A computational screen for mammalian pseudouridylation guide H/ACA RNAs

PETER SCHATTNER,[1] SERGIO BARBERAN-SOLER,[2] and TODD M. LOWE[1]

Departments of [1]Biomolecular Engineering and [2]Molecular, Cell, and Developmental Biology, UCSC RNA Center,
University of California–Santa Cruz, Santa Cruz, California 95064, USA

ABSTRACT

The box H/ACA RNA gene family is one of the largest non-protein-coding gene families in eukaryotes and archaea. Recently, we developed snoGPS, a computational screening program for H/ACA snoRNAs, and applied it to *Saccharomyces cerevisiae.* We report here results of extending our method to screen for H/ACA RNAs in multiple large genomes of related species, and apply it to the human, mouse, and rat genomes. Because of the 250-fold larger search space compared to *S. cerevisiae*, significant enhancements to our algorithms were required. Complementing extensive cloning experiments performed by others, our findings include the detection and experimental verification of seven new mammalian H/ACA RNAs and the prediction of 23 new H/ACA RNA pseudouridine guide assignments. These assignments include four for H/ACA RNAs previously classified as orphan H/ACA RNAs with no known targets. We also determined systematic syntenic conservation among human and mouse H/ACA RNAs. With this work, 82 of 97 ribosomal RNA pseudouridines and 18 of 32 spliceosomal RNA pseudouridines in mammals have been linked to H/ACA guide RNAs.

Keywords: **computational gene finding; H/ACA RNAs; pseudouridylation guides; comparative genomics**

INTRODUCTION

The family of box H/ACA RNA genes plays an important role in guiding the modification of RNA uridines into pseudouridines (Ψ) (for reviews, see Bachellerie et al. 2002; Kiss 2002; Bertrand and Fournier 2004). The H/ACA RNA family includes small nucleolar RNAs (snoRNAs) that guide eukaryotic ribosomal RNA (rRNA) Ψ formation, small Cajal body RNAs (scaRNAs) that guide the formation of eukaryotic spliceosomal nuclear RNA (snRNAs) Ψs (Darzacq et al. 2002), as well as a homologous class of RNAs in archaeal organisms (Tang et al. 2002). In the past, identifying all, or even most, H/ACA RNAs for a new species has not been possible. Recently, however, systematic experimental approaches have successfully detected numerous H/ACA RNAs (Huttenhofer et al. 2001; Kiss et al. 2004), although these methods are unlikely to be applied beyond selected model organisms because they are costly and labor intensive. Moreover, these techniques have difficulty in detecting weakly expressed H/ACA RNAs and provide only limited information regarding the identities of the cognate Ψ-sites. Consequently, the development of computational methods for identifying potential H/ACA RNAs and their associated targets remains important.

However, the computational identification of H/ACA guide RNAs is challenging. The H/ACA RNAs have two primary sequence motifs, the "H-box" and the "ACA-box" as well as a characteristic hairpin–hinge–hairpin–tail secondary structure (Kiss 2002). One or both hairpin stems contain an internal loop with sequences that base-pair for 3–10 nt on each side of the uridine to be modified in the target RNA (Kiss 2002). Developing an effective algorithm to computationally identify H/ACA RNAs is complicated by the facts that the H/ACA motifs are short, the H/ACA hairpin secondary structures exhibit considerable variation, and the target-guide duplexes are of varying lengths and may be imperfectly paired.

Recently, we developed a computer program, "snoGPS," capable of detection of H/ACA snoRNA genes in genomic sequences (Schattner et al. 2004). Using snoGPS, we were able to identify six new, experimentally verified H/ACA snoRNAs and make 14 new associations of rRNA Ψs with H/ACA RNA guides in *Saccharomyces cerevisiae*. In the present work, we have extended our methods to search for H/ACA RNAs in mammalian (human, mouse, and rat) genomes. Because mammalian genomes are significantly larger than yeast genomes (3 Gb vs. 12 Mb), we needed to add several new

Reprint requests to: Peter Schattner and Todd M. Lowe, Department of Biomolecular Engineering, UCSC RNA Center, University of California–Santa Cruz, 1156 High St., Santa Cruz, Santa Cruz, CA 95064, USA; e-mail: schattner@soe.ucsc.edu; lowe@soe.ucsc.edu.

Article and publication are at http://www.rnajournal.org/cgi/doi/10.1261/rna.2210406.

capabilities to our search algorithms. These enhancements included the use of whole-genome alignments, the scoring of cross-species sequence and secondary-structure conservation, and the use of high-speed, parallel computation. Here we demonstrate that with these enhancements, we are able to computationally detect H/ACA RNAs in mammalian-sized genomes. Our results include the detection and experimental verification of seven new mammalian H/ACA RNAs. We also propose 23 new assignments of H/ACA RNA guides to rRNA or snRNA Ψs, including four assignments for H/ACA RNAs for which no Ψ target had previously been identified. Our results indicate that computational identification of H/ACA RNAs is feasible in medium to large genomes when at least one comparative genome is known.

RESULTS

A computational screen can detect H/ACA RNAs in mammalian genomes

To identify candidate H/ACA RNA genes, we initially screened selected regions of the human and mouse genomes using the snoGPS program. For each hit, we evaluated the corresponding region in the other two mammalian species, again using snoGPS. This cross-species implementation of snoGPS, which we call snoGPS-C, is outlined in Figure 1 and is described in detail in the Materials and Methods section.

We applied snoGPS-C to 93 previously known H/ACA RNAs using all known mammalian rRNA and snRNA Ψ-sites as targets. This procedure was not intended to yield a rigorous estimate of algorithm sensitivity because data from all known H/ACA RNAs were used to train the algorithm. However, the procedure was important to establish cutoff scores for candidate sequences as well as providing estimates of expected snoGPS-C performance in searches for new H/ACA RNAs. Ultimately, we assessed the sensitivity of snoGPS-C only on the basis of our experimental tests of new predictions.

SnoGPS-C calibration tests showed that 44 of the previously known H/ACA RNAs (47%) had the highest snoGPS-C score within our data set of highly conserved human non-protein-coding sequences (the "hmr20 data set"; see Materials and Methods for definition) for at least one Ψ. In addition, 20 other H/ACA RNAs were among the second to fifth highest snoGPS-C scores in the hmr20 data set for at least one Ψ. In all, 64 H/ACA RNAs (69%) were detected by snoGPS-C—that is, they scored among the top five hits in the hmr20 data set for at least one target. In contrast, when single-genome snoGPS was applied to the human genome alone, only 44 (47%) of the previously known H/ACA RNAs ranked among the five highest snoGPS

