Published online 8 April 2009

Nucleic Acids Research, 2009, Vol. 37, No. 9 ***e69***
doi:10.1093/nar/gkp225

Mining small RNA sequencing data: a new approach to identify small nucleolar RNAs in Arabidopsis

Ho-Ming Chen[1,2,3] and Shu-Hsing Wu[1,2,3,*]

[1]Institute of Plant and Microbial Biology, Academia Sinica, [2]Molecular and Biological Agricultural Sciences Program, Taiwan International Graduate Program, National Chung-Hsing University and Academia Sinica, Taipei, 11529 and [3]Graduate Institute of Biotechnology and Department of Life Sciences, National Chung-Hsing University, Taichung, 402, Taiwan

Received December 15, 2008; Revised February 20, 2009; Accepted March 23, 2009

ABSTRACT

Small nucleolar RNAs (snoRNAs) are noncoding RNAs that direct 2′-*O*-methylation or pseudouridylation on ribosomal RNAs or spliceosomal small nuclear RNAs. These modifications are needed to modulate the activity of ribosomes and spliceosomes. A comprehensive repertoire of snoRNAs is needed to expand the knowledge of these modifications. The sequences corresponding to snoRNAs in 18–26-nt small RNA sequencing data have been rarely explored and remain as a hidden treasure for snoRNA annotation. Here, we showed the enrichment of small RNAs at Arabidopsis snoRNA termini and developed a computational approach to identify snoRNAs on the basis of this characteristic. The approach successfully uncovered the full-length sequences of 144 known Arabidopsis snoRNA genes, including some snoRNAs with improved 5′- or 3′-end annotation. In addition, we identified 27 and 17 candidates for novel box C/D and box H/ACA snoRNAs, respectively. Northern blot analysis and sequencing data from parallel analysis of RNA ends confirmed the expression and the termini of the newly predicted snoRNAs. Our study especially expanded on the current knowledge of box H/ACA snoRNAs and snoRNA species targeting snRNAs. In this study, we demonstrated that the use of small RNA sequencing data can increase the complexity and the accuracy of snoRNA annotation.

INTRODUCTION

Modifications of the noncoding RNAs, ribosomal RNAs (rRNAs) and spliceosomal small nuclear RNAs (snRNAs) are thought to influence RNA folding and/or their interactions with proteins for fine-tuning the activity of ribosomes and spliceosomes. rRNAs contain numerous modified nucleotides, of which some are conserved in eukaryotes (1). Two prevalent rRNA modifications are 2′-*O*-methylation at riboses and pseudouridylation of uridines. These two types of modifications are directed by two groups of small nucleolar RNAs (snoRNAs), box C/D and box H/ACA snoRNAs. Box C/D snoRNAs have two conserved motifs, boxes C and D at the 5′ and 3′ ends, respectively. The two motifs are brought together by a 3- to 4-bp terminal stem. The motifs (box C/D) and the stem together form a kink-turn structure (2). However, box H/ACA snoRNAs have two hairpins linked by a hinge and a short tail at the 3′ end. Box H is located at the hinge, and box ACA is usually 3 nt upstream of the 3′ terminus (3).

Box C/D snoRNAs guide 2′-*O*-methylation, whereas box H/ACA snoRNAs direct pseudouridylation, both through site-specific base-pairing of rRNAs and antisense elements on snoRNAs (4). In addition to rRNAs, snoRNAs guide modifications of snRNAs. snoRNAs received this nomenclature by their ability to guide modification of rRNA and U6 in the nucleolus. However, some reports have indicated that modifications of RNA polymerase II-transcribed snRNAs occur at the nucleoplasmic Cajal body in vertebrates by a group of small Cajal body-specific RNAs (scaRNAs) (5). Whether plants also adopt scaRNAs for the modification of some snRNAs remains to be studied. To expand the understanding of the rRNA/snRNA modifications, a comprehensive snoRNA repertoire must be built by identifying additional snoRNAs.

To date, the identification of snoRNAs has been largely achieved by conventional cloning-sequencing and computational prediction based on primary genomic sequences. The cloning-sequencing approach often lacks comprehensiveness and entails technical constrains. For example, snoRNAs identified by conventional sequencing sometimes lack well-defined termini (6). With the increasing availability of genome sequence information, multiple computational programs have been developed to predict box C/D and box H/ACA snoRNAs and have revealed

*To whom correspondence should be addressed. Tel/Fax: 886 2 27871178; Email: shuwu@gate.sinica.edu.tw

many snoRNAs in diverse species (7–12). However, computational snoRNA prediction is usually restricted to snoRNAs with known predicted targets, their secondary structures or their sequences being conserved among species. A revolutionary approach is needed for the discovery of species-specific snoRNAs and snoRNAs with noncanonical targets.

Next-generation sequencing technologies have became powerful tools for functional genomics research (13). High-throughput sequencing of short RNA fragments (18–26 nt) or RNA ends has greatly facilitated the discovery of Arabidopsis interfering RNAs and their targets (14–17). Although interfering RNAs contributed most of small RNA sequence data, a small proportion of small RNA fragments were derived from rRNAs, transfer RNAs, snRNAs and snoRNAs (17). Since these noncoding RNAs are usually longer than 60 bases, small RNA fragments from these transcripts likely result from RNA degradation processes and are usually discarded without further analysis. However, a recent discovery of a snoRNA-derived microRNA (miRNA) suggests that production of some small RNA fragments from these long noncoding RNAs may be through specific biogenesis pathways (18). It is thus worthwhile to further explore the hidden information in these small RNA data.

Here, by mining next-generation sequencing data, we show enriched small RNA fragments at the snoRNA termini and describe a computational approach to identify both box C/D and box H/ACA snoRNAs on the basis of this feature. In addition to revising the sequences of 48 known snoRNA transcripts, we used this approach to identify 44 novel snoRNAs. Newly predicted snoRNAs are supported by their conserved structures, conserved target sites on rRNAs and snRNAs or their expression by alternative approaches. This work presents an additional application of small RNA sequencing data in the annotation of noncoding RNAs other than interfering RNAs and further reveals the complexity of snoRNAs in Arabidopsis.

MATERIALS AND METHODS

Sequence data sets used in this study

Small RNA sequencing data obtained from various Arabidopsis genotypes, tissues and platforms were collected from the following public databases (Supplementary Table 1). Small RNAs cloned from Col-0 and mutants defective in small regulatory RNA pathways were downloaded from the Arabidopsis Small RNA Project database (ASRP, http://asrp.cgrb.oregonstate.edu/) (16,19,20). Small RNA data generated by the studies of Rajagopalan *et al.* and Axtell *et al.* (17,21) were retrieved from NCBI Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/). Thirteen small RNA libraries contributed by three separate studies were obtained from the Arabidopsis SBS database (http://mpss.udel.edu/at_sbs/) (14,22,23). Small RNAs ≥17 nt were pooled and mapped to Arabidopsis genome sequences released by the Arabidopsis Information Resource in 2004 (TAIR, http://www.arabidopsis.org/).

Known Arabidopsis snoRNA sequences were collected from the Scottish Crop Research Institute Plant snoRNA database (http://bioinf.scri.sari.ac.uk/cgi-bin/plant_snorna/home) (24), TAIR and GenBank (http://www.ncbi.nlm.nih.gov/Genbank/). The sequences overlapping with transposons annotated in TAIR8 were removed from the known snoRNA data set.

Data of parallel analysis of RNA ends (PARE) for Col-0 were downloaded from the Arabidopsis PARE website (http://mpss.udel.edu/at_pare/) (14).

