site (http://genome.ucsc.edu). The repeat families were removed by RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html). Sequences and annotation data for known human snoRNA genes (which were used in program training) were downloaded from snoRNA-LBME-db (17) on August 2005. UCSC KnownGene, RefGene and AceView annotation data for human protein genes and transcript units were downloaded from the UCSC Genome Bioinformatics site (http://genome.ucsc.edu).

rRNA and snRNA sequences were downloaded from snoRNA-LBME-db (17). 2′-O-methylation and pseudouridylation sites of human rRNA and snRNA were cited from snoRNA-LBME-db (17) and other reports (18,19). Sequences 4 nt upstream to 25 nt downstream of known methylation sites were extracted from the rRNA and snRNA sequences as target sequences for the CDseeker program. Sequences 15 nt upstream to 15 nt downstream of the known pseudouridylation sites were extracted from the rRNA and snRNA sequences as target sequences for the ACAseeker program.

Algorithm description

The CDseeker program combines probabilistic model (20), conserved primary and secondary structure motifs to search orphan and guide C/D snoRNAs in WGA sequences. Common algorithm components for guide and orphan snoRNA genes are box C, box D and the terminal stem base pairing. Two additional components are applicable to guide snoRNA genes, a region of sequence complementary to rRNA or snRNA and box D′ if the rRNA or snRNA complementary region is not directly adjacent to box D. For orphan snoRNA genes, two additional components are (i) predicted conserved functional region next to box D or box D′ and (ii) box D′ if the conserved functional sequence is not next to box D. The distance between components (e.g. the maximum distance is 100 nt between box C and box D) was also taken into account. The program searches box C, box D, terminal stem pairing and the antisense element step-by-step in WGA consensus sequences, and scores the corresponding elements with probabilistic models, then evaluates them based on the standard cutoff score of the training set. Candidates progress to the next evaluation only if the element score is higher than the cutoff. The examination of antisense is an optional criterion in the CDseeker program. The program assigns the candidates as guide snoRNAs or orphan snoRNAs using this evaluation. Finally, in order to rank the candidates, the program sums the scores of the motifs resulting in a final score. The standard cutoff score of a training set is also applied for selecting candidates. The structural model of C/D snoRNA genes for CDseeker is based on the canonical C/D structure shown in Figure 1A and the core algorithm workflow is shown using a schematic diagram in Figure 1B.

The ACAseeker program combines probabilistic model (20), conserved primary and secondary structure motifs to search orphan and guide H/ACA snoRNAs in WGA sequences. Common algorithm components for guide and orphan snoRNA genes are box H, box ACA, stem1 (hairpin I), stem2 (hairpin II) and hairpin–hinge–hairpin. For guide snoRNA genes, another component is taken into account which is the two regions of sequence complementary to rRNA or snRNA in a hairpin. The program searches box H and ACA in conserved sequences of WGA consensus sequences and scores them using probabilistic models. The candidates having a score higher than the standard cutoff progress to the next step which is an evaluation of secondary structure feature using a slightly arbitrary standard observed from known H/ACA snoRNAs. Similar to CDseeker, the final examination of antisense by ACAseeker program is an optional criterion. The program assigns the candidates as guide snoRNAs or orphan snoRNAs by this evaluation. The structural model of H/ACA snoRNA genes used by ACAseeker is based on the canonical H/ACA structure shown in Figure 1C and the core algorithm workflow is shown using a schematic diagram in Figure 1D.

Generating consensus sequences for searching. We generated consensus sequences from UCSC whole-genome pairwise alignments. The following annotations were assigned: (i) the same letter if the sequences were identical between two alignment sequences; (ii) a dot (.) if there was a point mutation between two alignment sequences; and (iii) a dash (-) if there was an insertion or deletion (indel) between alignment sequences. The two alignment sequences were transformed as a consensus sequence (Supplementary Figure S1). To efficiently scan the snoRNA genes, consensus motifs in consensus sequences were searched. From alignments of known human–mammalian alignment snoRNA sequences, we found that the regular expression for box C, box D, box H and box ACA are as follows: (i) Box C, [ACGT.][ATG]GA[TG]G[ATG.]; (ii) Box D, CTGA; (iii) Box H, A[ACGT.]A[ACGT.][ACGT.][G.]?A; and (iv) Box ACA, A[CT.]A. Dot (.) indicates a point mutation between human