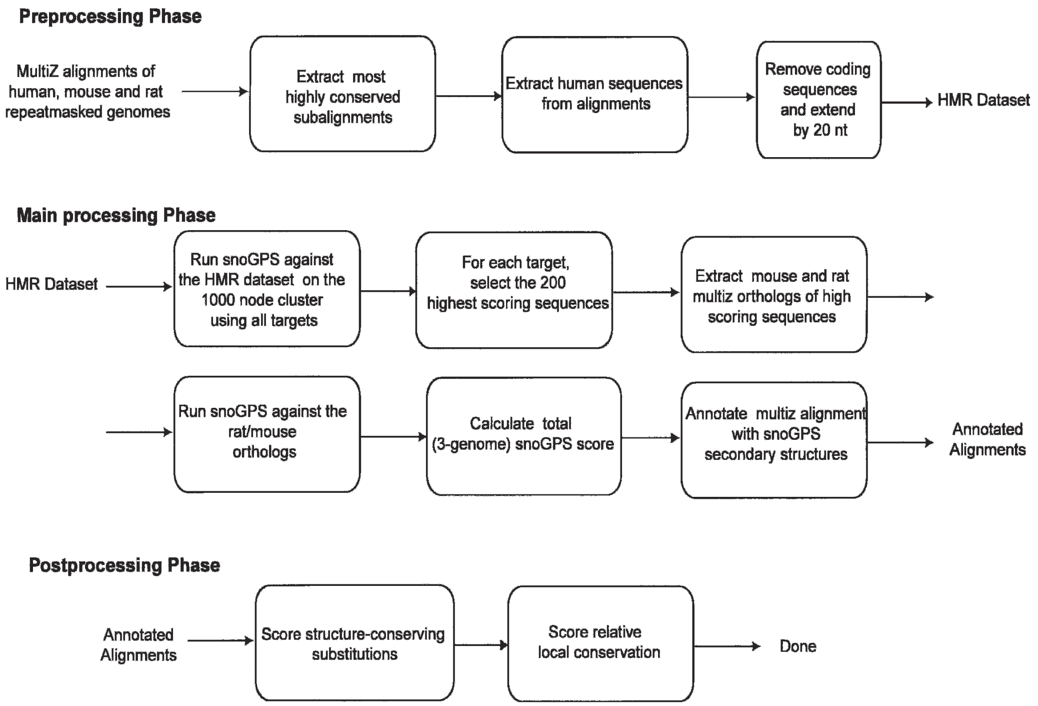
Preprocessing Phase
MultiZ alignments of human, mouse and rat repeatmasked genomes
Extract most highly conserved subalignments
Extract human sequences from alignments
Remove coding sequences and extend by 20 nt
HMR Dataset
Main processing Phase
HMR Dataset
Run snoGPS against the HMR dataset on the 1000 node cluster using all targets
For each target, select the 200 highest scoring sequences
Extract mouse and rat multiz orthologs of high scoring sequences
Run snoGPS against the rat/mouse orthologs
Calculate total (3-genome) snoGPS score
Annotate multiz alignment with snoGPS secondary structures
Annotated Alignments
Postprocessing Phase
Annotated Alignments
Score structure-conserving substitutions
Score relative local conservation
Done

FIGURE 1. Flowchart of the snoGPS-C algorithm. The algorithm is divided into three main stages. The initial stage is the generation of the most highly conserved non-protein-coding human and mouse sequences. The second stage involves running snoGPS on one of these two data sets and extracting the top 200 candidate sequences for each Ψ target. The final stage consists of extracting multiZ homologs of each of these sequences, testing each of these sequences with snoGPS, and averaging the human, mouse, and rat snoGPS scores.

scores for at least one target. This observation provided support for our decision to rank candidate sequences by composite snoGPS-C scores rather than by their single-genome snoGPS scores.

The 29 H/ACA RNAs that were not among the five highest scoring sequences for any target (and which consequently were not detected) were missed either because they have no known target site; because they are not well conserved among human, mouse, and rat; or because they lack some canonical H/ACA feature. Five (ACA11, ACA33, ACA39, ACA49, and ACA53) appear to be orphan H/ACA RNAs with no rRNA or snRNA target. Two others (ACA4 and ACA46) guide Ψs that were experimentally identified only after these tests were performed (Kiss et al. 2004). Five H/ACA RNAs (MBI-114, ACA21/MBI-3, ACA36/MBI-87, ACA2a/MBI-137, and U71a) were ranked among the top five snoGPS candidates in the human genome for at least one Ψ, but were missed by the snoGPS-C comparative genomic screen. In three of these cases (MBI-114, ACA21/MBI-3, ACA36/MBI-87), the multiZ whole-genome alignment procedure (Blanchette et al. 2004) was unable to identify a complete homologous sequence in mouse and rat and consequently snoGPS-C testing could not be performed at all. In the cases of ACA2a/MBI-137 and U71a, multiZ mouse and rat homologs were identified but scored poorly under snoGPS-C because they did not completely conserve some H/ACA RNA feature. Finally, in 17 cases (18%)—ACA10, ACA24, ACA29/MBI-39, ACA32, ACA35, ACA38, ACA41/MBI-83, ACA42, ACA48, ACA56, ACA59, ACA60, ACA61/MBI-164, U66, U70, U87, and U88—some atypical feature caused both snoGPS and snoGPS-C to score the H/ACA RNA lower than the top five hmr20 sequences for all Ψs. Examples of such unusual features included insertions/deletions, short terminal loop lengths, short (2 nt) left- or right-target-guide base-pairings, or target-guide base-pairings that contained multiple GU pairs. These omissions may indicate limitations of the current snoGPS H/ACA model. On the other hand, almost no guide-target assignments have been experimentally verified in mammals. Consequently, the fact that snoGPS "missed" a known H/ACA RNA may simply indicate that none of the true Ψ targets of this RNA have been identified yet. Figure 2 presents a graphical summary of the fraction of previously known H/ACA RNAs found by snoGPS-C. Scores and annotated alignments of the previously known H/ACA RNAs can be found in the Supplemental Material Files 1 and 2 (http://www.soe.ucsc.edu/~lowe/pubs/SupMat/RNA2005), respectively.

Most of the known H/ACA RNAs were found to have composite snoGPS-C scores >35 bits and target-guide complementarity 'pair scores' (see Materials and Methods for definition) in at least two mammalian species ≥8.7. Consequently, these values were typically used as minimum scores when evaluating new candidate sequences. In some cases (e.g., if the candidate sequence was located near a known snoRNA or in a ribosomal protein intron, or if its alignment had an unusually high number of substitutions that preserved the predicted secondary structure), candidates with lower overall scores were also considered.

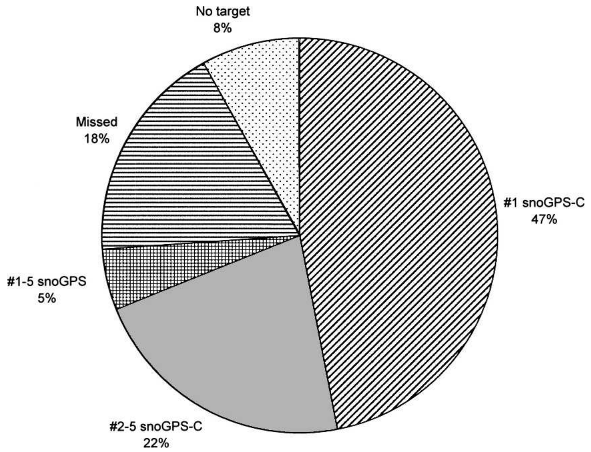

FIGURE 2. Fraction of previously known H/ACA RNAs detected by snoGPS and snoGPS-C. An H/ACA RNA was considered "detected" if it was among the five highest scoring hmr20 sequences for at least one rRNA or snRNA Ψ. The "No target" sequences included H/ACA RNAs with no predicted target as well as those whose associated Ψ was only experimentally located subsequent to these studies. The "#1–5 snoGPS" sequences include those that ranked among the five snoGPS-highest-scoring human sequences for at least one target, but that were not detected using snoGPS-C composite scores. "Missed" sequences did not rank among the top five for any target either for snoGPS or snoGPS-C.

New H/ACA RNAs detected by snoGPS are verified with Northern and primer extension analysis

To search for new H/ACA RNAs, we screened the most highly conserved non-protein-coding sequences between human and mouse, as well as the 3-kb regions adjacent to each of the known snoRNAs. All known mammalian rRNA and snRNA Ψs were used as targets. Our computational searches resulted in 45 candidate sequences (sequences and annotated alignments are available in Supplemental Material Files 3 and 4, http://www.soe.ucsc.edu/~lowe/pubs/SupMat/RNA2005). During the course of our computational screens, but before we performed any experimental verifications, we became aware of recent results of Kiss and colleagues (Kiss et al. 2004), which identified 10 of our candidates as H/ACA RNAs. The remaining 35 candidates were divided on the basis of snoGPS-C scores and putative secondary-structure conservation into a group of 11 top candidates and 24 secondary candidates.