Box C/D snoRNA prediction

We searched for small RNAs containing a box C motif and looked for downstream small RNAs that contain a box D motif and can form a 3–4-bp terminal stem with the upstream small RNAs. The box C should be located 4–5-nt downstream of the 5′ start of small RNAs, whereas the box D should be located 3–5-nt upstream of the 3′ terminus of small RNAs. The region defined from the start of a box C containing small RNA to the end of a box D containing small RNA should range from 65 to 300 nt and was further analyzed by the following criteria. First, the numbers of distinct small RNAs mapped to the 5′ and 3′ regions were denoted as N_5 and N_3. Second, the sum of N_5 and N_3 should be ≥3. Third, the number of distinct small RNAs mapped to the antisense strand of this region was indicated as N_{as}. Fourth, to examine the enrichment of small RNAs at both termini, we calculated the reads (*Xinner*) of small RNAs mapped to the positions ≤5 nt to each terminus and the reads (*Xouter*) of small RNAs mapped to the positions >5 and <19 nt from each terminus. The enrichment index (E) was calculated as $E = 0\text{-}(Xouter/Xinner)$.

From our analysis of small RNAs, we observed two characteristics associated with most of the known snoRNAs. First, the number of antisense small RNAs (N_{as}) does not exceed that of small RNAs at termini ($N_5 + N_3 – N_{as} \geq 0$). Second, small RNAs show at least 2-fold enrichment at their termini. Thus, E should be ≥−0.5. When no weighting of E was applied, a cutoff score of –0.5 ($N_5 + N_3 + E – N_{as}$) recovered 114 of the known 'subgroup I' box C/D snoRNA loci (see below), as well as 14 transposon loci. To discriminate snoRNA loci from transposons, which are also rich sources for small RNAs, we applied increased weighting for the enrichment index E. When the weighting of E between 1 and 30 was empirically tested, with a weighting of 10, the 114 known snoRNA loci were retained, but transposon loci were reduced to two. Thus, a score calculated by the following equation was used to evaluate each candidate genomic region for its potential as a box C/D snoRNA locus. For overlapping snoRNA candidates, the one with the best score was selected for each region.

$$ScoreCD = N_5 + N_3 + 10E - N_{as}$$

Box C/D snoRNAs were predicted separately in three subgroups on the basis of the motif sequences of boxes C and D described as follows. The sixth base

of box C and the first base of box D in each subgroup are complementary.

Subgroup I: box C: RTGA(NGA|TGN);
box D: CTGA; $ScoreCD \geq -0.4$;
subgroup II: box C: RTGATT(T|A|G);
box D: ATGA; $ScoreCD \geq 3$; and
subgroup III: box C: R(A|G|C)GATGA;
box D: CTGA; $ScoreCD \geq 3$.

The cutoff scores were determined by rounding the lowest score awarded for the known snoRNA that gave the best separation of known snoRNAs and transposons in each corresponding subgroup.

R indicates a purine base and N indicates any base. Parentheses mean either one of the letters separated by vertical lines within the parentheses. Candidates overlapping with transposons or containing highly repeated small RNAs (>100 genomics hits) were filtered out. The computational program is available upon request.

Box H/ACA snoRNA prediction

A 3′ end supported by at least two distinct small RNAs containing an A(A|T|C)A motif located 3–4-nt upstream of the 3′ terminus was selected for further analysis. A region 110–300-nt upstream of a selected 3′ end was searched for a potential 5′ start supported by at least two distinct small RNAs. The definitions of N_5, N_3 and N_{as} are as described in the prediction of box C/D snoRNAs. To examine the enrichment of small RNAs at both ends, we calculated the reads (*Xinner*) of small RNAs mapped to the positions ≤3 nt from each end and the reads (*Xouter*) of small RNAs mapped to the positions >3 and <10 nt from each terminus. A score calculated by the following equation was used to evaluate each candidate genomic region for its potential as a box H/ACA snoRNA locus.

$$ScoreHACA = N_5 + N_3 + 30E - N_{as}$$

A weighting of E between 1 and 50 was empirically tested in conjunction with the cutoff *ScoreHACA* set between 0 and 6. A combination of 30-fold E weighting and a cutoff score ≥6 retained 20 of 43 known box H/ACA snoRNAs. This process also reduced candidate sequences to 67 for manual structure examination with use of mfold v. 3.2 (http://mfold.bioinfo.rpi.edu/cgi-bin/rna-form1.cgi) (25). Candidates for box H/ACA snoRNA loci should have the folded hairpin-hinge-hairpin-tail structures being the best (lowest) free energy ones and should contain box H motifs (ANANNR) at the hinge region. The computational program is available upon request.

snoRNA target prediction

25S rRNA and 5.8S rRNA sequences were extracted from GenBank accession X52320, and the 18S rRNA sequence was extracted from GenBank accession X16077. Except for U5, experimentally identified spliceosomal snRNA sequences were obtained from the Arabidopsis Splicing Related Gene database (ASRG, http://gremlin3dev.gdcb.iastate.edu/SRGD/ASRG/) (26). The U5 sequence was extracted from GenBank accession X13012.

For box C/D snoRNAs, upstream sequences of box D or D′ were searched for complementarity to Arabidopsis rRNAs or snRNAs. Potential target sites should form at least 10-bp pairing with the 11-nt region located 1-nt upstream of box D or D′ of newly identified snoRNAs. No more than 1 G:U pair was allowed in the first 10 bp. Box D′ motifs could be CNGA or NTGA, and the distance of box D′ to both termini of box C/D snoRNAs had to be at least 25 nt. If more than one target site was predicted for one antisense element, only the best site was chosen for listing in Table 2. The presumptive nucleotides for 2′-*O*-methylation were those paired to the fifth nucleotide upstream of box D or D′ and were further examined for sequence conservation and experimental validation of 2′-*O*-methylation in humans and yeast. Data for human and yeast snoRNAs and RNA modification sites were extracted from snoRNA-LBME-db (http://www-snorna.biotoul.fr/) and the yeast snoRNA database at the University of Massachusetts-Amherst (http://people.biochem.umass.edu/sfournier/fournierlab/snornadb/) and shown as 'Homology' in Table 2 (27,28).

For box H/ACA snoRNAs, on the basis of their structures, pair sequences from internal loops in which the topmost nucleotide was located 13–16 nt upstream of box H or ACA were extracted and searched for complementarity to rRNAs and snRNAs. A potential pseudouridylation site, together with a downstream nucleotide, should be located at the top of the internal loops and flanked by a bipartite duplex of snoRNA and target sequences. The total length of a bipartite should be at least 9 bp and 3 bp for each stem in which no more than 1 G:U pair was allowed. If more than one target site was predicted for one antisense element, only the best site was chosen for listing in Table 3. The presumptive nucleotides for pseudouridylation were further examined for sequence conservation and experimental validation of pseudouridylation in humans and yeast as described above.

snoRNA northern blot analyses

Total RNA was extracted from 10-day-old seedlings, rosette leaves of 4-week-old plants and flowers by use of TRIZOL reagent (Invitrogen, Carlsbad, CA, USA). Ten micrograms of total RNA was separated by 6% or 15% denaturing polyacrylamide TBE-Urea gels (Invitrogen) and transferred to Hybond-N^+ membranes (GE Healthcare, Piscataway, NJ, USA) by use of a semidry transfer cell (Bio-Rad, Hercules, CA, USA). Membranes were UV-crosslinked and then baked at 80°C for 1 h. Antisense oligonucleotides complementary to predicted snoRNAs listed in Supplementary Table 2 were used as probes. The probes were end-labeled with [γ-^{32}P]ATP by use of T4 polynucleotide kinase (New England Biolabs, Ipswich, MA, USA). Hybridization was performed at 42°C overnight after pre-hybridization with ULTRAhyb-Oligo buffer (Ambion, Austin, TX, USA) for at least 1 h. After two washes with 2× SSC and 0.1% SDS for 10 min each at room temperature and one wash with 0.1× SSC

and 0.5% SDS for 1 min at 42°C, the membrane was exposed to Kodak BioMAX MS X-ray films for 1–3 days.