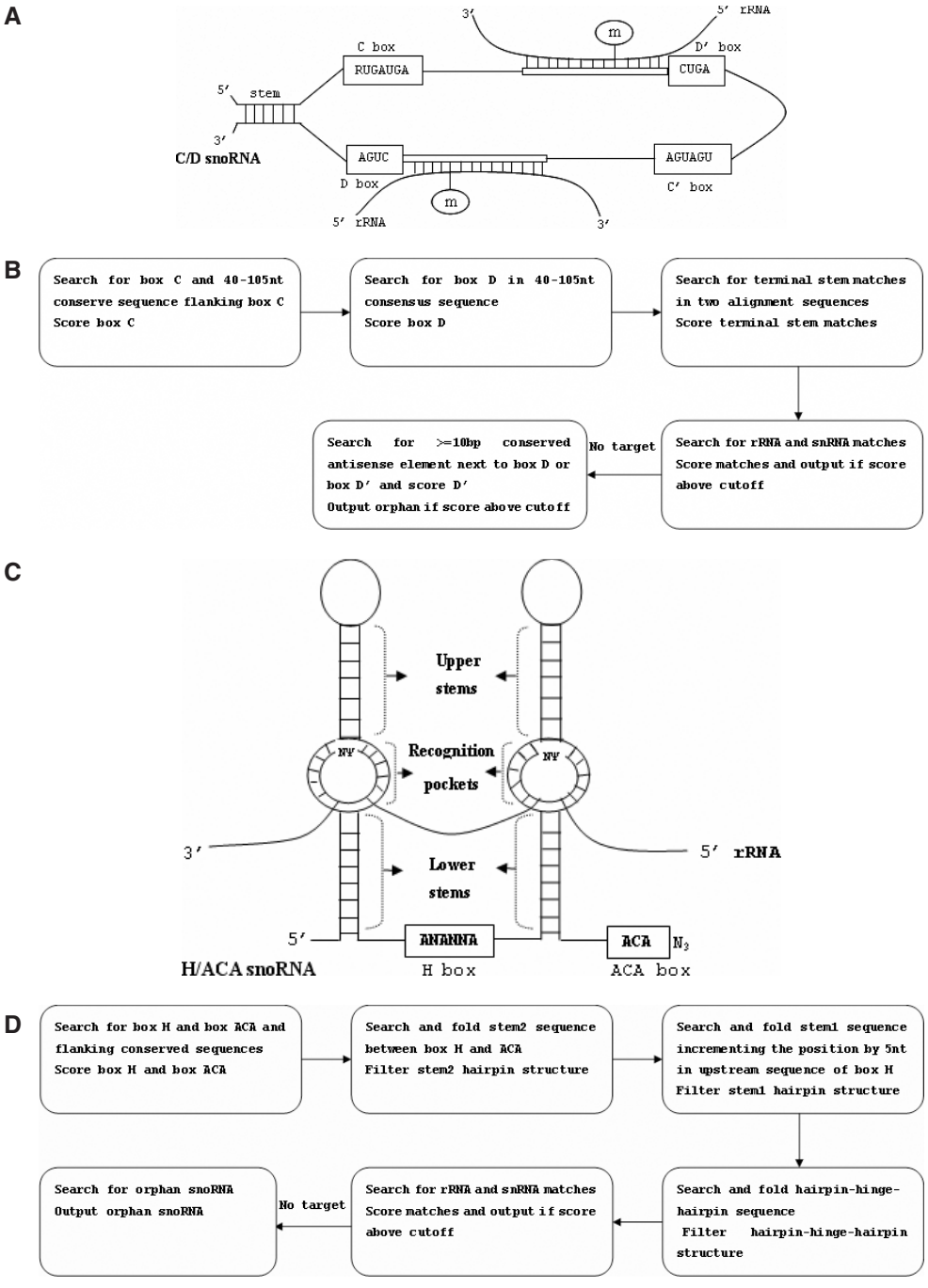

Figure 1. CDseeker and ACAseeker core algorithm workflow. (**A**) The C/D snoRNA model. The C/D box snoRNAs carry the conserved boxes C (RUGAUGA, R = purine) and D (CUGA) near their 5′ and 3′ ends, respectively. The two boxes are frequently folded together by a short (4–5 bp) terminal helix, to form a structure similar to a kink-turn. Often, imperfect copies of the C and D boxes, named C′ and D′, are located internally, in the order C/D′/C′/D. The 2′-O-ribose methylation of target RNAs is guided by one or two 10–21 antisense elements located upstream of the D and/or D′ boxes, so that the modified base is paired with the snoRNA nucleotide located precisely 5 nt upstream of the D or D′ box (17). (**B**) Schematic representation of the CDseeker algorithms. (**C**) The H/ACA snoRNA model. The H/ACA box snoRNAs consist of two hairpins and two short single-stranded regions, which contain the H box (ANANNA) and the ACA box. The latter is always located 3 nt upstream of the 3′ end of the snoRNA. The hairpins contain bulges, or recognition loops that form complex pseudoknots with the target RNA, where the target uridine is the first unpaired base. The position of the substrate uridine always resides 13–16 nt upstream of the H box (left recognition pocket) or of the ACA box (right recognition pocket) (17). (**D**) Schematic representation of the ACAseeker algorithms.

and other mammalians, letters are identical nucleotides between human and other mammalians, the symbol '?' indicates that the former symbol appear 0 or 1 time, and '[]' means that one of the letters located within the brackets appear only once in regular expression.

Indel and substitution models. As H/ACA snoRNAs are structural and functional RNAs, their sequences are constrained for maintaining the hairpin–helix–hairpin–tail secondary structure. We surveyed 100 known H/ACA snoRNAs and found that successive mutations in these RNAs were mostly

<7 nt and successive indels were mostly <5 nt in the alignment sequences (Supplementary Figure S2). We, therefore, defined maximum successive mutations as 7 nt and maximum successive indels as 5 nt. We applied these models to extract conserved segments from whole-genome pairwise alignments and found all known box H/ACA snoRNAs located in 120–1000 nt conserved segments (Supplementary Figure S3). The ACAseeker program was then applied to search for H/ACA genes whose lengths varied between 120 and 1000 nt in the conserved sequences of WGA.

Training for scoring standard with probabilistic model. (i) *Training for scoring standard of box element with hidden Markov model*. The hidden Markov model (HMM) has been widely used for searching protein-coding genes and non-coding RNA genes (10,20–22). The fixed-length HMM (or zero-order HMM) was used in training the box elements (including box C, box D, box D′, box H and box ACA) of known snoRNAs. In detail, box elements are trained by calculating the probability of test sequences with the HMM. According to the alignment sequences, the box H being trained, is ANANNA and ANANNGA, so there are two transition probabilities from position 5 to 6 of 0.90 and 0.10, respectively (the logarithm of the probability is −0.152 and −3.322, respectively; Supplementary Figure S6). The other transition probability is 1 (the logarithm of the probability is 0; Supplementary Figure S6). The emission probability is constructed by the following equation:

$$E_{ij} = \log_2\left(\frac{f(S_{ij}) + p}{f(S_{bj})}\right).$$

The E_{ij} is an emission score for the j-th symbol (base) of the i-th position of sequence S. $f(S_{ij})$ is a frequency for j-th symbol (base) of the i-th position of sequence S. p is a pseudo-count. $f(S_{bj})$ is a background frequency of the j-th symbol. Consequently, the HMM consists of an $n \times m$ matrix when the length of sequence S is n and the number of symbols is m.

(ii) *Training for scoring standard of terminal stem pairing with stochastic context-free grammars (SCFGs)*. Secondary structures in RNA are not local, similar to proteins; thus, it is necessary to use a more complex model than an HMM for modeling terminal stem pairing. Stochastic context-free grammars (SCFGs) (10,20), which are used here, can describe some long-range interactions, including most of those in RNA secondary structure. The SCFG production probabilities were estimated from a training set of C/D snoRNAs with terminal stem pairing. After surveying the terminal stem pairings of all known C/D snoRNA, we found that the pairs reached a maximum of 15 bp in the terminal stem. Hence the 15 nt flanking sequences of known snoRNA were extracted for folding the terminal stem and training. The score for terminal stem pairing is from 1.149 bits to 21.252 bits.

(iii) *Training for scoring standard of complementary regions with HMM*. The HMM model was also used in training the complementary regions between ribosomal RNA and snoRNAs (10,20). The probability of matches (Watson–Crick matches and GU matches) and mismatches in the complementarity is emission probability. After surveying the antisense–target–helix of all known snoRNA, we defined criteria for complementary region evaluation. For box C/D snoRNA genes, the criteria are as follows: (a) we selected the lowest score (13.0 bits) of the complementary region as a cutoff score; (b) the first mismatch next to box D or box D′ did not affect the stability of the complementary helix (the antisense and target; 0 bit was given for the first mismatch next to box D or D′); (c) the maximum mismatch was 1 nt except for the first mismatch; (d) the maximum GU pairs were 3 bp; and (e) the minimum complementary length was 10 nt. For box H/ACA snoRNA genes, the criteria are as follows: (a) we selected the lowest score (16.2 bits) of the complementary region as a cutoff score; (b) the maximum mismatch was 1 nt; (c) the minimum complementary length was 9 nt; (d) the minimum upper stem pairing was 4 bp and maximum bulge in upper stem pairing was 5 nt; and (e) the minimum lower stem pairing was 4 bp and maximum bulge in lower stem pairing was 2 nt.

Evaluating box H/ACA secondary structure. For evaluating the hairpin structure, we folded the hairpin II and hairpin I step-by-step using RNAfold (23) and then evaluated the structure with the following standards concluded from surveying all the known H/ACA snoRNAs: (i) maximum distance between H box and hairpin I was 5 nt; (ii) maximum distance between H box and hairpin II was 7 nt; (iii) maximum distance between ACA box and hairpin II was 5 nt; (iv) maximum pocket was 12 nt; (v) maximum bulge in lower and upper stems were 2 and 5 nt, respectively; (vi) minimum stem pairs were 14 bp; (vii) minimum and maximum loop sizes were 3 and 17 nt, respectively; (viii) maximum tail length was 4 nt; and (ix) hairpin MFE was less than −11.0 (kcal/mol) and larger than −43.0 (kcal/mol). For evaluating the hairpin–hinge–hairpin structure, we folded the whole candidate using MFOLD (24), and evaluated the candidate with the following standards concluded from surveying all the known H/ACA snoRNAs: (i) maximum hinge size was 15 nt and (ii) hairpin–hinge–hairpin MFE was less than −29.0 (kcal/mol).

