Supplemental Materials

Supplemental Text

Sequence Data Collection

Wormbase (www.wormbase.org; Chen et al. 2005) release WS170 annotated 122 snoRNA sequences, identified as 46 H/ACAs and 75 C/Ds through NCBI Blast (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi) and matches against relevant literature (Zemann et al. 2006; Deng et al. 2006; Huang et al. 2007). One gene remained unidentified (WBGene00045133), but its sequence identity is consistent with being a C/D snoRNA. We further collected sequences from the literature on *C. elegans* snoRNA identification and prediction (Zemann et al. 2006; Deng et al. 2006; Huang et al. 2007) and augmented the list by 16 H/ACAs and 14 C/Ds, bringing the totals to 62 and 87 (there were two ambiguously identified C/D sequences, see below), respectively. Some of these sequences have since been annotated in the more current versions (post-WS180) of Wormbase (Supplemental Table S1 provides a list of all previously annotated snoRNAs).

We refer to genes by their Wormbase ID (e.g. "WBGene00012345") where possible. Protein-coding genes are sometimes identified by their common gene name (such as *rpl-7*). In cases the genes have not been annotated in Wormbase, we used the name given in the gene's discovery paper. For example, CeN68 (Deng et al. 2006).

There is some discrepancy in the classification of three genes annotated in Wormbase. K12D12.6 (WBGene00045110), K12D12.7 (WBGene00045111) and K12D12.8 (WBGene00045112) all reside in the third intron of WBGene00010785 (K12D12.1). Given the sequence information and length of these annotated genes, they are more likely to be C/D snoRNAs than H/ACAs, but two of the references (Deng et al. 2006; Huang et al. 2007) identified regions overlapping these annotated genes as H/ACA snoRNAs, while another predicted WBGene00045110 as a C/D (Zemann et al. 2006). Although it is possible that this is an example of an H/ACA-C/D hybrid, such as the human U85 scaRNA (Henras et al. 2004), we did not include these genes in this study.

Sequence Set Selection

We characterized gene properties of the 36 host genes, containing parallel-nested H/ACAs, that we considered to be compact (Figures 1, 3). Where multiple variant transcripts exist, we selected the most

compact transcript to represent the host gene. For example, WBGene00009126 (F25H5.3) contains 5 annotated variant transcripts with alternate transcription start sites, but only one transcript (F25H5.3b) can be considered compact.

We used only the parallel-nested H/ACA snoRNAs to train the Bayesian Classifier. We also removed 11 sequences from the set of 54 sequences before training the Classifier. Some resided in non-compact host genes (WBGene00044928, WBGene00044946, WBGene00045120 and PsiCeSSU-970; see Supplemental Table S3). Others were considered atypical, raising a possibility that these predicted sequences may not represent genuine H/ACA snoRNAs. Two sequences, WBGene00045136 and WBGene00045154, were removed because they did not contain canonical H-boxes. The most likely H-box in these sequences were ATATTGA and AAATA respectively. Relaxation of the selection rules to allow these similar sequences introduces much background noise, so we set strict requirement for the H-box to follow the AnAnnA pattern exactly. Five other sequences (WBGene00044952, WBGene00045138, WBGene00045148, WBGene00045152, WBGene00045367) were removed because they were substantially longer than all other H/ACA snoRNAs. While all other known H/ACA snoRNAs were between 120 to 160 nt in length, these sequences are all longer than 210 nt. Furthermore, one hairpin is much longer than the other in these sequences, unlike the rest of the H/ACA genes. Additionally, two of these sequences (WBGene00044952 and WBGene00045367) do not fold into dual hairpin structures.

The Bayesian Classifier was thus trained on a positive training set of 43 H/ACA snoRNAs, which we refer to as the "RNA Training Set". We compiled a set of sequences that consisted of introns with length between 161 and 400 nt from all non-compact genes. This set of "Background Introns" represented the non-host intron sequences in the training of the Bayesian Classifier for two reasons: first, it is less likely to contain snoRNAs (we also trained using host-sized introns from compact genes, but the results were nearly identical), and second, the nucleotide usage in short introns (< 400 nt) appears to be appreciably different from that in longer introns (> 1000 nt), thus we considered the short introns from non-compact genes to be a more appropriate negative training set.

A negative control set was generated by taking 300-nt segments from the middle of 7,000 randomly selected long introns (> 1,000 nt), as they are also much less likely to contain ncRNAs than even the non-compact introns set. This set is referred to as "Long Introns". A second set of negative control sequences, referred to as "Short Introns", was generated by randomly selecting 7,000 sequences

from the Background Introns (Table 1A). The final numbers of sequences in both sets was less than 7,000, because the original selection of 7,000 sequences included duplicate introns from alternative transcripts. These duplicates were filtered out from the final reported numbers.