The 11 top candidates were tested experimentally in mouse by Northern blot and primer extension analyses, resulting in the verification of seven new H/ACA RNAs (ACA62–ACA68) (see Fig. 3; Tables 1, 2). Base-pairings of the newly confirmed H/ACA RNAs with their putative Ψ targets are shown in Figure 4A. The four candidates with

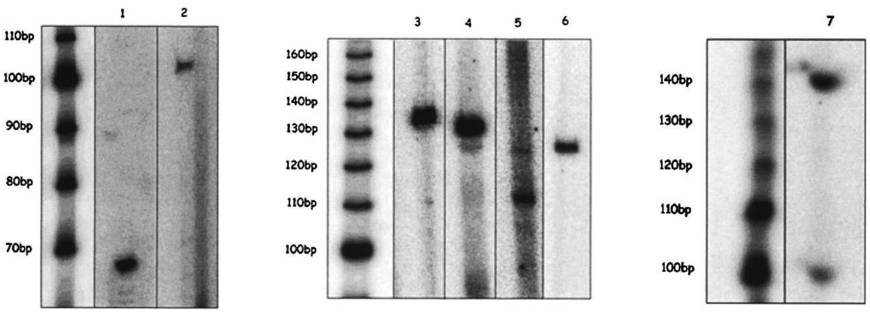

FIGURE 3. Experimental verification of predicted guide RNAs by primer extension. (Lanes *1*–*7*) ACA63, ACA65, ACA67, ACA64, ACA68, ACA66, and ACA62, respectively. Experimentally determined 5′-ends were always within 7 nt of predicted 5′-ends.

negative or inconsistent Northern results and the 24 untested candidates had lower snoGPS-C scores (average scores of 39.1 and 39.8, respectively) than the seven H/ACA RNAs that were experimentally confirmed (average snoGPS-C score of 53.4). This suggests that the identification of the human–rodent conserved H/ACA RNAs with known target sites may be close to completion. It is formally possible that some of our negative experimental results were due to the RNAs being expressed only at low levels or in tissue types other than those tested. Moreover, from manual inspection of their annotated alignments, we expect that the untested candidates include additional H/ACA RNAs. For example, Figure 5A shows the alignment of the 3′-hairpin subunit of the homologous human, mouse, rat, and chicken sequences of one untested candidate. The alignment shows four substitutions that conserve the proposed secondary structure, five others that are outside the proposed structure, and only a single substitution that conflicts with the structure. Supplemental Material File 3 (http://www.soe.ucsc.edu/~lowe/pubs/SupMat/RNA2005) lists the negative and untested candidate sequences.

As with the previously identified mammalian H/ACA RNAs (Kiss et al. 2004), the newly identified RNAs are located in introns of known genes or spliced expressed sequence tags (ESTs). Interestingly, ACA64 is also located only 500 nt from the 5′-end of the ribosomal protein RPS2, which has two other H/ACA RNAs, U64 and ACA10, encoded in its introns. However, ACA64 and RPS2 are encoded on opposite strands, making the biological significance of their proximity, if any, unclear.

snoGPS predicts guide assignments of H/ACA RNAs to rRNA and snRNAΨs

Our analyses predict 23 new associations of H/ACA RNAs to rRNA or snRNA Ψs. Table 3 lists these predicted assignments with their associated target-guide pair scores in human, mouse, and rat. Figure 4 shows the predicted target-guide base-pairings in human for these assignments. All assignments have target-guide base-pairings of at least 9 bp in at least two of the three genomes. We chose a minimum of nine pairings since experimentally verified target-guide associations with as few as 9 bp are known in yeast (Schattner et al. 2004). Of the sites with new guide assignments, 10 had not been previously assigned. The remaining 13 Ψs had previously been assigned to other H/ACA RNAs (Ganot et al. 1997; Huttenhofer et al. 2001; Kiss et al. 2004); snoGPS-C identified these assignments as well, although generally with lower scores (data not shown). Four of our assignments were for H/ACA RNAs (ACA54, ACA55, U99, and U100) that had previously been classified as "orphan snoRNAs" with no known Ψ target site (Kiss et al. 2004). U100 has been predicted to perform some guiding function at position U9 in the U6 snRNA (Vitali et al. 2003); however, this uridine is not pseudouridylated (Vitali et al. 2003). Interestingly, our assignments include predictions that U93 guides U5-Ψ53 and U100 guides U2-Ψ7, which is consistent with U93 (Kiss et al. 2002) and U100 (Vitali et al. 2003) having previously been demonstrated to be scaRNAs. With these new associations, 82 of the 97 Ψs in mammalian rRNA and 18 of 32 Ψs in mammalian snRNA now have putative guide H/ACA RNA assignments.

TABLE 1. snoGPS-C scores of new H/ACA RNAs

RNA	Target	snoGPS-C score	aRk	hsScore	hRk	rnScore	rnScore	Hs pScore
ACA62	18S.Ψ105	64.99	1	63.37	1	62.97	68.62	12.0
ACA63	18S.Ψ814	53.64	1	52.67	1	53.92	54.34	9.0
ACA64	23S.Ψ4321	47.51	1	47.19	5	48.62	46.71	10.0
ACA65	U6.Ψ86	52.92	1	55.15	1	50.89	52.72	13.0
ACA66	U12.Ψ28	47.95	1	51.17	1	42.75	49.94	11.0
ACA67	18S.Ψ109	49.58	1	47.63	5	53.38	47.72	11.0
ACA68	U12.Ψ19	57.44	1	61.07	1	59.13	52.11	10.0

snoGPS-C scores of experimentally confirmed H/ACA RNAs and the ranks of those scores in the hmr20 data set (aRk). The table also indicates the predicted Ψ-site, the human, mouse, and rat single-genome snoGPS scores (hsScore, mmScore, and rnScore, respectively), the rank of the human snoGPS score (hRk), and the pair score between the H/ACA RNA guide domain and the Ψ target in *H. sapiens* (Hs pScore).

TABLE 2. Locations of the new H/ACA RNAs

RNA	Ch	St	Predicted location	Predicted length	Host gene	Experimental size estimate (nt)
ACA62	11	W	107,344,896–107,345,028	133	EST	140
ACA63	16	C	17,713,849–17,713,974	126	Ranbp 1	125
ACA64	17	C	23,408,657–23,408,784	128	EST	133
ACA65	12	C	85,520,378–85,520,520	143	Gtf2a 1	140
ACA66	5	C	132,661,784–132,661,912	129	Wbscr2 2	131
ACA67	5	W	118,530,864–118,530,997	134	EST	135
ACA68	11	C	70,009,865–70,010,008	144	EST	147

Predicted and experimentally determined genomic coordinates in mouse (using mouse database mm3, NCBI build 30, February 2003) of the newly verified H/ACA RNAs. "Ch" and "St" indicate chromosome and strand, respectively. For each H/ACA RNA, the host gene of the RNA is shown as well. "EST" indicates that the RNA is located in an intron of an unnamed spliced EST. Experimental size estimates were determined from the primer extension results and the position of the predicted ACA motif.

Motifs and secondary structure are highly conserved among homologous mammalian H/ACA RNAs

One of the most striking features to emerge from our analyses was the high degree of motif and secondary-structure conservation among the mammalian H/ACA RNA homologs as defined by multiZ alignments. Of 44 H/ACA RNAs that snoGPS ranked among the top five in the human genome for at least one target, 38 (86%) were also detected in both rodent multiZ homologs. In three of the six remaining cases (ACA21, ACA36, and MBI-114), multiZ was unable to align a homolog in either rodent genome. For the other three cases (U71a, ACA2a, and ACA48), the homologous sequence(s) identified in mouse or rat identified by multiZ had gross violations of H/ACA motifs and/or secondary structure. It is possible that rodent homologs of these H/ACA RNAs are no longer present. It is also possible that multiZ was simply unable to detect them. Consequently, we also searched for rodent homologs with genome-wide BLAT (Kent 2002). However, we were unable to find any alternative structure-conserving homologs by this approach either.