Selecting candidates using locateGenome program. With the exception of U3, mgU2-25/61, mgU12-22/U4-8, mgU12 and U13, all snoRNAs are located in introns of spliced mRNAs. We investigated the size distribution of all known snoRNA-hosted introns and found their lengths mainly fell into the range of 100–600 nt (Supplementary Figure S4). The length of introns containing C/D RNAs varied from 157 to 44 874 nt (Supplementary Figure S4). The length of introns containing H/ACA RNAs varied from 213 to 103 658 nt (Supplementary Figure S4).

The locateGenome program first locates the candidates according to the UCSC AceView gene data of human genomic coordinates and orientations which includes >250 000 alt-splicings. The program then selects C/D candidates which are located within introns <50 000 bp and H/ACA candidates which are located within introns <110 000 bp.

Selecting candidates with conservation filter. We surveyed all the training snoRNA genes and found snoRNA sequences were more conserved than their flanking sequences (Figure 2A and B). We then evaluated candidates through the conservation percentage of flanking upstream and

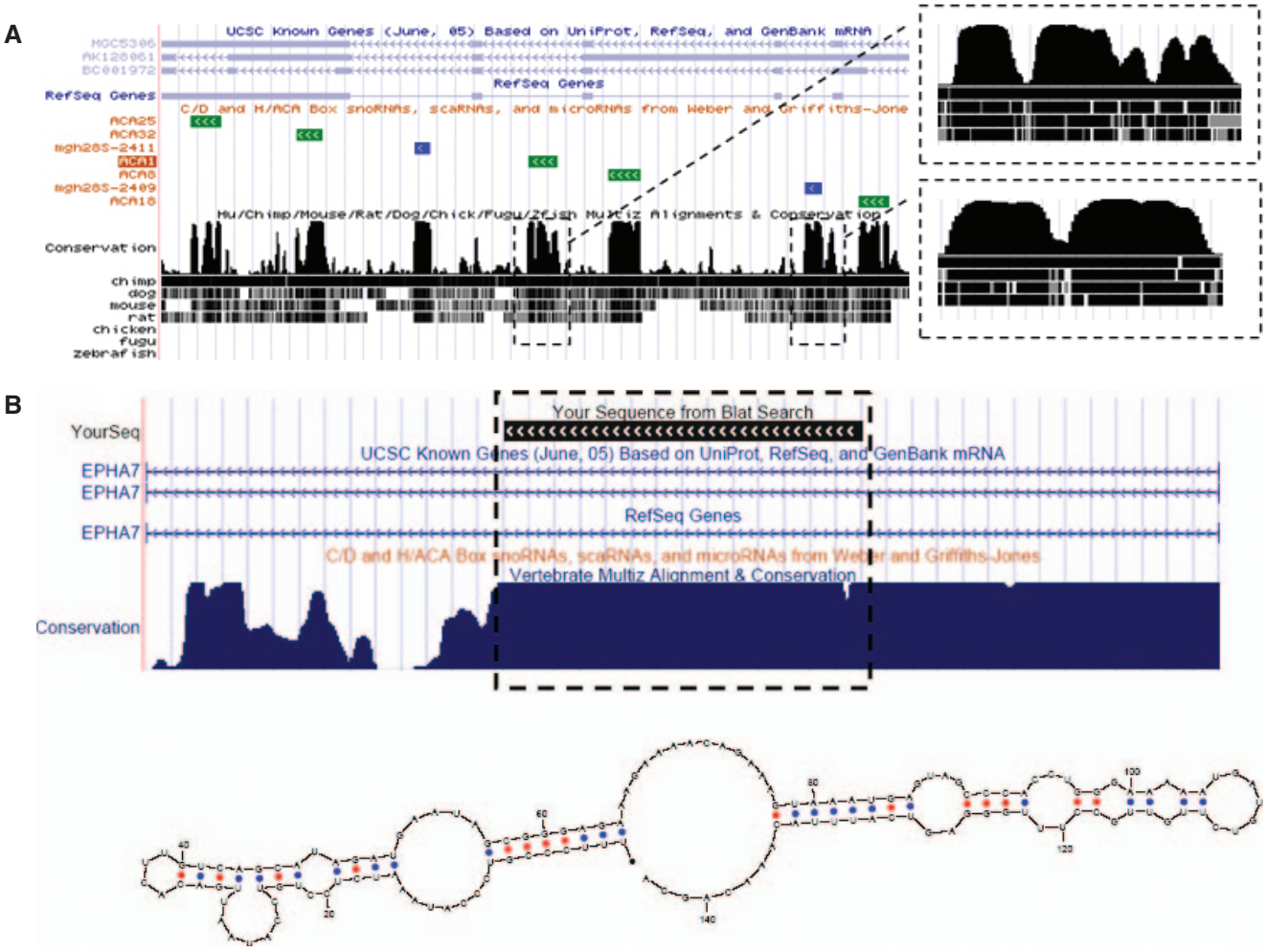

Figure 2. (**A**) SnoRNA conserved features on the human UCSC Genome Browser. C/D and H/ACA snoRNAs are colored blue and green, respectively. Conservation track reveals that sequences corresponding to snoRNAs are more highly conserved than those of flanking sequences. (**B**) A candidate box H/ACA RNA by computational method does not fit the conserved pattern. UCSC conservation track reveals that sequences corresponding to candidate box H/ACA RNA are less highly conserved than those of flanking sequences. Although this candidate can fold into a hairpin–hinge–hairpin–tail structure, its expression cannot be confirmed by northern blot and reverse transcription.

downstream 50 bp sequences and the UCSC genome browser high conserved track (25). We evaluated candidates by the following standards concluded from surveying all known snoRNAs:

$$\text{For box H/ACA}: (\text{IDhp1} - \text{IDup}) + (\text{IDhp2} - \text{IDdown}) > 19\%;$$

$$\text{For box C/D}: (\text{IDcd} - \text{IDup}) + (\text{IDcd} - \text{IDdown}) > 9\%.$$

The IDhp1 and IDhp2 are the conservation percentages of hairpins 1 and 2, IDup and IDdown are the conservation percentages of flanking upstream and downstream 50 bp sequences.

RNA isolation and northern blot

Total cellular RNA was isolated from adult male rat tissues and purified according to the method of guanidine thiocyanate/phenol–chloroform (26). Small RNA was purified from total RNA according to the PEG8000/NaCl protocol (27). Small RNA (25 μg/lane) from brain, thymus, heart, lung, liver, spleen, kidney and testis was analyzed by electrophoresis on 8% acrylamide/7 mol/l urea gels. Total RNAs were transferred on to nylon membranes (Hybond-N^+; Amersham) followed by UV irradiation for 5 min. Hybridization with 5′-labeled probes was performed as described previously (28).

Oligodeoxynucleotides

Oligonucleotides were synthesized and purified by Sangon Co. (Shanghai, China). The sequences of oligonucleotide probes for northern blotting and oligonucleotide primers for cDNA libraries construction and screening are shown in Supplementary Table S1. The probes used in northern blotting were 5′-end-labeled with [γ-^{32}P]ATP (Yahui Co.) and submitted to purification according to standard laboratory protocols as described previously (26).

RESULTS

The strategy and efficiency of the CDseeker program for screening of mammalian C/D snoRNA genes

In this study, we developed a computational program, CDseeker, specific to the screening of box C/D snoRNAs that are conserved in human–mammal genomes (human/mouse, human/rat, human/dog and human/cow). The CDseeker program searched and scored the conserved motifs, terminal stem pairing and antisense in a consensus alignment sequence. The output candidates were assigned as 'guide' and 'orphan' C/D RNAs according to the score assignment. The whole procedure is outlined in Figure 1B and described in detail in Materials and Methods.