RESULTS

Enrichment of small RNA fragments at the termini of known snoRNAs

To establish the sequence relationship of snoRNAs and snoRNA-derived small RNAs, we analyzed the position of small RNAs on 204 known snoRNAs in the Arabidopsis genome. We divided known snoRNAs into 5′-end, body and 3′-end regions. The 5′- and 3′-ends are genomic regions spanning 11 bases (–5 to +6) of the previously annotated snoRNA 5′ and 3′ termini (Figure 1A). For box C/D snoRNAs, the 5′ and 3′ regions accounted for an average of 22% (~11% for each end) of the total length of snoRNAs, and the remaining 78% is the body region (Figure 1B). However, on mapping the small RNAs to these snoRNAs, those with their 5′ starts mapped to the 5′ end region contributed to more than 40% of total small RNAs mapped to known box C/D snoRNAs in terms of distinct small RNAs (Figure 1B). The percentage increased to 60% if read numbers were used in the calculation (Figure 1B). We also found enrichment of small RNAs with their 3′ termini mapped to the 3′ end region

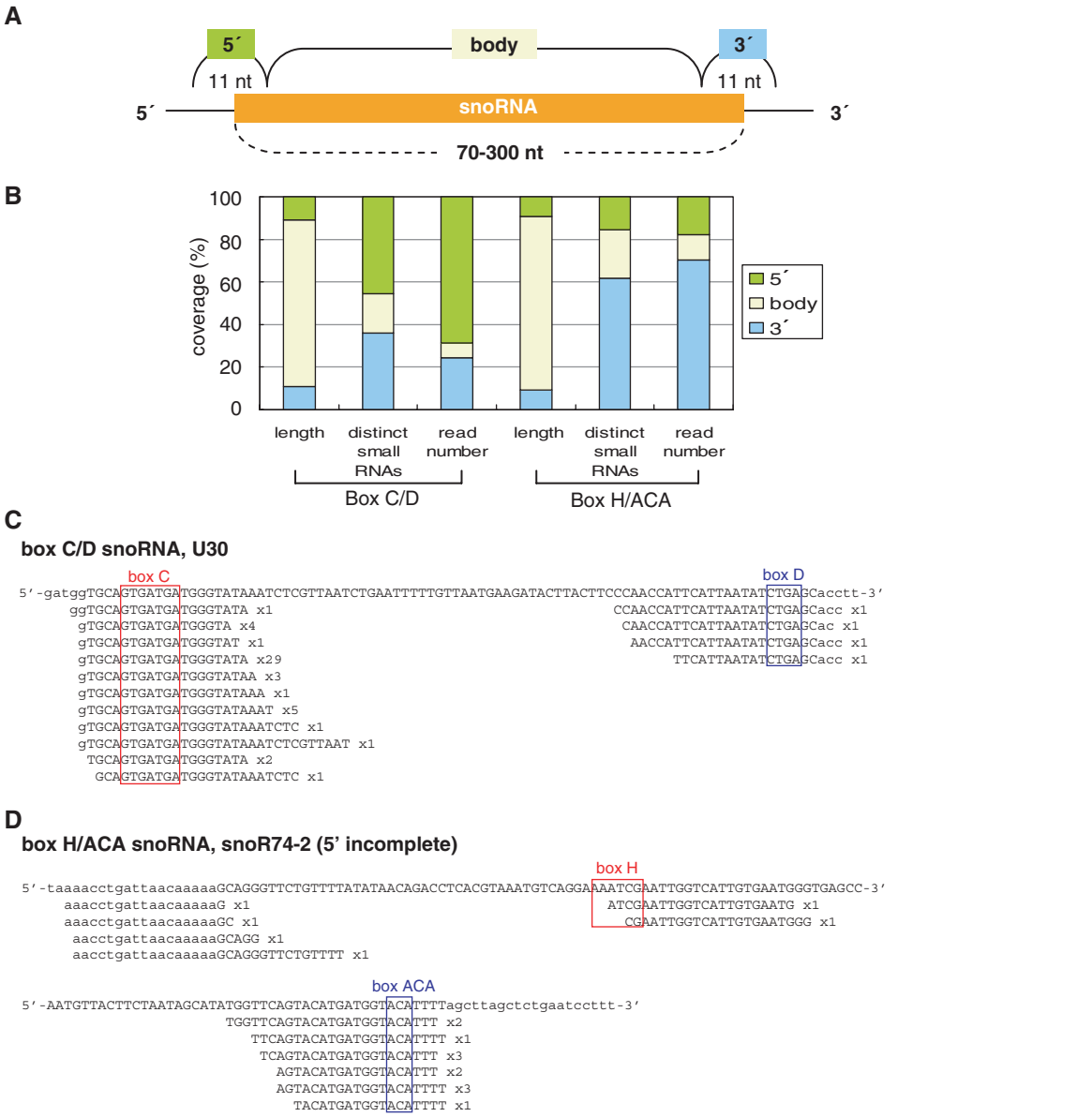

Figure 1. Enrichment of small RNAs at both termini of snoRNAs. (**A**) Definition of 5′-, body and 3′-end regions of a snoRNA. (**B**) The proportion of the sequence length, distinct small RNAs and total reads of small RNAs mapped to each region. (**C**) The distribution of small RNAs on a known box C/D snoRNA, U30. (**D**) The distribution of small RNAs on a known box H/ACA snoRNA, snoR74-2 with an incomplete 5′ end. Capital letters indicate the snoRNA sequences reported previously and lowercase letters indicate the flanking sequences. Short sequences are small RNAs followed by their read numbers. Conserved motifs are highlighted.

of box C/D snoRNAs but to a lesser extent than that in the 5′ end region. For box H/ACA snoRNAs, the enrichment of small RNAs was more significant at the 3′ end than at the 5′ end, which might be due to incomplete knowledge of the 5′ ends for most box H/ACA snoRNAs reported previously (6).

A close examination of the mapping positions of small RNAs revealed that those mapped to the 5′- and 3′-end regions of snoRNAs could accurately define the 5′ and 3′ termini of these snoRNAs. For instance, among 49 small RNA reads mapped to the 5′ end region of a box C/D snoRNA, U30, 45 reads began from the same position, which was 1-nt upstream (–1) from the reported U30 5′ terminus (Figure 1C). Moreover, three of four small RNA reads mapped to the 3′-end region terminated at the same position, which was 3-nt downstream (+4) from the reported U30 3′ terminus. Of interest, a canonical 4-bp terminal stem could be successfully formed if the 5′ (–1) and the 3′ (+3) ends defined by small RNAs were adopted for the mature U30 sequence (data not shown). This terminal stem is a standard feature for box C/D snoRNAs but could not be identified in the previously reported U30 sequence.

Small RNAs also well supported the 3′ terminus of a known box H/ACA snoRNA, snoR74-2 (Figure 1D), which was found to have an incomplete 5′ end in a previous report (6). A cluster of small RNAs was observed upstream of the previously reported 5′ terminus of snoR74-2. This small RNA-enriched site likely represents the true 5′ terminus of snoR74-2. Indeed, the size of snoR74-2 with the newly defined 5′ terminus was consistent with that previously observed with northern blot analysis (6). A canonical hairpin-hinge-hairpin-tail structure could also be formed for snoRNA74-2 with the extended 5′ terminus (data not shown).

Prediction of snoRNAs by small RNA data

The predominant occurrences of degraded small RNAs at the snoRNA termini prompted us to examine the possibility of using small RNA data for the annotation of snoRNAs similar to the use of expressed sequence tags to annotate genes. For this purpose, we developed a computational approach to identify snoRNAs by integrating the enriched behavior of small RNAs at snoRNA ends as described above (Figure 1) and conserved features of known snoRNAs. These features include: (i) small RNAs containing conserved snoRNA motifs (i.e. box C/D or box ACA) at desired positions; (ii) pairs of 5′- and 3′-end small RNAs within the typical length of snoRNAs; (iii) the ability to form conserved structures (i.e. terminal stem or hairpin-hinge-hairpin-tail); (iv) enrichment of small RNAs surrounding the termini of predicted snoRNAs; and (v) depletion of small RNAs from the antisense strand of predicted snoRNAs. The fifth criterion was adopted to distinguish snoRNA-derived small RNAs from small interfering RNAs (siRNAs), which are abundant throughout the Arabidopsis genome (17). Details of criteria used for predicting box C/D or box H/ACA snoRNAs are described in 'Materials and Methods' section.