We found a few additional examples in which the multiZ mouse or rat homolog of a human H/ACA RNA did not conserve an important H/ACA RNA motif. One striking example is ACA50. Figure 5B shows an annotated alignment of ACA50. Structure- and motif-violating substitutions, shown in red, include loss of the "ACA" motif, a deletion and loss of base-pairing in the 3′-hairpin, and loss of base-pairing in the Ψ-guiding domain in rat. Another example is U98/ACA16, for which the rat homolog has a single nucleotide deletion destroying the "H" motif as well as four substitutions in the Ψ-guiding domain (alignment in Supplemental Material File 2, http://www.soe.ucsc.edu/~lowe/pubs/SupMat/RNA2005). Three additional H/ACA RNAs (U71, ACA41/MBI-83, ACA56) have substitutions destroying functional motifs in mouse and/or rat Ψ-guiding domains only. If these motif-violating sequences are the correct homologs and the apparent motif violations are not the result of sequencing errors, it will be interesting to determine whether these H/ACA RNAs are nevertheless expressed, transported to the nucleolus, and function as guides for the predicted pseudouridine modifications.

Most homologous H/ACA RNAs are syntenic in human and mouse and have homologous host genes

We sought to determine whether multiZ H/ACA RNA homologs are syntenic, that is, whether their neighboring genomic regions are also homologous between human and mouse. We determined synteny using the chain-net algorithm (Kent et al. 2003) as implemented on the "net" track of the UCSC Genome Browser. This algorithm has the advantage of determining synteny solely on the basis of genomic sequence alignment without requiring gene annotation data (Kent et al. 2003). Using this approach, we determined that 89 of 105 H/ACA RNAs (85%) are syntenic between human and mouse. In the remaining cases, the chain-net algorithm indicated a "microrearrangement" as a result of which the homologous human and mouse H/ACA RNAs (generally along with their host genes) were embedded in genomic sequence that was not homologous between the two species.

We also investigated whether the host genes of homologous H/ACA RNAs are themselves homologous. Because many mammalian genes have functional and/or pseudogene paralogs, determining the correct mouse homolog to a human host gene can be difficult. To minimize the effect of ambiguous homologs, we restricted our analysis to 69 human H/ACA RNAs with protein-coding host genes in SWISS-PROT and Trembl that were not annotated as "fragments," "hypothetical," or "open reading frames." For these host genes, we searched for mouse homologs with two different methods: BLASTP, using the UCSC genome database table of best BLASTP hits to known genes (Altschul et al. 1997) and translated BLAT (Kent 2002), using a cutoff of 60% identity. (The low BLAT cutoff score was used to detect possible cases in which the mouse or human host gene had been duplicated and subsequently diverged in sequence from the host gene of the H/ACA RNA in the other species.)

With this procedure, mouse homologs of 84% of human H/ACA RNAs (59/69) were found to be located in mouse homologs of the human host genes. For five human H/ACA RNAs—ACA16 (PNAS-123), ACA61 (PNAS-123), ACA44 (PNAS-123), ACA13 (TIGA1), and ACA47 (PRO0872) (human host genes shown in parentheses)—neither translated BLAT nor BLASTP identified a homologous gene in mouse to the human host gene. For the five remaining

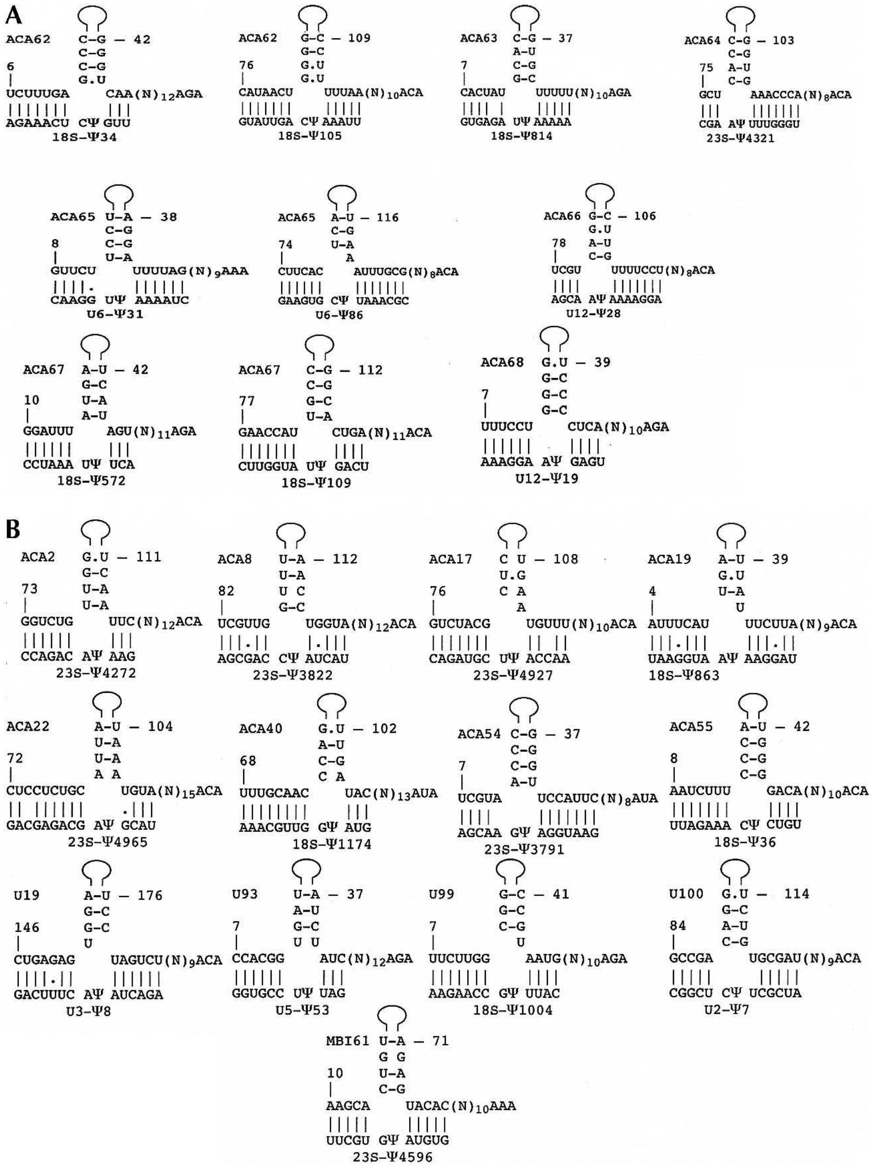

FIGURE 4. Putative base-pairings between Ψ target regions and H/ACA RNAs. Pairings are shown for (*A*) the newly identified Ψ guide H/ACA RNAs; (*B*) new assignments of previously known H/ACA RNAs.

H/ACA RNAs— ACA37 (NBD2), ACA58 (MRPL3), ACA33 (RPS12), ACA63 (ATP2B4), ACA66 (USP32)—homologous regions in mouse of the human host genes were found at locations different from those of the embedded H/ACA RNA. For these cases, we searched for possible mouse H/ACA RNAs in the introns of the mouse homologs of the human host genes with snoGPS and snoGPS-C. No strong H/ACA candidates were found, suggesting that these H/ACA RNAs had, indeed, changed host genes since the human–mouse divergence. Figure 6 illustrates one such H/ACA RNA with apparently nonhomologous hosts in human and mouse. From the figure, one sees that the mouse homolog of the human host (USP32) of ACA66 is on Chromosome 11, while the multiZ homolog of ACA66 itself is on Chromosome 5 (where it is located in an intron of Wbscr22).

For comparison, we performed an identical analysis of 86 C/D-box RNAs and their host genes derived from the snoRNA track of the UCSC Genome Browser. The results were quite similar to the H/ACA RNA case. All but five (U96b, U105, HBI-420, HBII-13, HBII-436) of 86 C/D snoRNAs (94%) have homologous host genes in human and mouse. In conclusion, we find that the large majority of H/ACA- and C/D-box RNAs appear to have homologous hosts in both human and mouse.