To date, 252 C/D snoRNAs, including 119 imprinted C/D snoRNAs and excluding U3, have been identified from the human genome through experimental and conservative (homologous to mouse snoRNAs) detections (17). To test the sensitivity of this computational program on C/D snoRNA genes in human, we chose 133 C/D snoRNA genes and 6 imprinted C/D snoRNA genes as a dataset for training the CDseeker program (if the imprinted snoRNA had multiple gene isoforms, only one isoform was selected to avoid the overfit of the program to these large-amount-repeated imprinted snoRNA genes). We mainly focused on the score assignment and feature evaluation of the known C/D snoRNAs with respect to the following elements: conserved box C, box D and box D′ motifs, terminal stem pairing, conserved antisense elements, indel and substitution pattern and conservation of flanking sequence (Figure 2A). We then applied CDseeker to the training set of snoRNA gene alignments, which were extracted from four human–mammalian WGAs.

As a result, 124 out of 139 known human snoRNA genes (90%) including 6 imprinted snoRNA genes were detected by CDseeker. Fifteen box C/D snoRNA genes were missed due to various reasons, including lack of conservation among any human–mammal genome alignments, location within repeatmask regions, lack of conserved motifs or a large length (>120 nt) (Supplementary Table S2). The training result showed that most of the orphan C/D snoRNAs had scores >26.53 bits and the guide C/D snoRNAs had scores >39.0 bits (Figure 3A and B). We therefore established the cutoff scores of 26.53 bits for orphan C/D snoRNAs and 39.0 bits for guide C/D snoRNAs to identify novel candidates of human C/D snoRNA genes.

The strategy and efficiency of the ACAseeker program for screening of mammalian H/ACA snoRNA genes

We also developed another computational program, ACAseeker, specific to the screening of box H/ACA snoRNAs that are conserved in human–mammal genomes. Similar to CDseeker, ACAseeker first searched conserved motifs in a consensus alignment sequence before evaluating the secondary structure features of box H/ACA snoRNAs. The program then searched and scored the functional antisenses to define the candidates as 'guide' H/ACA RNAs or 'orphan' H/ACA RNAs. The whole procedure is outlined in Figure 1D and described in detail in Materials and Methods.

The same strategy as in CDseeker was applied to test the sensitivity of this computational program on human H/ACA

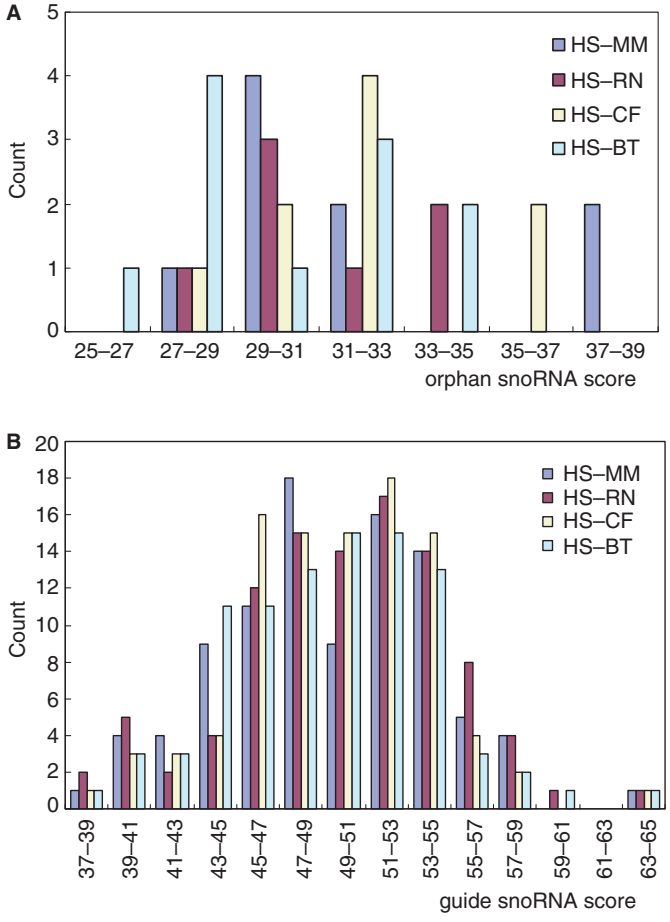

Figure 3. Computational identification of box C/D snoRNAs. (**A**) The distribution of CDseeker scores for known 'orphan' C/D snoRNA genes. (**B**) The distribution of CDseeker scores for known guide C/D snoRNA genes. (HS, MM, RN, CF and BT are abbreviations of human, mouse, rat, dog and cow, respectively. HS-MM represent for human–mouse WGA sequences.)

snoRNA genes. In brief, 100 H/ACA snoRNA genes, which were previously identified through experimental and conservative (homologous to mouse snoRNAs) detection (17), were used to serve as a dataset for training the ACAseeker program. The score assignment and feature evaluation of the known H/ACA snoRNAs were considered according to the following elements: box H, box ACA motifs, secondary structure, minimum free energy, conserved antisense elements, indel and substitution pattern and conservation of flanking sequence (Figure 2A). ACAseeker was then applied to the training set of snoRNA gene alignments which were extracted from four human–mammalian WGAs.

The test results showed that 75 out of 100 known box H/ACA snoRNA genes (75%) were detected by ACAseeker. The remaining 25 box H/ACA RNAs were missed due to reasons similar to those box C/D genes undetected in the CDseeker program (Supplementary Table S3). We found that most of known orphan H/ACA snoRNAs and guide H/ACA snoRNAs had a standard hairpin–hinge–hairpin–tail structure, H box score >0.7 bits, ACA box score >1.63 bits and target score >16.2 bits for guide H/ACA snoRNA. These results provided the cutoff scores and standards for the searching of novel candidates of human H/ACA snoRNA genes.

A genome-wide search of mammalian snoRNA genes with snoSeeker identifies 37 novel human snoRNA genes and 17 novel isoforms of known snoRNA genes

After the training tests of the two programs on known snoRNA genes, we applied the programs to the human genome for an overall search for snoRNA genes of the two families. The whole procedure is outlined in Figure 4A and B.

To search for C/D snoRNA genes, we applied CDseeker to four human–mammal WGA sequences (Figure 4A). About 300 candidates were obtained from the analysis of each human–mammal WGA. The second step was to locate the