We divided the box C/D snoRNA prediction into three subgroups based on specific combinations of box C and box D. Boxes C and D are brought together by a terminal stem, and the sixth nucleotide of box C is opposite to the first nucleotide of box D (2). The pairing of these two nucleotides, together with neighboring pairings, form a short stem that is essential for the binding of nucleolar proteins (29). Currently, without the constrains of base-pairing of these two nucleotides, algorithms used to predict box C/D snoRNAs usually evaluate the box C and box D motifs independently. By considering the co-occurrence of these two base-pairing nucleotides in different subgroups (GC or AT pairs) as described in 'Materials and Methods' section, our algorithm may improve the sensitivity and specificity of box C/D snoRNA prediction.

With our approach, we could identify 124 of 161 known box C/D snoRNAs and 27 candidates for novel box C/D snoRNAs (Tables 1 and 2 and Supplementary Tables 3 and 4), which indicates the robustness of the small RNA-aided prediction methodology. Among the 37 known box C/D snoRNAs not identified by our approach, most lacked small RNA fragments mapped to their 5′ end and/or 3′ end. The under-representation of these small RNAs might be due to the higher stability or lower expression of these snoRNAs in the sequence libraries we used in the current analyses. The prediction results also amended the 5′ and/or 3′ termini of 30 known box C/D snoRNAs (Supplementary Table 3). The newly annotated sites deviated >3 nt from the termini reported previously.

Our small RNA-based approach also successfully identified 20 of 43 known box H/ACA snoRNAs (Table 1). We could not identify the remaining known box H/ACA snoRNAs by our approach mostly because of the absence or under-representation of small RNAs at snoRNA ends. The mapping of small RNAs successfully extended the 5′-end boundaries of 18 known H/ACA snoRNAs (Supplementary Table 3), which were originally found to lack complete 5′ ends (6). The resulting full-length snoRNA sequences are thus able to form intact hairpin-hinge-hairpin-tail structures (data not shown). Therefore, our method of defining the 5′ ends by small RNAs will greatly improve the current annotations of box H/ACA snoRNAs. Moreover, 17 candidates of novel box H/ACA snoRNAs were revealed by our prediction method (Supplementary Table 4) and showed structural resemblance to known box H/ACA snoRNAs (Figure 2). Furthermore,

Table 1. Number of snoRNAs identified by small RNA sequencing data

snoRNA	Known	Novel
Box C/D	124 (30)/161	27
Subgroup I	114 (26)	24
Subgroup II	8 (3)	2
Subgroup III	2 (1)	1
Box H/ACA	20 (18)/43	17

The denominators of known snoRNAs are the total numbers of snoRNAs reported in previous studies.
The numbers of snoRNAs with predicted 5′ and/or 3′ ends deviating more than 3 nt from those in previous reports are in parentheses.

Table 2. Characteristics of novel box C/D snoRNAs predicted by small RNA sequencing data

Name	Size (nt)	Location[a]	Target site[b]	Homology[c]
U27-2	84	IGR, cluster	Am42/U6 (D) Am28/18S (D′)	Am47 (Hs: mgU6-47) Am27 (Hs: U27); Am28 (Sc: snR74)
U46-1	79	Intron of At3g51800	Am2246/25S (D)	Am3739 (Hs: U46); Am2256 (Sc: snR63)
U46-2	83	IGR, cluster	Am2246/25S (D)	Am3739 (Hs: U46); Am2256 (Sc: snR63)
snoR102-2	165	3′ UTR of At4g30993	Gm41/U5 (D)	
snoR113	103	IGR	Gm1446/25S (D′)	
snoR114	88	IGR	Gm1191/25S (D) Gm1590/18S (D′)	
snoR115-1	68	IGR, cluster	Um602/18S (D′)	
snoR115-2	69	AS of At1g18740, cluster	Um602/18S (D′)	
snoR116	86	IGR, cluster	Um123/18S (D′)	Um121 (Hs: mgh18S-121, Z17B)
snoR117	72	IGR, cluster	Um2103/25S (D′)	
snoR118	93	IGR, cluster	Cm1280/25S (D′)	
snoR119	75	Intron of At5g01350, cluster	Am778/18S (D′)	
snoR120	74	IGR, cluster	Um44/25S (D)	
snoR121	85	IGR	Cm1510/25S (D)	Cm2409 (Hs: mgh28S-2409)
snoR122	74	AS of At1g18740, cluster	Um168/18S(D′)	Um172 (Hs: U54A)
snoR123a	80	Intron of At4g31980, cluster	Um2483/25S(D′)	
snoR123b	79	Intron of At4g31980, cluster	Um2483/25S(D′)	
snoR124	68	IGR, cluster	Gm244/18S(D′)	
snoR125	200	IGR	Am31/U2 (D)	Am30 (Hs: mgU2-19/30)
snoR126	105	5′UTR of At5g65860	Gm75/U6 (D) Gm84/U6 (D′)	
snoR127	138	Intron of At5g27720	Gm20/U2 (D′)	Gm19 (Hs: mgU2-19/30)
snoR128	92	IGR, cluster		
snoR129	88	IGR, cluster		
snoR130	120	Intron of At3g07860		
snoR131	77	IGR, cluster		
snoR132	82	IGR, cluster		
snoR133	92	IGR, cluster		

[a]The location of snoRNAs is described as IGR (intergenic region), AS (antisense), UTR (untranslated region) and intron of coding genes. Cluster indicates potential polycistronic snoRNA.
[b]The target site is indicated as the methylated nucleotide followed by its position on rRNAs or snRNAs.
The location of antisense elements that are complementary to rRNAs or snRNAs is indicated in parentheses as D or D′.
[c]Homology indicates equivalent nucleotides on which methylation has been experimentally validated in humans (Hs) and/or yeast (Sc). Equivalent nucleotides are followed by snoRNAs guiding the modification in parentheses.

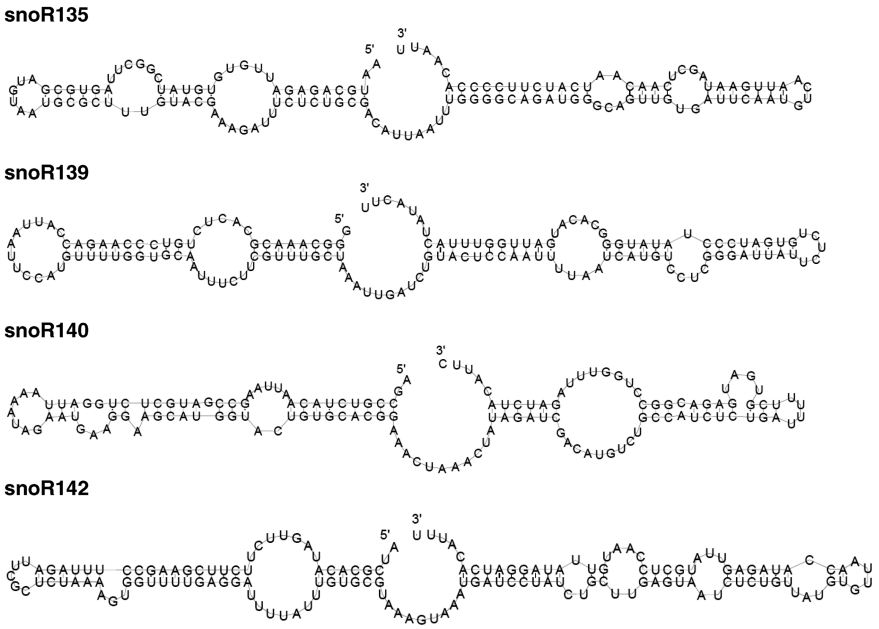

Figure 2. Structures of novel box H/ACA snoRNAs predicted by small RNA sequencing data. The program of mfold was used to predict structures of newly identified snoR135, snoR139, snoR140 and snoR142 as described in 'Materials and Methods' section.