Distinct H/ACA RNAs may guide identical pseudouridine pairs

Earlier studies have shown that several mammalian H/ACA RNAs have multiple close paralogs (Kiss et al. 2004). Similarly, one of the new H/ACA RNAs, ACA67, also has two close paralogs in each of the human, mouse, and rat genomes, which we refer to as ACA67b and ACA67c. ACA67b and ACA67c have overall sequence similarities of 83% and 85% to ACA67 and conserve all the H/ACA motifs to guide 18S-Ψ109 and 18S-Ψ572 in all three genomes.

Surprisingly, we also found that two of the new H/ACA RNAs, ACA62 and ACA67, appear to have distant paralogs among previously identified RNAs: ACA50 and ACA42, respectively. For example, both ACA62 and ACA50 appear to guide Ψs 18S-Ψ34 and 18S-Ψ105. In principal, it is possible that ACA62, which we verified solely in mouse, is simply the mouse ortholog of ACA50, which has been confirmed only in human (Kiss et al. 2004). However, this seems very unlikely. ACA62 and ACA50 have only 65% overall sequence similarity. Moreover, ACA50 in human and its multiZ partner in mouse are syntenic (flanking genes are NDRG4 and GOT2), while ACA50 in human and ACA62 in mouse are not syntenic. Indeed, when ACA50 was initially identified, it was annotated as not having any paralogs (Kiss et al. 2004). Without experimental verification we cannot determine whether ACA50 or ACA62 (or both) actually guides 18S-Ψ34 and 18S-Ψ105. However, it is interesting to note that the multiZ rat homolog of ACA50 does not conserve several charac-

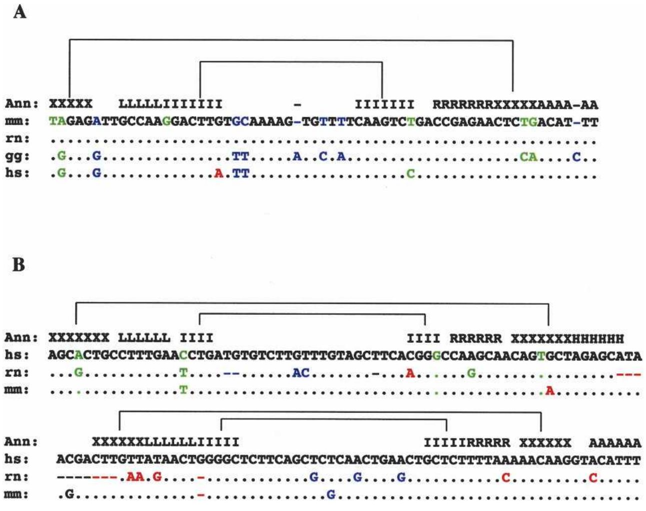
A
Ann: XXXXX   LLLLLIIIIIII   -   IIIIIII RRRRRRXXXXXAAAA-AA
mm: TAGAGATTGCCAAGGACTTGTGCAAAAG-TGTTTTCAAGTCTGACCGAGAACTCTGACAT-TT
rn: ....
gg: .G...G...............TT.....A..C.A....................CA....C..
hs: .G...G.............A.TT..................C.....................
B
Ann: XXXXXXX LLLLLL IIII   IIII RRRRRR XXXXXXXHHHHHH
hs: AGCACTGCCTTTGAACCTGATGTGTCTTGTTTGTAGCTTCACGGGCCAAGCAACAGTGCTAGAGCATA
rn: ...G...........T....--......AC.......-...A.....G...............---
mm: ...............T...A..........
Ann: XXXXXXLLLLLLIIIII   IIIIIRRRRR XXXXXX AAAAAA
hs: ACGACTTGTTATAACTGGGGCTCTTCAGCTCTCAACTGAACTGCTCTTTTAAAAACAAGGTACATTT
rn: -------.AA.G....-..............G....G....G...........C.........C.....
mm: .G...............-...............G................................

FIGURE 5. Annotated sequence alignments. Dots (. . .) in the alignments indicate bases that are identical to those in the reference sequence. The letters in the annotation string indicate H/ACA features identified in the sequence: (H and A) the H and ACA motifs, respectively. (L and R) The two components of the rRNA guide sequence. Bases annotated with X and I refer to the lower (or "external") and upper (or "internal") stems of either the 5′- or 3′-hairpins. (*A*) Annotated alignment of an unverified candidate sequence (S35) showing the 3′-hairpin of the human, mouse, rat, and chicken homologs. The alignment shows four substitutions (in green) that conserve the proposed secondary structure, seven others do not affect any of the proposed H/ACA motifs (in blue), and a single substitution (in red) that conflicts with the proposed H/ACA structure. (*B*) Alignment of ACA50 showing substitutions in the multiZ-identified rat homolog that violate the snoGPS-predicted RNA structure. Note the single nucleotide substitution in the ACA motif and the deletion in the 3′-hairpin in the rat sequence.

teristic H/ACA features (Fig. 5B), whereas the multiZ homologs of ACA62 do (alignments are available in the Supplemental Material File 4 http://www.soe.ucsc.edu/∼lowe/pubs/SupMat/RNA2005), suggesting that, at least in rat, ACA62 is the functional guide RNA.

The situation with ACA67 and ACA42 is similar to that between ACA62 and ACA50. Both ACA67 and ACA42 appear to guide 18S-Ψ109 and18S-Ψ572. However, their overall sequence identity is only 78%, and they are not syntenic between human and mouse. Interestingly, in this case as well, snoGPS-C alignments show that the H/ACA motif and secondary structure are not equally well conserved in human, mouse, and rat; namely, the mouse and rat homologs of ACA67 conserve all H/ACA motifs, while the mouse homolog of ACA42 does not preserve the H motif.

Characteristics of H/ACA-RNA host genes and introns

We sought to identify signatures that might characterize H/ACA RNA host genes or host-gene introns. To this end, we determined lengths of 96 introns containing known human H/ACA RNAs, as well as the total number of introns per host gene. We also investigated whether host introns of H/ACA RNAs are located preferentially near the 5′- or 3′-ends of the host genes and whether the H/ACA RNA was located at any characteristic position within its host intron.

Average length of introns containing H/ACA RNAs was 7282 nt (standard deviation = 32,599 nt, median length = 846 nt) as compared to an average length of 3365 nt (median = 1023 nt) for all human intron-containing genes (Lander et al. 2001). The average total number of introns in the H/ACA RNA host genes is 13 (standard deviation = 13, median number = 8). In contrast, the average number of introns for all human intron-containing genes is 7.8 (median = 6) (Lander et al. 2001). That is, H/ACA RNA hosts have somewhat more introns than average genes; however, because of the large standard deviations, intron number is not likely to be a significant signature in identifying H/ACA RNA host genes.

A histogram of distances of the 96 H/ACA RNAs to their upstream and downstream host-intron splice sites is shown in Figure 7A. From the figure, one sees that while H/ACA RNAs are usually located within 300 nt of their adjacent 3′- and 5′-splice sites, in numerous cases these distances can be 1 kb or more. This is in contrast to C/D-box RNAs that typically have distances of 65–100 nt to their downstream splice sites (Hirose and Steitz 2001). We also searched for systematic preferences for H/ACA RNAs to be in introns near the 5′- or 3′-ends of the host gene. As shown in Figure 7B, H/ACA RNAs are somewhat more commonly found toward the 3′-end of their host genes, but there are numerous exceptions as well. In summary, we were unable to identify any defining signatures of the host genes or their introns that would indicate that a specific gene or intron would be particularly likely to contain an H/ACA RNA gene.

DISCUSSION

Recently, we have demonstrated that a computational screen is capable of detecting H/ACA RNA sequences in the *S. cerevisiae* genome (Schattner et al. 2004). In the present work, we extend this screen to computationally detect mammalian H/ACA RNAs, including ones that have evaded large-scale experimental screens. We have also shown that, in most cases, H/ACA RNA motifs, predicted secondary structures, and host genes are well conserved among human, mouse, and rat.