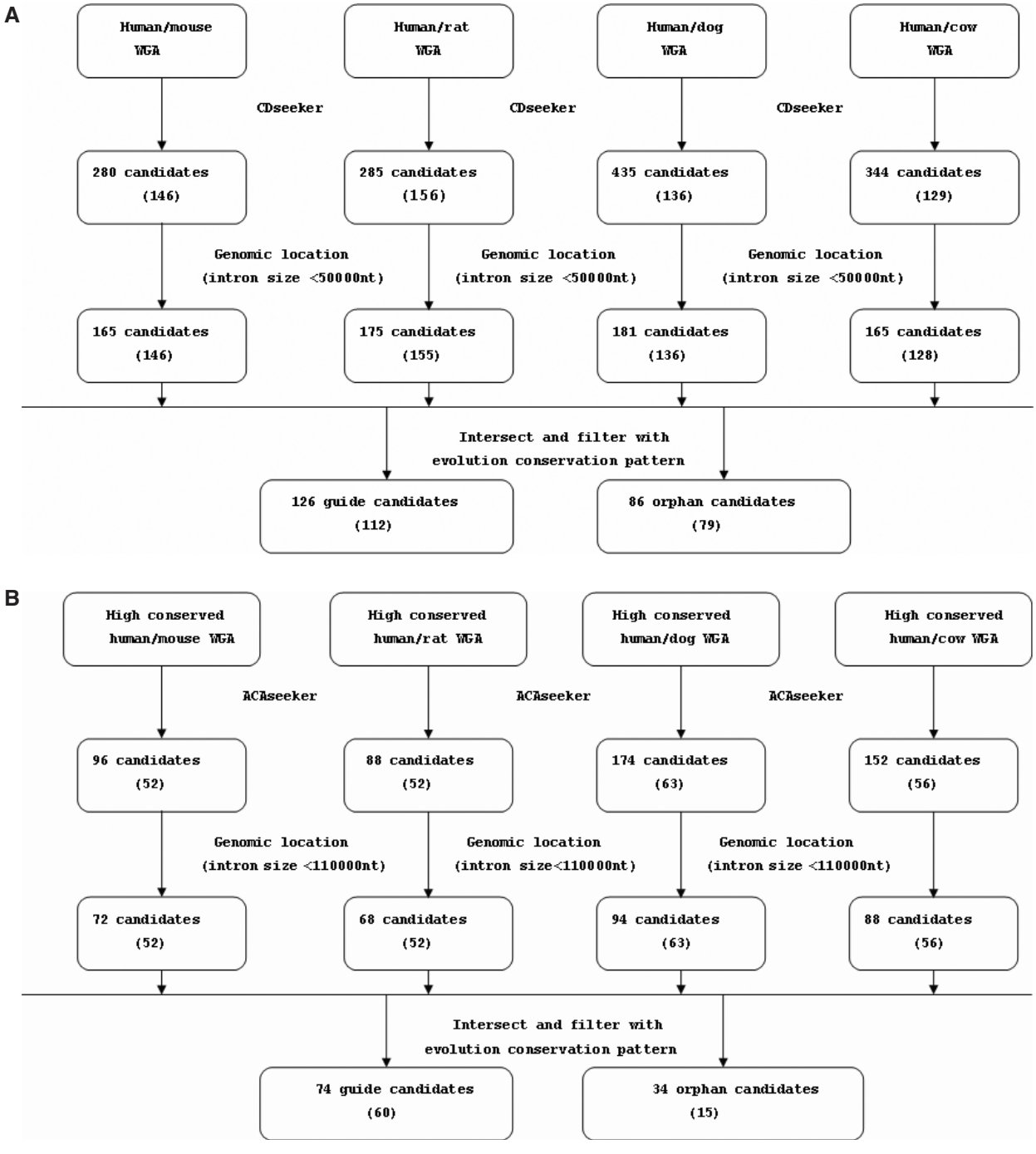

Figure 4. Flowchart of the CDseeker and ACAseeker algorithms. (**A**) The flowchart of the CDseeker algorithm is divided into three main stages. The initial stage is a scan of the four WGA sequences by the CDseeker core program. The second stage is location of the genome using the locateGenome program. The final stage is to intersect the four results and filter the candidate sequence with an evolution conservation pattern. The number of known snoRNAs found at different stages of analysis is shown in parentheses. (**B**) The flowchart of the ACAseeker algorithm is divided into three main stages. The initial stage is a scan of the four high WGA sequences by the ACAseeker core program. The second stage is location of the genome using the locateGenome program. The final stage is to intersect the four results and filter the candidate sequence with an evolution conservation pattern. The number of known snoRNAs found at different stages of analysis is shown in parentheses.

candidates in the human genome with the locateGenome program as described in detail in Materials and Methods. Only candidates located within introns, whose lengths were <50 000 nt, were accepted and the number of candidates for each human–mammal WGA was reduced to <200 after this step. Finally, we intersected the results of the four human–mammal WGA analysis and applied another filter on the results to eliminate false positive candidates. The filter was focused on the conservation of the flanking sequences of the candidates. In total, 212 C/D candidates including 86 orphan candidates were computationally identified from the human genome (Supplementary Table S5; sequences and annotated alignments are available at http://genelab.zsu.edu.cn/HSsnoRNA.html). Of these candidates, 191 were previously identified snoRNA genes and the remaining 21 candidates, including 5 novel orphan genes and 2 novel isoforms of known orphan snoRNA genes, were novel snoRNA candidates.

We similarly applied ACAseeker for searching another family of snoRNA genes, H/ACA genes, from the human genome. As H/ACA snoRNAs are structural RNAs with less successive mutations and indels, we used the conserved alignment sequences as source data to reduce the running time. The H/ACA program was then applied to the conserved alignment sequences in four WGAs, respectively (Figure 4B). About 100 candidates were obtained from each dataset of human–mammal WGA. With the same strategy applied in CDseeker, we located the H/ACA candidates in the human genome. All the candidates within introns with lengths <11 000 nt were selected. We also intersected the four human–mammal WGA results and applied another filter in a manner similar to the search of C/D snoRNA genes with CDseeker. In total, 108 human H/ACA candidates including 11 novel orphan genes and 8 novel isoforms of known orphan snoRNA were obtained (Supplementary Table S5; sequences and annotated alignments are available at http://genelab.zsu.edu.cn/HSsnoRNA.html). Among them, 75 candidates were previously identified snoRNA genes and the remaining 33 candidates, including 13 orphan ones, were novel candidates.

In summary, a total of 54 novel candidates were reported in this study (Tables 1 and 2, and Supplementary Table S4). In addition, >75% of the known snoRNA genes of the two families were detected in our computational scans. More importantly, a large number of orphan snoRNAs, including 26 novel orphan candidates and 94 known orphan genes were detected with the programs developed in this study, showing the efficiency of the programs as snoRNA screening tools. In addition to the ability of snoRNA gene detection, these two programs can also be applied for guiding function prediction of snoRNAs. Twenty-four novel guiding functions were presented in this study with the sequence pairing of the guide RNAs and the target RNAs shown in Supplementary Figure S5. Along with the previous results (29–32), 107 out of 110 ribosomal RNA 2′-O-methylations and 17 out of 25 spliceosomal RNA 2′-O-methylations in human have been assigned to C/D guide RNAs; 86 out of 97 ribosomal RNA pseudouridines and 20 out of 32 spliceosomal RNA pseudouridines in human have been assigned to H/ACA guide RNAs. Our results provide a more comprehensive understanding of the post-transcriptional modification-guided RNA machinery in the human genome.

Expression of the novel snoRNA genes in different tissues

To further confirm the candidates identified computationally, all novel snoRNA candidates, with the exception of eight