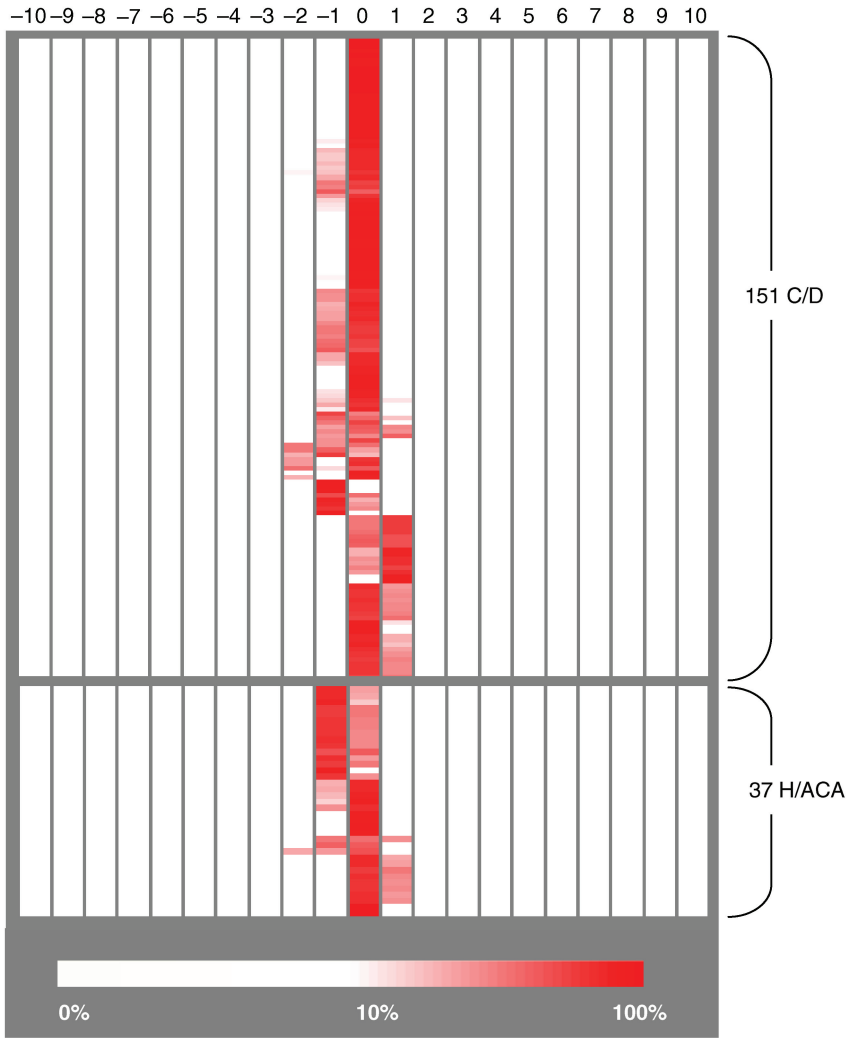

Figure 3. Validation of the predicted snoRNA 5′ ends by PARE data. The positions of predicted 5′ ends of 151 box C/D snoRNAs and 37 box H/ACA snoRNAs were set as 0. Positions upstream of the predicted 5′ ends were set as negative values and downstream as positive values. The read number of 5′ ends starting at each position (–10 to 10) was normalized to the total PARE read number of the 21-nt region. A gradient of red color was used to represent the frequency of 5′ ends starting at each position.

similar to known box H/ACA snoRNAs, the 5′ termini of the newly predicted box H/ACA snoRNAs have 1 or 2 nt protruding from the 5′ hairpin (3).

Our results indicate that the methodology we developed could effectively identify both known and novel snoRNAs and could efficiently improve the annotation of snoRNAs.

Validation of predicted snoRNAs by PARE data and northern blot analysis

Two distinct approaches were used to verify the authenticity of snoRNAs predicted by our methodology. We first cross-examined our results of the 5′ termini of snoRNAs with a high-throughput data set of 5′ ends sequenced by PARE (14). PARE revealed the 5′ ends of uncapped transcripts and was first developed to identify miRNA targets. Although most plant pre-snoRNAs are independently transcribed by RNA polymerase II or excised from introns (30), the activity of endonucleases and exonucleases is required for the production of mature snoRNA 5′ ends in yeast (31,32). Therefore, similar to the cleavage products of miRNA target mRNAs, snoRNAs are likely to be cap-free and could be targets of PARE analyses. Indeed, we found that the PARE reads corresponding to the 5′ termini of many known snoRNAs could surpass those of known miRNA targets (data not shown).

To validate whether our prediction results really reflect the 5′ termini of endogenous snoRNAs, we then extracted PARE reads at positions of snoRNA 5′ termini predicted by small RNAs and their neighboring sequences. The frequency of occurrence at each position was then plotted against the relative positions to the predicted 5′ terminus for each snoRNA. As shown in Figure 3 and Supplementary Table 5, both known and newly predicted box C/D and box H/ACA snoRNAs had the highest PARE reads at the predicted 5′ termini or the +/–1 positions. This finding indicates the high accuracy of our method to predict the 5′ termini of snoRNAs. The analysis

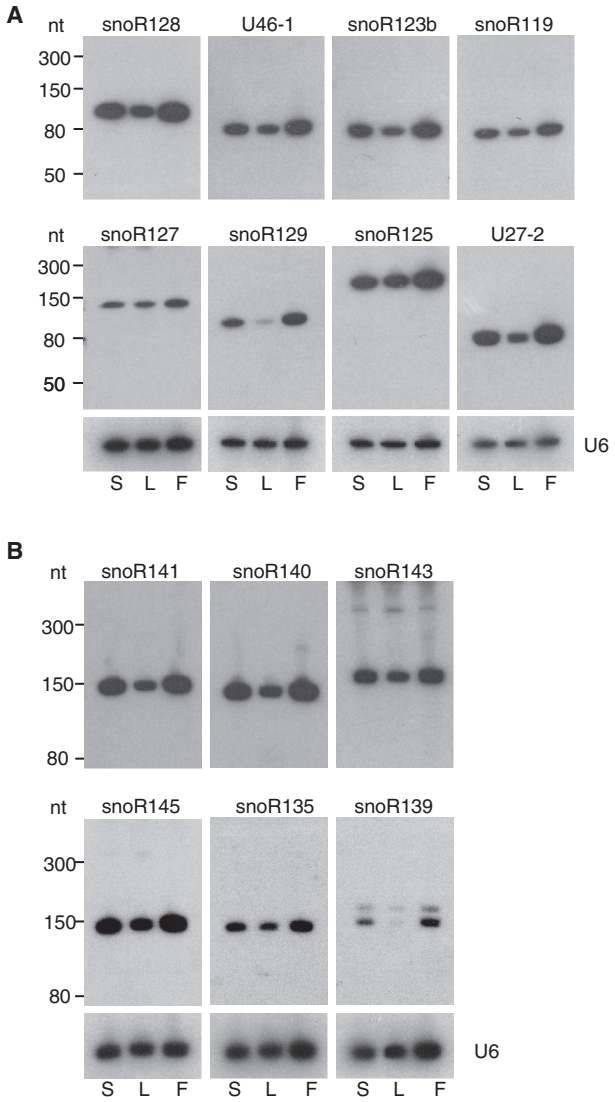

Figure 4. Northern blot analysis of novel snoRNAs predicted by small RNA sequencing data. (**A**) Aliquots of 10 μg total RNA for each lane were separated on 15% TBU gel for the northern blot analysis of eight new box C/D snoRNAs. snoR128: 92 nt; U46-1: 79 nt; snoR123b: 79 nt; snoR119: 75 nt; snoR127: 138 nt; snoR129: 88 nt; snoR125: 200 nt; U27-2: 84 nt. (**B**) Aliquots of 10 μg total RNA for each lane were separated on 6% TBU gels for the northern blot analyses of six new box H/ACA snoRNAs. snoR141: 151 nt; snoR140: 142 nt; snoR143: 171 nt; snoR145: 145 nt; snoR135: 141 nt; snoR139: 150 nt; Spliceosomal snRNA U6 was used as the loading control. S, 10-day-old seedlings; L, rosette leaves; F, flowers.

also suggests that the snoRNA species predicted by our method are expressed thus exist in the sequence libraries used for the PARE analysis.