With this work, we now have H/ACA RNA guide assignments of 100 of the 129 known mammalian rRNA and snRNA Ψs. Guides for the remaining 29 Ψs may not have been found because the associated H/ACA RNAs have atypical feature (e.g., large insertions/deletions) or because they have not been conserved in all three mammalian genomes. Some mammalian uridines may even be pseudouridylated by a guide-independent enzyme as occurs for U2-Ψ35 and U2-Ψ44 in yeast (Ma et al. 2003, 2005)

However, the statement that 100 of the 129 mammalian rRNA and snRNA Ψs are now assigned to a guide H/ACA

TABLE 3. Predicted Ψ guide assignments

Target	ID	h	Pr	snoGPS-C score	Rk	Hs pScore	Mm pScore	Rn pScore
18S-rRNA								
Ψ34	ACA62	5	HK	60.84	1	10.0	10.5	10.5
Ψ36	ACA55	5	G	45.40	1	11.0	11.0	14.5
Ψ105	ACA62	3	K	64.99	1	12.0	12.0	12.0
Ψ109	ACA67	3	K	49.58	1	11.0	11.0	11.0
Ψ572	ACA67	5	K	48.79	1	9.0	10.0	9.0
Ψ814	ACA63	5	HK	53.64	1	8.7	12.0	12.0
Ψ863	MBI51 ACA19	5	HK	36.48	2	12.0	12.0	12.0
Ψ1004	U99	5	K	41.03	1	11.0	11.0	11.0
Ψ1174	MBI142 ACA40	3	—	45.65	2	11.0	9.0	11.0
23S-rRNA								
Ψ3791	ACA54	5	—	49.78	1	9.7	9.7	9.7
Ψ3822	ACA8	3	G	38.22	3	10.0	10.0	10.0
Ψ4272	ACA2 MBI137	3	K	50.14	1	9.0	9.0	9.0
Ψ4321	ACA64	3	K	47.51	1	10.0	10.0	10.0
Ψ4596	MBI61	5	—	44.98	1	10.0	10.0	10.0
Ψ4927	MBI31 ACA17	3	K	46.42	2	10.0	10.0	10.0
Ψ4965	ACA22	3	G	40.60	2	10.2	9.5	9.5
snRNAs								
U2Ψ7	U100	3	—	37.67	1	9.0	9.0	9.0
U3Ψ8	U19	3	—	46.15	1	12.5	13.0	13.5
U5Ψ53	U93	5	—	53.60	1	9.0	9.0	9.0
U6Ψ31	ACA65	5	—	52.18	1	10.5	10.0	11.0
U6Ψ86	ACA65	3	—	52.92	1	13.0	12.0	12.0
U12Ψ19	ACA68	5	—	57.44	1	10.0	10.0	10.0
U12Ψ28	ACA66	3	—	47.95	1	11.0	7.7	10.0

For each Ψ guide assignment, the Ψ-site and the guide H/ACA RNA are listed. The column marked "h" indicates the H/ACA hairpin (5′ or 3′) that includes the predicted guide sequence. The "Pr" column indicates whether a different predicted assignment for the Ψ has been previously published: (G) Ganot et al. 1997; (H) Huttenhofer et al. 2001; (K) Kiss et al. 2004. The snoGPS-C score, the rank of that score for that site within the hmr20 data set ("Rk"), and the pair-scores in human (Hs), mouse (Mm), and rat (Rn) are also shown.

RNA should be viewed with caution. In contrast to the situation in yeast, where 43 of the 44 predicted guide-target assignments have been functionally verified (Schattner et al. 2004; Torchet et al. 2005), in mammals only two guide assignments have been functionally verified (Bortolin et al. 1999; Jady and Kiss 2001). From our experience in *S. cerevisiae*, we expect that in most cases, base-pairings of 9 or more nucleotides between a Ψ-site and an H/ACA RNA do indicate that the H/ACA RNA is involved in the formation of the associated Ψ; however, in yeast, exceptions to this rule exist as well (Schattner et al. 2004).

We have found 13 examples where—either because of longer target-guide base-pairings or because the base-pairings were more conserved among human, mouse, and rat—snoGPS-C assigned a guide RNA to a target that previously had been associated with a different H/ACA RNA (Ganot et al. 1997; Huttenhofer et al. 2001; Kiss et al. 2004). Moreover, there are already 12 other examples in the literature of mammalian Ψs (Ψ1723, Ψ3731, Ψ4512, and Ψ4956 on 23S-rRNA; Ψ34, Ψ105, Ψ814, Ψ863, Ψ1238, and Ψ1625 on 18S-rRNA; Ψ46 on U5 and Ψ40 on U6) that had been assigned to multiple H/ACA RNAs (Ganot et al. 1997; Huttenhofer et al. 2001; Kiss et al. 2004). Without experimental verification, one cannot determine which of these assignments, if any, is correct, or if some of these guide RNAs may be redundant. Recently, several in vitro H/ACA

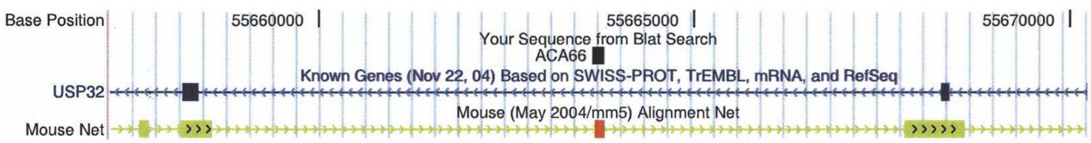
Base Position
55660000
55665000
55670000
Your Sequence from Blat Search
ACA66
Known Genes (Nov 22, 04) Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
USP32
Mouse (May 2004/mm5) Alignment Net
Mouse Net

FIGURE 6. ACA66 in human and mouse have nonhomologous host genes. Screen shot of the UCSC Genome Browser showing ACA66 with its host gene USP32 on *Homo sapiens* Chromosome 17 (NCBI Build hg17, May 2004). The yellow color along most of the mouse net alignment track indicates that the homologous region of USP32 in mouse is on Chromosome 11. The red color at the location of ACA66 in the mouse net alignment track indicates that the multiZ homolog of ACA66 itself is on Chromosome 5 (where it is located in an intron of Wbscr22).

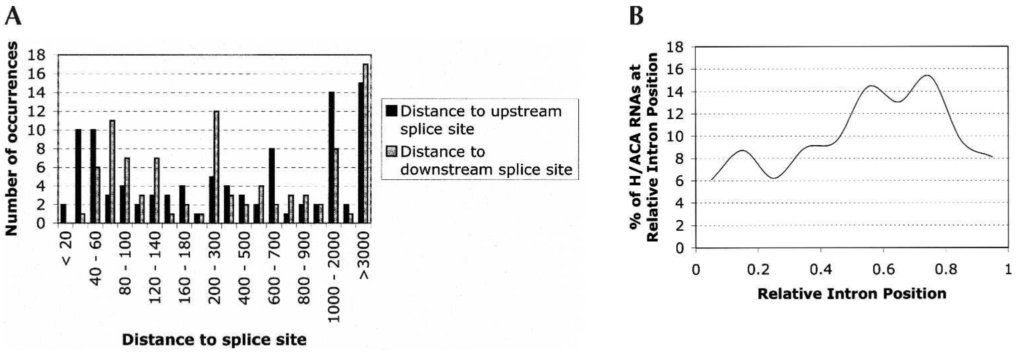

FIGURE 7. Positions of H/ACA RNAs within host introns and host genes. (*A*) Histogram of distances of H/ACA RNAs to their upstream and downstream host-intron splice sites. (*B*) Histogram of relative positions of introns containing H/ACA RNAs within their host genes. Relative positions near zero (one) indicate introns near the 5′ (3′) ends of the host gene, respectively.