Table 1. Novel box C/D snoRNA genes

snoRNA name	Len	Location	Exp.	Homology	Modification	Antisense element	Host gene/comments
Novel guide							
SNORD117	94	chr3:52699794–52699886	N.blot	MM RN CF BT	18S-Gm683	15 nt (3′)	GNL3
SNORD118	101	chr14:44649828–44649928	N.blot	MM RN CF BT	18S-Gm1447	11 nt (3′)	PRPF39
SNORD119	96	chr20:2391598–2391693	N.blot	MM RN CF BT	28S-Am4560	16 nt (3′)	SNRPB
SNORD120	84	chrX:20064093–20064185	N.blot	MM RN BT	U2-Am30	13 nt (3′)	EIF1AX
SNORD121A	91	chr9:33942762–33942852	N.blot	MM RN CF	28S-Gm4607	12 nt (5′)	UBAP2
SNORD121B	93	chr9:33924286–33924378	N.blot	MM RN CF	28S-Gm4607	12 nt (5′)	UBAP2
Novel isoform							
SNORD41B	94	chr19:12675401–12675494	Iso(U41)	MM RN CF BT	28S-Um4276	14 nt (3′)	TNPO2
SNORD12B	103	chr20:47330257–47330359	Iso(HBII-99)	MM RN CF BT	28S-Gm3878	13 nt (5′)	HSUP1
SNORD111B	94	chr16:69120906–69120999	Iso(HBII-82)	MM CF BT	28S-Gm3923	16 nt (5′)	SF3B3
SNORD58C	79	chr18:45269603–45269692	Iso(U58)	RN BT	28S-Um4197	15 nt (5′)	RPL17
SNORD11B	112	chr2:202864285–202864396	Iso(HBII-95)	MM RN CF BT	18S-Gm509	13 nt (3′)	NOP5/NOP58
SNORD105B	92	chr19:10081425–10081516	Iso(U105)	MM RN CF	18S-Um799	15 nt (3′)	P2Y11
Novel orphan							
SNORD122	100	chr2:29004342–29004441	N.blot	MM RN CF BT	Unknown	Unknown	WDR43
SNORD123	88	chr5:9601939–9602026	N.blot	MM RN CF BT	Unknown	Unknown	Hs.34447
SNORD124	104	chr17:35437321–35437424	N.blot	MM RN CF BT	Unknown	Unknown	THRAP4
SNORD125	96	chr22:28059152–28059247	N.blot	CF BT	Unknown	Unknown	AP1B1
SNORD126	99	chr14:19864440–19864538	N.blot	RN CF	Unknown	Unknown	CCNB1IP
Novel isoform							
SNORD116@	106	chr15:22881615–22881718	Iso(HBII-85)	MM RN CF BT	Unknown	Unknown	SNURF-SNRNP-UBE3A
SNORD114@	93	chr14:100534548–100534640	Iso(14q(II))	CF BT	Unknown	Unknown	MEG8

'Iso': is isoforms; 'Len': length of the snoRNA gene (as the program extends 5′ and 3′ stems by 15 nt, the predicted snoRNAs may be 20 nt larger than corresponding snoRNAs confirmed by northern blot); 'Exp': expression situation. 'N.blot' indicate the snoRNA was identified by northern blotting analysis in our work. In the column 'host gene', the protein-coding host genes are denoted by their symbols. In column 'location', the genomic locations are shown. In the column 'modification', a nucleotide with 'm' represents the rRNA or snRNA methylation site that is conserved in mammals. HS, MM, RN, CF, and BT are abbreviations of human (hg18, March 2006), mouse, rat, dog and cow, respectively.

Table 2. Novel box H/ACA snoRNA genes

snoRNA name	Len. (nt)	Location	Exp.	Homology	Modification	Antisense element	Host gene/ comments
Novel guide							
SCARNA26	145	chr1:153915523–153915667	N.blot	CF	U4-Ψ78	6 + 7 nt (5′)	YY1AP1
SNORA82	123	chr3:187986808–187986930	N.blot	MM RN	28S-Ψ4491	3 + 7 nt (5′)	EIF4A2
SCARNA27	126	chr6:8031640–8031765	N.blot	CF BT	U4-Ψ4	7 + 3 nt (5′)	EEF1E1
SNORA83A	135	chr7:64168351–64168485	N.blot	RN	18S-Ψ1367	5 + 7 nt (5′)	LOC441242
SNORA83B	135	chr7:64862474–64862608	N.blot	RN	18S-Ψ1367	5 + 7 nt (5′)	LOC441241
SNORA77B	122	chr22:18493925–18494047	N.blot	MM RN CF	18S-Ψ814	6 + 5 nt (5′)	RANBP1
SNORA77A	123	chr1:201965332–201965454	N.blot (Schattner, ACA63)	MM RN CF BT	18S-Ψ814	6 + 5 nt (5′)	ATP2B4
SNORA80B	135	chr7:6023034–6023168	Schattner.	CF	18S-Ψ109-Ψ572	6 + 4nt (5′) 7+ 4nt(3′)	JTV1
SNORA80A	136	chr21:32671367–32671502	Schattner. (ACA67)	MM BT	18S-Ψ109-Ψ572	6 + 4nt (5′) 7+ 4nt(3′)	C21orf108
SNORA79	140	chr14:80738792–80738931	Schattner. (ACA65)	MM RN	U6-Ψ31-Ψ86	5 + 6nt (5′) 5+ 7nt(3′)	GTF2A1
SCARNA21	138	chr17:7750166–7750303	Schattner. (ACA68)	CF	U12-Ψ19	6 + 4 nt (5′)	CHD3
SNORA76	132	chr17:59577431–59577562	Schattner. (ACA62)	MM RN CF BT	18S-Ψ34-Ψ105	7 + 3nt(5′) 7 + 5nt(3′)	EST cluster
SCARNA22	132	chr5:82395779–82395910	Gu.(U109)	MM RN CF	U1-Ψ6	4 + 5nt (3′)	MGC23909
Novel isoform							
SNORA58B	134	chr1:152498829–152498962	Iso(ACA58)	RN BT	28S-Ψ3823	5 + 9 nt (5′)	UBAP2L
Novel orphan							
SNORA84	133	chr9:94094564–94094696	N.blot(Washietl)	MN RN CF	Unknown	Unknown	IARS
SNORA85	130	chr15:63364852–63364981	N.blot	MM RN CF BT	Unknown	Unknown	PARP16
SNORA86	132	chr7:64163814–64163945	N.blot	MM RN CF BT	Unknown	Unknown	BX649060
SNORA45	127	chr11:8663564–8663690	Gu.(ACA3-2)	MM CF BT	Unknown	Unknown	RPL27A
SNORA12	144	chr10:101986903–101987046	Gu.(U108)	CF	Unknown	Unknown	CWF19L1
Novel Isoform							
SNORA36C	129	chr2:69600679–69600807	Iso(ACA36)	CF	Unknown	Unknown	AAK1
SNORA38B	132	chr17:63167248–63167379	Iso(ACA38)	BT	Unknown	Unknown	NOL11
SNORA70B	134	chr2:61497883–61498016	Iso(U70)	MN BT	Unknown	Unknown	USP34
SNORA70C	134	chr9:118983166–118983299	Iso(U70)	BT	Unknown	Unknown	ASTN2
SNORA11B	128	chr14:90662523–90662650	Iso(U107)	MM BT	Unknown	Unknown	C14orf159
SNORA11C	128	chrX:47132993–47133120	Iso(U107)	MM CF BT	Unknown	Unknown	ZNF157
SNORA11D	127	chrX:51823183–51823309	Iso(U107)	BT	Unknown	Unknown	MAGED4
SNORA11E	127	chrX:51950458–51950584	Iso(U107)	BT	Unknown	Unknown	MAGED4

'Iso': is isoforms; 'Len': length of the snoRNA gene, 'Exp': expression situation. 'N.blot' indicate the snoRNA was identified by northern blotting analysis in our work. 'Schattner', 'Gu' and 'Washietl' indicates the confirmed expression of snoRNAs in other works (ref. 13, 16, 21). In the column 'host gene', the protein-coding host genes are denoted by their symbols. In column 'location', the genomic locations are shown. In the column 'modification', a nucleotide with 'Ψ' represents the rRNA or snRNA pseudouridine site that is conserved in mammals. HS, MM, RN, CF, and BT are abbreviations of human (hg18, March 2006), mouse, rat, dog and cow, respectively.

candidate H/ACA snoRNAs confirmed by three recently published articles (14,16,33) and 20 isoforms of known snoRNAs, were selected for northern blotting analyses. These included 15 novel snoRNA candidates that were conserved between human and rat (rat was chosen for this analysis to facilitate examination of a diverse range of tissues) and 11 novel snoRNA candidates that were conserved between human and mouse or dog or cow (human cell lines: 293T, HeLa, Jurkat and K562 for confirming novel candidates not conserved between human and rat). Total RNAs of eight rat tissues and four human cell lines were used in the experiments. Positive northern blots were obtained for 18 novel snoRNAs (Figure 5A–C) and of these 18 novel snoRNAs, 8 were orphan snoRNAs and 2 (SNORD126 and SNORD123) exhibited tissue-specific or restricted expression. The SNORD126 snoRNA was only expressed in thymus and spleen, while the SNORD123 snoRNA was strongly expressed in lung and heart, with weak expression in spleen and testis (Figure 5A). The remaining six orphan snoRNAs were ubiquitously expressed even though the accumulation levels were not the same in different tissues. It has been generally thought that all guide snoRNAs are expressed ubiquitously as housekeeper RNA genes. Surprisingly, our study showed that one guide snoRNA, SNORD121, was specifically expressed in the thymus (Figure 5A). Another guide snoRNA, SNORD120, was strongly expressed in thymus with lower levels of expression in spleen, lung, kidney, and testis.