We also examined the sizes and expression of several newly predicted snoRNAs by northern blot analysis. The sizes of eight box C/D snoRNAs are consistent with those predicted by our computational analyses. All these snoRNAs could be detected in the tissues examined, including 10-day-old seedlings, 4-week-old leaves and flowers (Figure 4A). Northern blot analysis also validated the predicted sizes of six novel box H/ACA snoRNAs (Figure 4B). In addition to a band at the expected size, a probe complementary to snoR139 revealed a transcript with a slightly larger size, which could be derived from pre-snoR139 or was a result of cross-hybridization of a closely related species. Similar to box C/D snoRNAs, five novel box H/ACA snoRNAs had similar expression level among the three tissues. Of note, snoR129 and snoR139 were the only two snoRNAs with noticeably lower expression in leaves than in other tissues.

Both PARE data and northern blot analysis verified the prediction of snoRNAs by small RNAs, which suggests that small RNA data can be used to precisely predict the 5′ end and full-length sizes of snoRNA species.

Genome organization of loci encoding novel snoRNAs

We next investigated whether the new snoRNA loci exhibit unique features in their genome organization. The results summarized in Tables 2 and 3 indicate that 37 of 44 novel snoRNAs are produced from intergenic regions or introns, as most known Arabidopsis snoRNAs are. In some cases, snoRNAs reside in genes with related functionality for the modification of rRNAs or snRNAs. For example, previous results showed that two box C/D snoRNAs, U60.1 and U60.2, are located in the intron of two genes encoding fibrillarins, which are nucleolar proteins associated with box C/D snoRNAs (33). In our study, snoR140 is clustered with a known box H/ACA in the intron of a gene encoding an H/ACA ribonucleoprotein complex subunit 2-like protein (At5g08180). Moreover, a new box C/D snoRNA, snoR127, which was predicted to target the spliceosomal snRNA U2, is located in the intron of At5g27720, which encodes a protein similar to small nuclear ribonucleoprotein. In addition to the introns of genes involved in translation or splicing, novel snoRNAs are also found to locate in introns of genes encoding a nuclear DNA-binding protein and three unknown proteins.

Five novel snoRNAs are located in the regions annotated as the 5′ untranslated region (UTR), the 3′ UTR, the coding region or the antisense strand of a coding gene. All these genes are annotated to encode unknown proteins, except for the one producing snoR126. The transcript of snoR126 is part of the 5′ UTR or the alternative-spliced intron of a gene encoding an ankyrin repeat family protein (At5g65860).

Targets of novel snoRNAs

Among 27 newly predicted box C/D snoRNAs, 21 were predicted to target rRNAs or snRNAs on the basis of the target prediction criteria described in 'Materials and Methods' section (Table 2). The methylation sites of some target sites are evolutionarily conserved and have been experimentally validated in yeast and/or humans. Five new box C/D snoRNAs were predicted to target spliceosomal snRNAs. Among them, snoR126 has dual antisense elements that may target two neighboring sites on the spliceosomal snRNA U6. Several snoRNAs with dual targets on the same rRNAs were proposed to interact with both sites simultaneously (34). snoR126 may be another example of concurrent targeting of a snoRNA to a single

Table 3. Characteristics of novel box H/ACA snoRNAs predicted by small RNA sequencing data

Name	Size (nt)	Location[a]	Target site[b]	Homolgy[c]
snoR80-2	140	IGR, cluster	Ψ1130/25S (P1)	
			Ψ999/25S (P2)	Ψ990 (Sc: snR49)
snoR80-3	138	IGR, cluster	Ψ1130/25S (P1)	
			Ψ999/25S (P2)	Ψ990 (Sc: snR49)
snoR86-2	142	IGR, cluster	Ψ2555/25S (P1)	
			Ψ360/18S (P2)	Ψ406 (Hs: U71)
snoR97-2	134	IGR, cluster	Ψ2964/25S (P1)	
			Ψ2707/25S (P2)	Ψ4263 (Hs: ACA2)
snoR103-2	132	Intron of At4g30680	Ψ50/U5 (P1)	Ψ46 (Hs: U85, U89)
			Ψ47/U5 (P2)	Ψ43 (Hs: ACA57)
snoR134	144	IGR, cluster	Ψ1104/18S (P1)	
			Ψ1192/18S (P2)	Ψ1248 (Hs: ACA13); Ψ1191 (Sc: snR35)
snoR135	141	At2g46192 (ncRNA), cluster	Ψ2181/25S (P2)	Ψ3674 (Hs:unknown); Ψ2191 (Sc: snR32)
snoR136	155	IGR, cluster	Ψ2833/25S (P2)	Ψ4390 (Hs: E3)
snoR137	146	IGR	Ψ1208/18S (P1)	
snoR138	156	IGR	Ψ1479/18S (P1)	
snoR139	150	IGR, cluster	Ψ762/18S (P1)	Ψ815 (Hs: ACA28); Ψ759 (Sc: snR80)
snoR140	142	Intron of At5g08180, cluster	Ψ677/18S (P1)	
			Ψ1215/18S (P2)	
snoR141	151	IGR, cluster	Ψ1050/25S (P1)	Ψ1769 (Hs: ACA9); Ψ1042 (Sc: snR33)
			Ψ2093/25S (P2)	
snoR142	154	IGR, cluster	Ψ26/U6 (P1)	Ψ31 (Hs: ACA65)
			Ψ2884/25S (P2)	Ψ4441 (Hs: ACA1)
snoR143	171	CDS of At1g31835	Ψ38/U2 (P2)	Ψ37 (Hs: ACA45)
snoR144	144	IGR, cluster	Ψ1897/25S (P1)	
			Ψ1255/25S (P2)	
snoR145	145	At2g46192 (ncRNA), cluster		

[a]The location of snoRNAs is described as ncRNA (noncoding RNA), IGR, CDS (coding region), and intron of coding genes.
Cluster indicates potential polycistronic snoRNA.
[b]The target site of new box H/ACA snoRNAs is indicated as a pseudouridine, followed by its position on rRNAs or snRNAs.
The location of antisense elements that are complementary to rRNAs or snRNAs is indicated in parentheses as P1 (the 5′ internal loop) or P2 (the 3′ internal loop).
[c]Homology indicates equivalent nucleotides on which pseudouridylation has been experimentally validated in humans (Hs) and/or yeast (Sc).

snRNA species. Two sites individually targeted by snoR125 and snoR127 on the spliceosomal snRNA U2 are conserved in humans and Arabidopsis. The modifications of these two sites have been experimentally validated in humans (35). However, a scaRNA, mgU2-19/30, is the guide RNA responsible for the methylation of these two sites (36). That snoR125 and snoR127 target U2 implies that they are potential scaRNAs in Arabidopsis.