Preliminary Search for Box Motifs

Given a sequence, we first performed a fast scan for a valid ACA-box within the "allowed" region (at least 120 nt downstream of the 5' splice site and at least 6 nt upstream from the 3' splice site). We considered a valid ACA-box to be an ACAnnn or ATAnnn sequence which scored above the minimum threshold (95% of the lowest score from the RNA Training Set) using the ACA-box position weight matrix. For each valid ACA-box, we looked for a sequence matching the H-box consensus (AnAnnA, but excluding AAAAAA) between 90 and 50 nt upstream from the start of the ACA-box and at least 60 nt downstream from the 5' splice site. If both motifs were found, we designated the intervening sequence as a candidate hairpin 2 (HP2) and the 60 nt upstream from the H-box as candidate hairpin 1 (HP1). Thus, a single intron may contain multiple candidate H- and ACA-box pairs. Given that the ACA-boxes are always positioned three nucleotides from the 3' end of the snoRNA (Bachellerie et al., 2002), we were able to generate an alignment of the ACA-box motifs from the RNA Training Set and build a usage frequency-based position weight matrix extended to six nucleotide positions (Figure 2B). We tested whether this expanded definition of the ACA-box was shared by *Drosophila melanogaster*. The sequence logo derived from the aligned ACA-boxes of the fly (Supplemental Figure S2) shows no preference for any of the three nucleotides downstream of ACA. This further suggests that clade-specific training may be desirable.

Naive Bayesian Classifier

Setting up the Classifier

The Classifier employed four features: the starting hexamer of the host intron (5' splice site), the ACA-box, and two features relating to nucleotide composition of the RNA sequences (Figure 2D). We used the H-box in the preliminary search but not as a feature in the Classifier, because H-box sequences are indistinguishable from many A-rich sequences in introns, and thus the inclusion of H-box as a feature did not help with the Classifier performance. We used the Classifier to distinguish between H/ACA snoRNAs and RNA-less introns. We used the RNA Training Set as a positive training set, and the

Background Introns to approximate the negative training set (Table 1A). Although we cannot be certain that the Background Introns set contains no H/ACA genes at all, we assumed that the vast majority of introns do not contain H/ACA snoRNAs.

Using Bayes' Theorem, we can calculate the probability of a sequence S with features set $\{f_j\}$ belonging to class C_i ($P(C_i|\{f_j\})$) as follows:

$$P(C_i|\{f_j\}) = \frac{P(C_i)P(\{f_j\}|C_i)}{\sum_k P(C_k)P(\{f_j\}|C_k)} \quad \text{(Eq. 1)}$$

The denominator can be ignored as it is a constant. We therefore only need to calculated the numerator, which can be further simplified using the naïve (feature-independent) assumption, which transforms the numerator into the following:

$$P(C_i)P(\{f_j\}|C_i) = P(C_i)\prod_j P(f_j|C_i) \quad \text{(Eq. 2)}$$

where $P(C_i)$ is the probability of finding an H/ACA snoRNA or an intron sequence. The sequence S is then assigned to the class with highest probability.

Estimating the value of $P(C_i)$

$P(C_i)$ is the probability of finding a sequence belonging to a particular class C_i, or *class prior*. If we take this to be the probability of finding a host intron from the set of candidate introns ("Target Set", Table 1A), then the value is lower-bound at around 50/7000 (number of known H/ACA over number of possible introns in compact genes), and may be higher if there remain unannotated H/ACA genes. The Classifier, however, does not evaluate entire intron sequences, only the regions between possible H- and ACA-boxes, and some introns do not contain any possible H- and ACA-box pairs. This problem is further complicated by the fact that many of the known H/ACAs contain multiple H-boxes in the same region, thus a sequence may have multiple H- and ACA-box pairs. We therefore decided to leave the class priors as variables. For our final Classifier parameters, we used $P(C = RNA) = 1000/7000$.

Features of the Classifier

Feature 1 (5' splice site) assigned sequences to one of two categories: (1) for GTnnGT, (0) for other sequences. For the positive training set, we used the 5'-hexamers from the host introns of the RNA Training Set. For the negative set, we used the 5'-hexamers from Background Introns.

Feature 2 (ACA-box) assigned sequences to one of six categories: (1) ACAATT, (2) other ACAnTT's, (3) ACAAAn, (4) other ACAnnn's, (5) ATAnTT, (6) other sequences. The positive training set consisted of 43 ACA-boxes taken from the RNA Training Set, while the background set consisted of all hexamers in Background Introns which passed the ACA-box validation (i.e. start with "ACA" or "ATA" and score greater than the minimum threshold, previously defined as 95% of the lowest score from the RNA Training Set).