DISCUSSION

Recently, numerous orphan snoRNAs have been experimentally identified from different eukaryotes, notably mammals, suggesting that a large group of snoRNAs with unknown function is still hidden in mammalian genomes. With the finding of a novel function for the orphan snoRNA, HBII-52, there has been an increasing trend in orphan snoRNA research in mammalian genomes. In this study, a computational package, snoSeeker, which includes two programs, CDseeker and ACAseeker, was developed for the screening of both guide and orphan snoRNAs. We successfully detected 120 orphan snoRNAs as well as 200 guide snoRNA genes from the human genome using new programs. The programs incorporate a new strategy by taking complementary antisense as an optional criterion for candidate detection. Candidates having a high score (higher than a cutoff score) and complementary antisenses are assigned as guide snoRNAs, while those having a low score (lower than a cutoff score)

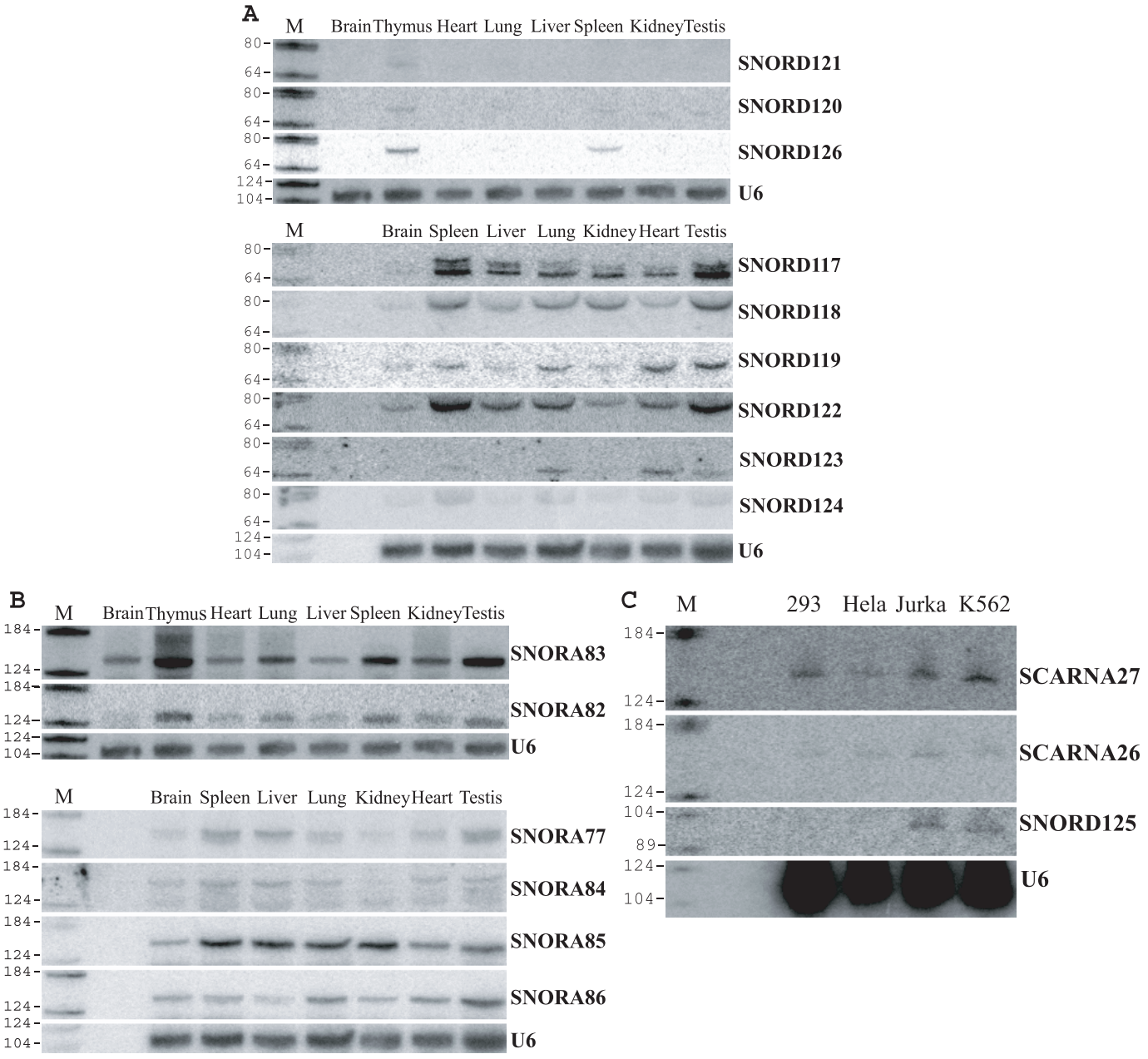

Figure 5. Northern blotting analysis of the expression patterns of novel snoRNAs. Lane M, molecular weight markers (pBR322 digested with HaeIII and 5′-end-labeled with [γ-^{32}P]ATP). The samples of different rat tissues are indicated by the names of tissues. U6 snRNA were analyzed as a control. (**A**) Expression pattern of novel C/D snoRNAs. (**B**) Expression pattern of novel H/ACA snoRNAs. (**C**) Expression pattern of novel snoRNAs in human cell lines. The names of human cell lines are indicated.

and lack of complementary antisenses by computational evaluation are assigned as orphan snoRNAs. In comparison to the existing programs for snoRNA screening which essentially focus on searching guide snoRNAs, our package provides an advanced search engine for an overall analysis of snoRNA genes in mammalian genomes.

It has been well demonstrated that, under selection pressure, functional RNAs [both protein-coding RNAs (ORF) and non-coding RNAs], exhibit highly conserved sequences between phylogenetically related species. Searching functional RNAs within the conserved intronic and/or intergenic sequences has been a key strategy to systematically identify snoRNAs (16), miRNAs (34,35) and other structural non-coding RNAs (33,36). Generally, non-coding RNAs receive higher selection pressure than their flanking sequences. Consistent with this point we found that snoRNA coding regions were more highly conserved than flanking sequences in the first survey of training snoRNA set and resembled a 'hill' in the UCSC genome browser (Figure 2A). Our experimental analyses further supported this notion. For example, some false positive candidates, which had highly conserved coding regions and flanking regions, were unable to be detected in northern blot analysis (Figure 2B and other experimental data not shown). Therefore, we used a conservation filter in

the programs to distinguish the false positive candidates from the authentic candidates based on the difference for the conservation between the coding region of snoRNAs and their flanking regions. By using this filter, the false positive candidates of snoRNA genes were efficiently eliminated in our analyses.