RNAP reconstitution systems have become available, which should facilitate the experimental testing of H/ACA RNA guide assignments (Wang and Meier 2004; Baker et al. 2005; Charpentier et al. 2005; Ma et al. 2005). However, until such experimental verifications have been performed, one needs to be cautious in accepting target-guide assignments.

We have demonstrated that snoGPS can detect most H/ACA RNAs in small genomes (e.g., *S. cerevisiae*) and, when complemented with comparative genomic data, in large (e.g., mammalian) genomes as well (~75% of the known H/ACA RNAs were detected in the present study). For these studies, the locations of the target Ψ-sites were known. However, snoGPS can also search for H/ACA RNA sequences even when the target sites are not known. This can be done, for example, by using all rRNA and snRNA uridines as candidate Ψ-sites. Such a procedure does significantly add to the computational demands on the program and also increases the number of false-positive "hits" that will be produced, as there are, for example, ~14 times as many uridines as pseudouridines in mammalian rRNA and snRNA. Nevertheless, for small to intermediate size (e.g., <200 Mb) genomes, such an approach should be feasible with computer resources comparable to those used in the current analysis.

MATERIALS AND METHODS

Data sources

Genome sequence data were taken from the NCBI April 2003 (hg15), February 2003 (mm3), and June 2003 (rn3) assemblies of the human, mouse, and rat genomes, respectively. The search space was limited to the 20% of the human and mouse genomes that are most highly conserved among human, mouse, and rat using a data set kindly provided by M. Blanchette using the methods of Margulies and colleagues (Margulies et al. 2003). This data set was repeat-masked (http://ftp.genome.washington.edu/RM/RepeatMasker.html) and restricted to sequences not overlapping exons of known protein-coding genes or other known features in the UCSC Human Genome Browser database (Kent et al. 2002). Each of the resulting sequence fragments was extended by 100 nt because an snoRNA might not be entirely conserved and hence might be truncated in the set of conserved sequences. This process created two data sets (referred to as the hmr20 and mhr20 data sets for human and mouse, respectively), each with ~3 million sequences with average length of ~430 nt. In addition to scanning the hmr20 and mhr20 data sets, snoGPS searches were also performed on the genomic sequences 3000 nt upstream and downstream of the known H/ACA- and C/D-box RNAs (with less stringent score requirements). Analyses of H/ACA RNA synteny, host-gene conservation, and host-intron statistics were carried out using the May 2004 (hg17) human and (mm5) mouse NCBI assemblies on the UCSC database.

rRNA sequence and Ψ data were taken from the rRNA database (Wuyts et al. 2001, 2002) and the literature (Maden 1990; Ofengand and Fournier 1998). snRNA sequence and Ψ data were taken from the snRNA database (http://mbcr.bcm.tmc.edu/smallRNA) and Massenet et al. (1998). Sequences of 20 nt surrounding each of the known Ψs were extracted from the rRNA and snRNA sequences for use as targets for the snoGPS program. At the time when these studies were carried out, only 16 Ψs in the 18S subunit of mammalian rRNA were known to single-nucleotide resolution (Maden 1990). Another 14 sites were known to within 4-nt resolution (Maden 1990). In the latter cases, all uridines within the range of possible Ψ locations were used as potential targets. This resulted in a total of 105 rRNA target sites. rRNA Ψ numbering conventions used in the references Maden (1990), Ofengand and Fournier (1998), and Huttenhofer et al. (2001) were followed. Note that this numbering differs from that used in the references Ganot et al. (1997) and Kiss et al. (2004) by 4 nt in the 18S subunit of mammalian rRNA and by 9 to 11 nt in 28S rRNA. snRNA Ψ numbering follows that used in Massenet et al. (1998). Sequence and annotation data for the known H/ACA RNAs, which were used in program training, were taken from the literature (Ganot et al. 1997; Huttenhofer et al. 2001; Vitali et al. 2003; Kiss et al. 2004).

Algorithm description

The search algorithm consisted of scanning the hmr20 and mhr20 data sets with snoGPS and subsequently scanning the homologs of the highest scoring candidates in the other two mammalian species. Homologs were identified from whole-genome alignments generated with the multiZ program (Schwartz et al. 2003; Blanchette et al. 2004). Figure 1 depicts this approach schematically.

The snoGPS program itself has been described previously (Schattner et al. 2004). Most parameters used in the snoGPS program were identical to that used in the phase III of our earlier study (Schattner et al. 2004). However, the snoGPS scoring matrices were retrained with data from the mammalian H/ACA RNAs that were known at the start of this investigation (Ganot et al. 1997; Huttenhofer et al. 2001). In addition, the following modifications were incorporated into the snoGPS parameters and descriptor files:

1. A decrease in the minimum allowable distance between two halves of a target recognition motif from 25 nt to 20 nt.
2. A decrease in the minimum allowable distance from the start of the H motif to the start of the 3′-stem from 10 nt to 8 nt.
3. An increase in the maximum allowable size of insertions at base of hairpins from 2 nt to 3 nt.
4. Removal of the test for a T-rich region immediately 3′ to the candidate sequence.
5. Modification of the H/ACA-motif test for improved motif detection in sequences with multiple occurrences of potential H or ACA motifs.

Candidates were ranked on the basis of their overall snoGPS as well as their target-guide region "pair scores." These pair scores were calculated by counting complementary bases between the predicted H/ACA RNA guide region and the bases surrounding the target Ψ (Watson-Crick matches scoring +1, GU pairs scoring +0.5, and mismatches scoring −1.3). For computational efficiency, the snoGPS program was installed on the UCSC 1000 node Linux cluster. Running snoGPS with 105 Ψ targets against the hmr20 data set and mhr20 data sets each required ~2.5 yr of CPU time on the UCSC cluster and 24 h of total elapsed time.

For each target site, genomic coordinates of the 200 highest-scoring snoGPS candidates were used to extract multiZ-based orthologous sequences in the other two mammalian genomes from the UCSC browser databases (Kent et al. 2002). The multiZMm3Rn3Gg0 alignment table of the hg16 database (which also includes chicken sequences from early assemblies of the chicken genome) and the multiZHg15Rn3 alignment table of the mm3 database were used for mouse and human candidate sequences, respectively. Sequences for each of the species in the alignment, extended by an additional 20 nt in both the 5′ and 3′ directions, were extracted from the UCSC database (Kent et al. 2002). Each sequence was then tested with snoGPS, and a composite snoGPS-C score was calculated by averaging the snoGPS scores of the homologous sequences.

Homologous sequences that ranked among the top five in the genome for at least one target in human, mouse, and rat were aligned with the T-Coffee program (Notredame et al. 2000), and the alignment was annotated with the predicted H/ACA secondary structure. The aligned sequences were then scored for substitutions that conserved the proposed secondary structure and for local sequence conservation. Substitutions involving both members of a base pair (e.g., CG → AU) scored +1, while structure-preserving single-nucleotide substitutions (e.g., GU → GC) scored +0.5. Counts were normalized by dividing by the length of the conserved feature. Local structure conservation was computed from the ratio of the frequency of substitutions within the biologically important parts of the candidate sequence (i.e., the H, ACA motifs, the target-guide regions, and the hairpins) to the frequency of substitutions in the other parts of the sequence and in the regions 20 nt upstream and downstream of the candidate.

Experimental verification

Experimental verification of candidate hits was performed using Northern analysis and primer extensions. For Northern analysis, total RNA (25 μg) was separated on a denaturing polyacrylamide gel (7 M urea, 1× TBE), electrophoretically transferred onto a nylon membrane (Hybond-N$^+$; Amersham) and cross-linked to the membranes with ultraviolet light. The membranes were treated for 1 h at 42°C in the formamide buffer (50% [w/v] deionized formamide, 5× SSC, 5× Denhardts, 0.2% SDS 200 μg/mL of sheared Herring sperm DNA, and 50 mM Tris-HCl at pH 8.0). Membranes were probed for 20 h in formamide buffer at 42°C with ^{32}P end-labeled 60-nt oligonucleotides corresponding to each candidate. Post-hybridization washes were done at 42°C (twice with 2× SSPE and 0.5 SDS for 30 min). Hybridization patterns were determined using a PhosphorImager (Molecular Dynamics).