Features 3 and 4 employed a classifier-within-classifier approach. Both features were based on nucleotide composition of the RNA sequence. Feature 3 calculated the *class-specific tetramer usage* (see below), while feature 4 calculated the dinucleotide and trinucleotide usage frequency in individual sequences. Both contained a relatively high number of features (256 for feature 3 and 80 for feature 4), however, so we could not incorporate them directly into the final Classifier. Instead, we generated a naive Bayesian Classifier for both of these features using only their own sets of features, but rather than using them to classify a given sequence, we calculated the ratio between the values $\prod_j P(f_j|C_i)$ (i.e. without *class priors*) for C_i = positive set (RNA Training Set, RNA) and C_i = negative set (Background Introns, Bg) to obtain a single value $R(S)$.

$$R_{f3,\,f4}(S)=\frac{\prod_j P(f_j|C_{RNA})}{\prod_k P(f_k|C_{Bg})} \qquad \text{(Eq. 3)}$$

We then calculated the R(S) value for all sequences in the positive and negative training sets and divided the total range of values into equal segments in order to discretize the values. This approach allowed us to combine large groups of features of very different nature into a single classifier in such a way that we found, after extensive testing, to be most sensitive and accurate (data not shown).

We called **Feature 3** the "in-class tetramer usage". It employed a simplified version of the word-based naive Bayesian Classifier used in the Ribosomal Database Project Classifier (Wang et al. 2007), in which a sequence is classified according to the set of *n*-mer's it contains (where *n* is a fixed integer chosen for optimal performance). First, the *word-specific priors* were calculated for all words with length *n*:

$$P(w_i|C)=\frac{n(w_i)+0.5}{N_C+1} \qquad \text{(Eq. 4)}$$

where C = RNA Training Set (snoRNA) or Background Introns (Bg), N_C = number of sequences in

class C, $n(w_i)$ = number of sequences in C which contain the word w_i, and the values 0.5 and 1 ensure that $0 < P < 1$. Each word is used as an independent feature in the classifier. Any given sequence S can then be represented by the set of words it contains: $V_S = \{v_1, v_2, ..., v_k\}$ $(V \subseteq W)$. As the model assumes independence between features, the probability of observing V_S in C is estimated by the joint probability of observing each individual word v_i independently in C: $P(V_S \mid C) = \Pi\, P(v_i \mid C)$. The value $R(S)$ calculated for this feature is:

$$R_{f3}(S) = \frac{\prod_{v_i \in V} P(v_i | C_{snoRNA})}{\prod_{v_j \in V} P(v_j | C_{Bg})} \qquad \text{(Eq. 5)}$$

The values were made discrete by dividing the total range covered by the training sets into ten equal segments. We tested tetramers, pentamers and hexamers, but found that the longer words have strong tendencies to over-train the Classifier, especially given that our positive training set is relatively small (43 snoRNAs) and consists of relatively short sequences (120 – 160 nt). We therefore subsequently only used tetramers for all calculations.

Feature 4 was based on the dinucleotide and trinucleotide frequencies in the H/ACA sequences. The difference between this feature and feature 3 is that, feature 3 calculated the likelihood of a particular oligomer (tetramer) to be found in an H/ACA or intron sequence regardless of how frequently it appears in a single sequence, but feature 4 examined how frequently a particular dinucleotide or trinucleotide is present in a single sequence belonging to either set. Feature 3 is more useful for longer words while feature 4 is more appropriate for short motifs. We called this feature "in-sequence di-/tri-nucleotide frequency distribution."

Every di-/tri-nucleotide was considered as an independent feature. We made each feature (x_i) discrete by setting a threshold value ($V(x_i)$) such that for a given sequence in which x_i appears n times, then it is assigned a value of 0 if $n <= V(x_i)$ and 1 if $n > V(x_i)$. The threshold value for each feature was determined individually; it was the value at which the proportion of one training set (snoRNAs or introns) below the threshold is equal (or as close as possible) to the proportion of the other training set above the threshold. For example, the threshold for AA dinucleotide is 17: 94% of RNA sequences have 17 or fewer AA dinucleotides, while 94% of Background Introns have 18 or more AA dinucleotides.