The programs developed in this study were also time-efficient and did not require sophisticated computational equipment. The search duration was ~96 h using a personal computer (CPU 3 GHz) to complete the whole process, including searching the four WGAs of human/mouse, human/rat, human/dog and human/cow. With the low requirement of computational equipment, these programs can therefore be easily popularized.

Frequently, two or even more snoRNAs are encoded within different introns of the same host gene. Interestingly, some isoforms are located in the flanking introns in the same host gene (37,38). The new isoforms of HBII-95, U105, HBII-99, HBII-82, U58 and U41 predicted by the computational methods developed in this study and the two novel snoRNAs, SNORD121A and SNORD121B, are new examples of intragenic snoRNA duplication.

Most, if not all, of the snoRNAs in mammals are intron-encoded. Their host genes, such as ribosomal protein genes and snoRNA binding protein genes, are ubiquitously expressed housekeeper genes and mainly involved in ribosomal biogenesis. Only a few examples of imprinted orphan snoRNAs are found to have a brain-specific expression pattern (39). Surprisingly, we found that two (SNORD121 and SNORD120) newly identified guide snoRNAs exhibited an obvious tissue-specific or restricted expression pattern in this study (Figure 5A). This observation was further supported by the expression patterns of the host genes, UBAP2 and EIF1AX, of these two snoRNAs (in UCSC gene sorter http://genome.ucsc.edu). The increasing number of tissue-specific expressed snoRNAs implies a regulatory role of snoRNAs in gene expression as has been showed for HBII-52 (8,9).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Xiao-Hong Chen for her technical assistance and Dr Peter Schattner for a critical reading of the manuscript and helpful suggestions to improve the manuscript. We would also like to thank Professor Mohsen Ghadessy and Dr Roxana S. Ghadessy for improving the text. This research is supported by the National Natural Science Foundation of China (key project 30230200), and funds from the Ministry of Education of China and Guangdong Province (IRT0447 and NCET-04-0788, NSF-05200303) and the National Basic Research Program (No. 2005CB724600). Funding to pay the Open Access publication charges for this article was provided by the National Basic Research Program (No. 2005CB724600) from the Ministry of Science and Technology of China.

Conflict of interest statement. None declared.

REFERENCES

1. Kiss,T. (2002) Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell*, **109**, 145–148.
2. Balakin,A.G., Smith,L. and Fournier,M.J. (1996) The RNA world of the nucleolus: two major families of small nucleolar RNAs defined by different box elements with related functions. *Cell*, **86**, 823–834.
3. 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.
4. Bachellerie,J.P., Cavaille,J. and Huttenhofer,A. (2002) The expanding snoRNA world. *Biochimie*, **84**, 775–790.
5. 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.
6. Huttenhofer,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.
7. Huang,Z.P., Zhou,H., He,H.L., Chen,C.L., Liang,D. and Qu,L.H. (2005) Genome-wide analyses of two families of snoRNA genes from *Drosophila melanogaster*, demonstrating the extensive utilization of introns for coding of snoRNAs. *RNA*, **11**, 1303–1316.
8. Kishore,S. and Stamm,S. (2006) The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science*, **311**, 230–232.
9. Vitali,P., Basyuk,E., Le Meur,E., Bertrand,E., Muscatelli,F., Cavaille,J. and Huttenhofer,A. (2005) ADAR2-mediated editing of RNA substrates in the nucleolus is inhibited by C/D small nucleolar RNAs. *J. Cell Biol.*, **169**, 745–753.
10. Lowe,T.M. and Eddy,S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, **283**, 1168–1171.
11. 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.
12. Edvardsson,S., Gardner,P.P., Poole,A.M., Hendy,M.D., Penny,D. and Moulton,V. (2003) A search for H/ACA snoRNAs in yeast using MFE secondary structure prediction. *Bioinformatics*, **19**, 865–873.
13. Huttenhofer,A., Cavaille,J. and Bachellerie,J.P. (2004) Experimental RNomics: a global approach to identifying small nuclear RNAs and their targets in different model organisms. *Methods Mol. Biol.*, **265**, 409–428.
14. Gu,A.D., Zhou,H., Yu,C.H. and Qu,L.H. (2005) A novel experimental approach for systematic identification of box H/ACA snoRNAs from eukaryotes. *Nucleic Acids Res.*, **33**, e194.
15. Fedorov,A., Stombaugh,J., Harr,M.W., Yu,S., Nasalean,L. and Shepelev,V. (2005) Computer identification of snoRNA genes using a Mammalian Orthologous Intron Database. *Nucleic Acids Res.*, **33**, 4578–4583.
16. Schattner,P., Barberan-Soler,S. and Lowe,T.M. (2006) A computational screen for mammalian pseudouridylation guide H/ACA RNAs. *RNA*, **12**, 15–25.
17. 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**, D158–162.
18. Maden,B.E. (1990) The numerous modified nucleotides in eukaryotic ribosomal RNA. *Prog. Nucleic Acid Res. Mol. Biol.*, **39**, 241–303.
19. Ofengand,J. and Bakin,A. (1997) Mapping to nucleotide resolution of pseudouridine residues in large subunit ribosomal RNAs from representative eukaryotes, prokaryotes, archaebacteria, mitochondria and chloroplasts. *J. Mol. Biol.*, **266**, 246–268.
20. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.
21. Rabiner,L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE.*, **77**, 257–286.
22. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
23. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.

24. Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*., **9**, 133–148.
25. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al*. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*., **15**, 1034–1050.
26. Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual, 2nd edn*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
27. Park,W., Li,J., Song,R., Messing,J. and Chen,X. (2002) CARPEL FACTORY, a Dicer homolog and HEN1, a novel protein, act in microRNA metabolism in *Arabidopsis thaliana*. *Curr*. *Biol*., **12**, 1484–1495.
28. Zhou,H., Chen,Y.Q., Du,Y.P. and Qu,L.H. (2002) The *Schizosaccharomyces pombe* mgU6-47 gene is required for 2′-O-methylation of U6 snRNA at A41. *Nucleic Acids Res*., **30**, 894–902.
29. 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.
30. 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.
31. Ganot,P., Bortolin,M.L. and Kiss,T. (1997) Related site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. *Cell*, **89**, 799–809.
32. 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.
33. Washietl,S., Hofacker,I.L., Lukasser,M., Huttenhofer,A. and Stadler,P.F. (2005) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat*. *Biotechnol*., **23**, 1383–1390.
34. Berezikov,E., Guryev,V., van de Belt,J., Wienholds,E., Plasterk,R.H. and Cuppen,E. (2005) Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, **120**, 21–24.
35. Lai,E.C., Tomancak,P., Williams,R.W. and Rubin,G.M. (2003) Computational identification of *Drosophila* microRNA genes. *Genome Biol*., **4**, R42.
36. Glazov,E.A., Pheasant,M., McGraw,E.A., Bejerano,G. and Mattick,J.S. (2005) Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. *Genome Res*., **15**, 800–808.
37. Cecconi,F., Mariottini,P., Loreni,F., Pierandrei-Amaldi,P., Campioni,N. and Amaldi,F. (1994) U17XS8, a small nucleolar RNA with a 12 nt complementarity to 18S rRNA and coded by a protein S8 gene. *Nucleic Acids Res*., **22**, 732–741.
38. Cecconi,F., Crosio,C., Mariottini,P., Cesareni,G., Giorgi,M., Brenner,S. and Amaldi,F. (1996) A functional role for some *Fugu* introns larger than the typical short ones: the example of the gene coding for ribosomal protein S7 and snoRNA U17. *Nucleic Acids Res*., **24**, 3167–3172.
39. 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.