For 5′-end-mapping by primer extension analysis, total mouse embryo RNA (80 μg) was heat-denatured in the presence of 10^6 cpm of each 5′-end-labeled oligonucleotide. Reverse-transcription reactions were carried out in a final volume of 20 μL for 60 min at 42°C with the following components: 1× First-strand buffer, 0.01 M DTT, 0.5 mM dNTPs, and 200 U of Superscript II RT (Invitrogen). The reaction was inactivated by heating to 70°C. Samples were ethanol-precipitated, and half of the reaction was loaded onto an 8% sequencing gel (7 M urea, 1× TBE). A 10-bp ladder (Invitrogen) was used. Extension patterns were determined using a PhosphorImager (Molecular Dynamics). The primer sequences used in the primer extension analysis were:

ACA62: GCCTGTAGCTAGCGGTTAAAAGG;
ACA63: TGGACTGTTTCTTTGGCAGATGG;
ACA64: AATGTAGGGGCCTGGGTTTC;
ACA65: ACCCAATATCTTGATATTCTCTGAGTG;
ACA66: GTCTGAGGCCGGAAACAAGG;
ACA67: TTCTGTGCGACCCACTTCAG; and
ACA68:TCTTCAAACACCTCTCTCTC.

Further information can be found in the Supplemental Material (http://www.soe.ucsc.edu/~lowe/pubs/SupMat/RNA2005), which includes six files: a table of snoGPS scores of previously known H/ACA RNAs, annotated alignments of the previously known H/ACA RNAs, a table of the sequences with negative or inconclusive Northern results and the untested candidates, annotated alignments of the candidate sequences and newly verified H/ACA RNAs, a FASTA file of the newly verified H/ACA RNAs, and the snoGPS source code.

ACKNOWLEDGMENTS

We are grateful to Matthieu Blanchette for providing us with the hmr20 and mhr20 data sets and to Tamas Kiss for sharing pre-publication access to sequences of the H/ACA RNAs identified by his group. We thank Carrie Davis, Andy Pohl, Mark Diekhans, Hiram Clawson, and members of the Lowe Laboratory for numerous stimulating discussions and two anonymous reviewers for helpful comments. This work was supported in part by an Alfred P. Sloan Research Fellowship (T.M.L.).

Received August 29, 2005; accepted October 4, 2005.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Bachellerie, J.P., Cavaille, J., and Huttenhofer, A. 2002. The expanding snoRNA world. *Biochimie* **84:** 775–790.

Baker, D.L., Youssef, O.A., Chastkofsky, M.I., Dy, D.A., Terns, R.M., and Terns, M.P. 2005. RNA-guided RNA modification: Functional organization of the archaeal H/ACA RNP. *Genes & Dev.* **19:** 1238–1248.

Bertrand, E. and Fournier, M.J. 2004. The snoRNPs and related machines: Ancient devices that mediate maturation of rRNA and other RNAs. In *The nucleolus* (ed. M. Olson), pp. 225–261. Landes Bioscience, Georgetown, TX.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14:** 708–715.

Bortolin, M.L., Ganot, P., and Kiss, T. 1999. Elements essential for accumulation and function of small nucleolar RNAs directing site-specific pseudouridylation of ribosomal RNAs. *EMBO J.* **18:** 457–469.

Charpentier, B., Muller, S., and Branlant, C. 2005. Reconstitution of archaeal H/ACA small ribonucleoprotein complexes active in pseudouridylation. *Nucleic Acids Res.* **33:** 3133–3144.

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.

Ganot, P., Bortolin, M.L., and Kiss, T. 1997. Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. *Cell* **89:** 799–809.

Hirose, T. and Steitz, J.A. 2001. Position within the host intron is critical for efficient processing of box C/D snoRNAs in mammalian cells. *Proc. Natl. Acad. Sci.* **98:** 12914–12919.

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.

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.

Kent, W.J. 2002. BLAT—The BLAST-like alignment tool. *Genome Res.* **12:** 656–664.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res.* **12:** 996–1006.

Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. 2003. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci.* **100:** 11484–11489.

Kiss, T. 2002. Small nucleolar RNAs: An abundant group of noncoding RNAs with diverse cellular functions. *Cell* **109:** 145–148.

Kiss, A.M., Jady, B.E., Darzacq, X., Verheggen, C., Bertrand, E., and Kiss, T. 2002. A Cajal body-specific pseudouridylation guide RNA is composed of two box H/ACA snoRNA-like domains. *Nucleic Acids Res.* **30:** 4643–4649.

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.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Ma, X., Zhao, X., and Yu, Y.T. 2003. Pseudouridylation (Ψ) of U2 snRNA in *S. cerevisiae* is catalyzed by an RNA-independent mechanism. *EMBO J.* **22:** 1889–1897.

Ma, X., Yang, C., Alexandrov, A., Grayhack, E.J., Behm-Ansmant, I., and Yu, Y.T. 2005. Pseudouridylation of yeast U2 snRNA is catalyzed by either an RNA-guided or RNA-independent mechanism. *EMBO J.* **24:** 2403–2413.

Maden, B.E. 1990. The numerous modified nucleotides in eukaryotic ribosomal RNA. *Prog. Nucleic Acid Res. Mol. Biol.* **39:** 241–303.

Margulies, E.H., Blanchette, M., Haussler, D., and Green, E.D. 2003. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13:** 2507–2518.

Massenet, S., Mougin, A., and Branlant, C. 1998. Posttranscriptional modification in the U small nuclear RNAs. In *Modification and editing of RNA* (eds. H. Grosjean and R. Benne), pp. 201–227. ASM Press, Washington, DC.

Notredame, C., Higgins, D.G., and Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302:** 205–217.

Ofengand, J. and Fournier, M.J. 1998. The pseudouridine residues in rRNA: Number, location, biosynthesis and function. In *Modification and editing of RNA* (eds. H. Grosjean and R. Benne), pp. 229–253. ASM Press, Washington, DC.

Schattner, P., Decatur, W.A., Davis, C.A., Ares Jr., M., 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.

Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. 2003. Human–mouse alignments with BLASTZ. *Genome Res.* **13:** 103–107.

Tang, T.H., Bachellerie, J.P., Rozhdestvensky, T., Bortolin, M.L., Huber, H., Drungowski, M., Elge, T., Brosius, J., and Huttenhofer, A. 2002. Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. *Proc. Natl. Acad. Sci.* **99:** 7536–7541.

Torchet, C., Badis, G., Devaux, F., Costanzo, G., Werner, M., and Jacquier, A. 2005. The complete set of H/ACA snoRNAs that guide rRNA pseudouridylations in *Saccharomyces cerevisiae*. *RNA* **11:** 928–938.

Vitali, P., Royo, H., Seitz, H., Bachellerie, J.P., Huttenhofer, A., and Cavaille, J. 2003. Identification of 13 novel human modification guide RNAs. *Nucleic Acids Res.* **31:** 6543–6551.

Wang, C. and Meier, U.T. 2004. Architecture and assembly of mammalian H/ACA small nucleolar and telomerase ribonucleoproteins. *EMBO J.* **23:** 1857–1867.

Wuyts, J., De Rijk, P., Van de Peer, Y., Winkelmans, T., and De Wachter, R. 2001. The European large subunit ribosomal RNA database. *Nucleic Acids Res.* **29:** 175–177.

Wuyts, J., Van de Peer, Y., Winkelmans, T., and De Wachter, R. 2002. The European database on small subunit ribosomal RNA. *Nucleic Acids Res.* **30:** 183–185.