All, but one, novel box H/ACA snoRNAs were predicted to target rRNAs or snRNAs (Table 3). Roughly 60% of box H/ACA snoRNAs have potential targets for both antisense elements. This is similar to the previous finding that, for 50% of Arabidopsis box C/D snoRNA, both D and D′ have potential targets (34). The pseudouridylation of 13 predicted target sites have been experimentally validated in yeast and/or humans. Both newly predicted snoR103-2 and previously reported snoR103-1 were predicted to target two sites of snRNA U5 (Table 3) (6). Pseudouridylation of equivalent sites on human U5 was experimentally validated and predicted to be directed by three scaRNAs (37–39). Among them, human Ψ46/U5 is presumably directed by U85 and U89 (37,39), whereas Ψ43/U5 is directed by ACA57 (38). Interestingly, human U85 and U89 are composed of both box C/D and box H/ACA motifs while snoR103-1 and snoR103-2 contain only the box H/ACA motif. Arabidopsis snoR143 was predicted to direct U2 Ψ38 modification (Table 3). The equivalent site of human U2 was guided by a scaRNA ACA45 (38). A recent paper demonstrated that a small fraction of ACA45 is processed by Dicer to generate a small RNA with miRNA-like function (18). Such phenomenon was not observed for snoR143 in the current data sets we analyzed.

Of note, no rRNAs or snRNAs are identified as potential targets of seven new snoRNA we identified (Tables 2 and 3). This finding could be due to the more stringent criteria we applied in the target prediction. Alternatively, these snoRNAs might target RNA species other than rRNAs or snRNAs.

DISCUSSION

Our analysis of more than 30 million reads of small RNA data revealed enriched small RNAs at termini of snoRNAs. By the use of a computational approach integrating this characteristic and conserved motifs/structures of snoRNAs, we could re-annotate 48 known snoRNAs and identify 44 novel snoRNAs in the Arabidopsis genome. The newly identified snoRNAs especially expand the knowledge of Arabidopsis box H/ACA snoRNAs, snoRNAs targeting snRNAs for modification and snoRNAs without canonical targets. This study describes the first application of small RNA data in the study of snoRNAs that are 70–300-nt long.

With our successful application, small RNA data are new resources for the discovery and annotation of snoRNAs. Our work also demonstrates that, with appropriate mining tools, the analysis of small RNA data generated by next-generation sequencing will increase our understanding of longer noncoding RNAs in addition to miRNAs or siRNAs.

The observation of small RNA fragments enriched at snoRNA termini raises the questions of how these small RNAs are generated and whether they have biological functions. snoRNA ends might be protected from degradation because they and/or their neighboring sequences, which contain conserved motifs, are bound by nucleolar proteins. Alternatively, there might be unknown RNA degradation or RNA processing pathways that prefer to direct endonucleolytic cleavage near snoRNA ends. For example, the production of miRNA-like small RNAs from a human box H/ACA snoRNA depends on Dicer, the RNase III enzyme responsible for the biogenesis of miRNAs and siRNAs (18). Further studies will help to determine whether small RNAs from Arabidopsis snoRNA termini are generated from similar pathways and have silencing activity similar to the miRNA-like small RNA from the human snoRNA.

Similar to the use of expressed sequence tags to annotate genes, the use of small RNAs to identify snoRNAs may help uncover snoRNAs with atypical motifs or structures. Nevertheless, the abundance of snoRNA-derived small RNAs may heavily depend on the tissues sampled and the depth of small-RNA sequencing. To decrease the false-positive rate, our approach required at least two distinct small RNAs to support the potential termini of snoRNAs. Some known snoRNAs missed by our approach have only single small RNAs at each terminus or only small RNAs for one of the termini. This drawback can be overcome by integrating small RNA data with other snoRNA-predicting computational programs based mainly on genomic sequences. For example, single small RNA fragments can be used as seeds to initiate the computational prediction of snoRNAs. The relaxation of the number requirement of small RNAs at termini will likely increase the prediction sensitivity of our algorithm.

The presence of small RNAs may replace the knowledge of known targets and evolutionary conservation, which was usually required in previous prediction programs. This information will especially improve the discovery of snoRNAs that do not target rRNAs and snRNAs or are species specific. The utilization of small RNA data may also allow for the search for noncanonical snoRNAs with less stringent thresholds for conserved motifs/structures. Our method could be easily adapted for the annotation of snoRNAs in species other than Arabidopsis.

The 5′ termini of snoRNAs identified by small RNA data were validated by PARE data, which were generated from high-throughput sequencing of 5′ termini of transcripts lacking the 5′ cap (14). According to the experimental procedure in generating PARE data, the 5′ termini sequenced by PARE technology theoretically should come from transcripts with poly(A) tails (14). As mature snoRNAs lack poly(A) tails, PARE reads from 5′ ends of snoRNAs might have been generated from low-efficiency annealing of the dT priming oligo to the 3′- end of the RNA transcripts. Since the 5′ termini of most snoRNAs have been sequenced more than 50 times in the PARE data set, PARE data provide a great opportunity to study the maturation of snoRNA 5′ termini (Supplementary Table 5). The analysis of PARE reads and snoRNA genes may reveal clues to understanding how the snoRNA 5′ termini are defined and how precise the maturation is. Although PARE was first developed to identify miRNA targets, PARE data also facilitate the annotation of other non-coding RNAs and, potentially, their maturation process.

Our results increase the number of snoRNAs targeting snRNAs from 4 to 12 and thus provide more candidates for the study of this group of snoRNAs (Tables 2 and 3). The predicted target sites on Arabidopsis U2, U5 and U6 for newly identified snoRNAs have previously been validated in *Vicia faba* or *Pisum sativum* (40). Among these 12 snoRNAs, nine were predicted to target RNA polymerase II-transcribed snRNAs, U2 and U5, and are presumably localized in the Cajal body as human scaRNAs. However, the localization of plant scaRNAs and the determinants of their localization have not been well characterized. Moreover, our study did not identify any potential Arabidopsis scaRNAs with the signatures for both box C/D and box H/ACA scaRNAs as were previously described for U85, U87, U88 and U89 in humans (37,39).

In contrast to the intergenic localization of most Arabidopsis snoRNAs targeting rRNAs, eight of the nine snoRNAs targeting U2 and U5 are located in genes (Tables 2 and 3). This finding suggests that snoRNAs targeting snRNAs may evolve differently from those targeting rRNAs in Arabidopsis. Further investigation of snoRNAs with snRNA targets in other plant species such as rice and poplar may shed light on snoRNA evolution.

Our current target prediction failed to yield rRNA or snRNA targets for six box C/D snoRNAs and one box H/ACA snoRNA. Although snoRNAs largely target rRNAs or snRNAs for modification, in rare cases, box C/D snoRNAs target tRNAs for methylation or pre-mRNA for alterative splicing (41,42). When including tRNAs in the target search for the newly identified box C/D snoRNAs without identified rRNA/snRNA targets, none of these snoRNAs could target tRNAs on the basis of the criteria we applied (data not shown). We have also performed genome-wide target prediction for all new box C/D snoRNAs. With our current criteria, hundreds to thousands of potential target sites throughout the genome could be identified (data not shown). More stringent criteria should be applied, but with only one preceding report of mRNA targeted by snoRNAs, deriving such criteria remains immature.

Because of the decreasing cost and increasing throughput, next-generation sequencing is becoming a popular or even a regular approach to profile mRNAs or small RNAs in a genome-wide scale. RNA degradation is a crucial process to regulate RNA homeostasis within a cell. Therefore, degraded RNA fragments will be seen in any RNA sequence data set to various degrees.

As demonstrated in this study, the mining of degraded RNA sequence data can provide unforeseen opportunities to study noncoding RNAs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We thank the Wu lab members for helpful discussions.

FUNDING

A research grant from Academia Sinica (to S-H Wu, Foresight Project L20-2). Funding for open access charge: Academia Sinica Grant 034002.

Conflict of interest statement. None declared.

REFERENCES

1. Maden,B.E. (1990) The numerous modified nucleotides in eukaryotic ribosomal RNA. *Prog. Nucleic Acid Res. Mol. Biol.*, **39**, 241–303.
2. Watkins,N.J., Segault,V., Charpentier,B., Nottrott,S., Fabrizio,P., Bachi,A., Wilm,M., Rosbash,M., Branlant,C. and Luhrmann,R. (2000) A common core RNP structure shared between the small nucleoar box C/D RNPs and the spliceosomal U4 snRNP. *Cell*, **103**, 457–466.
3. 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.
4. Bachellerie,J.P., Cavaille,J. and Huttenhofer,A. (2002) The expanding snoRNA world. *Biochimie*, **84**, 775–790.
5. Jady,B.E., Darzacq,X., Tucker,K.E., Matera,A.G., Bertrand,E. and Kiss,T. (2003) Modification of Sm small nuclear RNAs occurs in the nucleoplasmic Cajal body following import from the cytoplasm. *EMBO J.*, **22**, 1878–1888.
6. Marker,C., Zemann,A., Terhorst,T., Kiefmann,M., Kastenmayer,J.P., Green,P., Bachellerie,J.P., Brosius,J. and Huttenhofer,A. (2002) Experimental RNomics: identification of 140 candidates for small non-messenger RNAs in the plant Arabidopsis thaliana. *Curr. Biol.*, **12**, 2002–2013.
7. Yang,J.H., Zhang,X.C., Huang,Z.P., Zhou,H., Huang,M.B., Zhang,S., Chen,Y.Q. and Qu,L.H. (2006) snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. *Nucleic Acids Res.*, **34**, 5112–5123.
8. 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.
9. Schattner,P., Barberan-Soler,S. and Lowe,T.M. (2006) A computational screen for mammalian pseudouridylation guide H/ACA RNAs. *RNA*, **12**, 15–25.
10. Lowe,T.M. and Eddy,S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, **283**, 1168–1171.
11. 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.
12. Chen,C.L., Chen,C.J., Vallon,O., Huang,Z.P., Zhou,H. and Qu,L.H. (2008) Genomewide analysis of box C/D and box H/ACA snoRNAs in Chlamydomonas reinhardtii reveals an extensive organization into intronic gene clusters. *Genetics*, **179**, 21–30.
13. Wold,B. and Myers,R.M. (2008) Sequence census methods for functional genomics. *Nat. Methods*, **5**, 19–21.
14. German,M.A., Pillay,M., Jeong,D.H., Hetawal,A., Luo,S., Janardhanan,P., Kannan,V., Rymarquis,L.A., Nobuta,K., German,R. *et al.* (2008) Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat. Biotechnol.*, **26**, 941–946.
15. Addo-Quaye,C., Eshoo,T.W., Bartel,D.P. and Axtell,M.J. (2008) Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome. *Curr. Biol.*, **18**, 758–762.
16. Fahlgren,N., Howell,M.D., Kasschau,K.D., Chapman,E.J., Sullivan,C.M., Cumbie,J.S., Givan,S.A., Law,T.F., Grant,S.R., Dangl,J.L. *et al.* (2007) High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes. *PLoS ONE*, **2**, e219.
17. Rajagopalan,R., Vaucheret,H., Trejo,J. and Bartel,D.P. (2006) A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes Dev.*, **20**, 3407–3425.
18. Ender,C., Krek,A., Friedlander,M.R., Beitzinger,M., Weinmann,L., Chen,W., Pfeffer,S., Rajewsky,N. and Meister,G. (2008) A human snoRNA with microRNA-like functions. *Mol. Cell*, **32**, 519–528.
19. Howell,M.D., Fahlgren,N., Chapman,E.J., Cumbie,J.S., Sullivan,C.M., Givan,S.A., Kasschau,K.D. and Carrington,J.C. (2007) Genome-wide analysis of the RNA-dependent RNA polymerase6/dicer-like4 pathway in Arabidopsis reveals dependency on miRNA- and tasiRNA-directed targeting. *Plant Cell*, **19**, 926–942.
20. Kasschau,K.D., Fahlgren,N., Chapman,E.J., Sullivan,C.M., Cumbie,J.S., Givan,S.A. and Carrington,J.C. (2007) Genome-wide profiling and analysis of Arabidopsis siRNAs. *PLoS Biol.*, **5**, e57.
21. Axtell,M.J., Jan,C., Rajagopalan,R. and Bartel,D.P. (2006) A two-hit trigger for siRNA biogenesis in plants. *Cell*, **127**, 565–577.
22. Lister,R., O'Malley,R.C., Tonti-Filippini,J., Gregory,B.D., Berry,C.C., Millar,A.H. and Ecker,J.R. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**, 523–536.
23. Gregory,B.D., O'Malley,R.C., Lister,R., Urich,M.A., Tonti-Filippini,J., Chen,H., Millar,A.H. and Ecker,J.R. (2008) A link between RNA metabolism and silencing affecting Arabidopsis development. *Dev. Cell*, **14**, 854–866.
24. 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.
25. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
26. Wang,B.B. and Brendel,V. (2004) The ASRG database: identification and survey of Arabidopsis thaliana genes involved in pre-mRNA splicing. *Genome Biol.*, **5**, R102.
27. Piekna-Przybylska,D., Decatur,W.A. and Fournier,M.J. (2007) New bioinformatic tools for analysis of nucleotide modifications in eukaryotic rRNA. *RNA*, **13**, 305–312.
28. 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.
29. Watkins,N.J., Dickmanns,A. and Luhrmann,R. (2002) Conserved stem II of the box C/D motif is essential for nucleolar localization and is required, along with the 15.5K protein, for the hierarchical assembly of the box C/D snoRNP. *Mol. Cell Biol.*, **22**, 8342–8352.
30. 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.
31. Chanfreau,G., Legrain,P. and Jacquier,A. (1998) Yeast RNase III as a key processing enzyme in small nucleolar RNAs metabolism. *J. Mol. Biol.*, **284**, 975–988.
32. Qu,L.H., Henras,A., Lu,Y.J., Zhou,H., Zhou,W.X., Zhu,Y.Q., Zhao,J., Henry,Y., Caizergues-Ferrer,M. and Bachellerie,J.P. (1999) Seven novel methylation guide small nucleolar RNAs are processed from a common polycistronic transcript by Rat1p and RNase III in yeast. *Mol. Cell Biol.*, **19**, 1144–1158.
33. Barneche,F., Steinmetz,F. and Echeverria,M. (2000) Fibrillarin genes encode both a conserved nucleolar protein and a novel small nucleolar RNA involved in ribosomal RNA methylation in Arabidopsis thaliana. *J. Biol. Chem.*, **275**, 27212–27220.
34. Barneche,F., Gaspin,C., Guyot,R. and Echeverria,M. (2001) Identification of 66 box C/D snoRNAs in Arabidopsis thaliana: extensive gene duplications generated multiple isoforms predicting

new ribosomal RNA 2′-O-methylation sites. *J. Mol. Biol.*, **311**, 57–73.
35. Westin,G., Lund,E., Murphy,J.T., Pettersson,U. and Dahlberg,J.E. (1984) Human U2 and U1 RNA genes use similar transcription signals. *EMBO J.*, **3**, 3295–3301.
36. Tycowski,K.T., Aab,A. and Steitz,J.A. (2004) Guide RNAs with 5′ caps and novel box C/D snoRNA-like domains for modification of snRNAs in metazoa. *Curr. Biol.*, **14**, 1985–1995.
37. 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.
38. 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.
39. 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.
40. Massenet,S., Mougin,A. and Branlant,C. (1998) Posttranscriptional modifications in the U small nuclear RNAs. In Grosjean,H. and Benne,R. (eds), *Modification and Editing of RNA*. ASM Press, Washington DC, pp. 201–227.
41. 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.
42. Kishore,S. and Stamm,S. (2006) The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science*, **311**, 230–232.