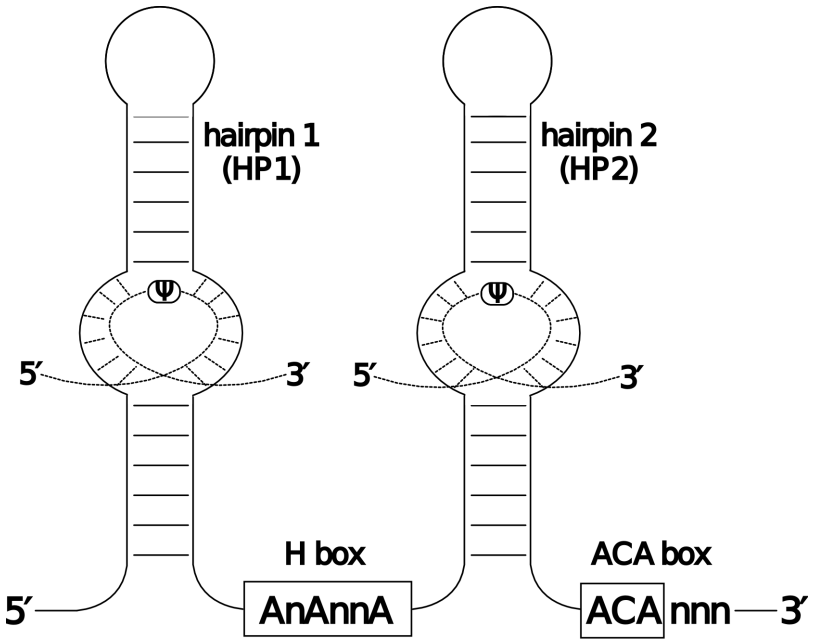
Random Re-sampling of Training Sets

For each sequence, we calculated conditional probability, $P(C_i)\prod_j P(f_j|C_i)$, of it belonging to either the class of snoRNAs (C_i = snoRNA) or to the background (C_i = Background Introns). In each re-sampling round, instead of using the entire positive or background set, only a subset of sequences was used to calculate the individual feature probabilities. In other words, in each re-sampling round, the Classifier was trained on randomly selected subsets of the RNA Training Set and Background Introns. The fraction of RNA Training Set used in each re-sampling round was determined after testing and optimization. We finally set this fraction at 66%, or 30 out of 43 sequences. An equal number of sequences was also randomly selected from Background Introns. We estimated the confidence of an H/ACA prediction by the number of times out of 1,000 re-sampling rounds in which it was classified as an H/ACA.

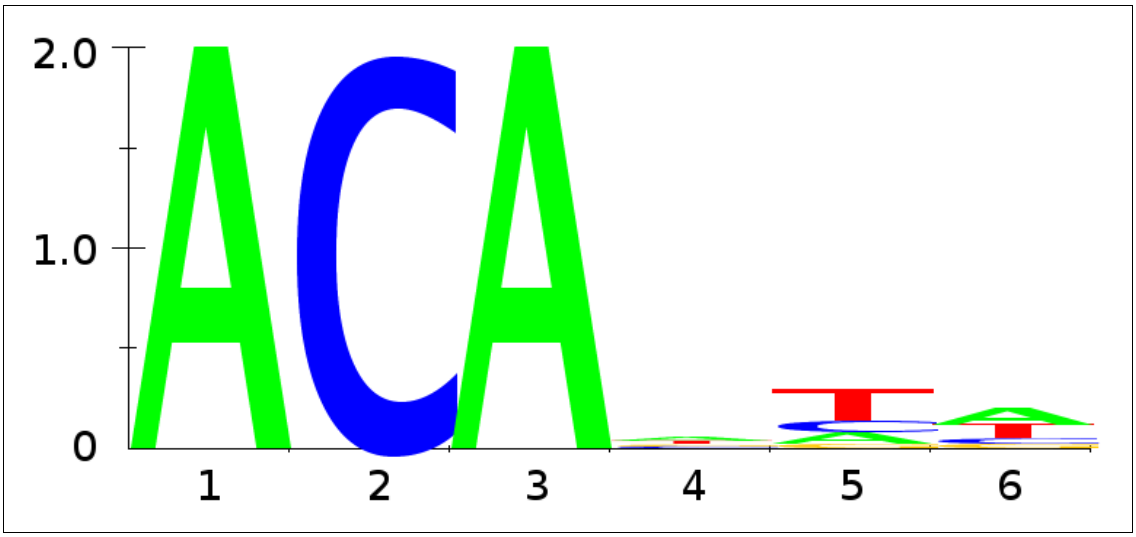
Leave-1-out Cross Validation

We evaluated the robustness of the Classifier training using "leave-1-out" cross validation method. We first trained the Classifier on N-1 (where N was the total, i.e. 43) sequences from the RNA Training Set, and then tested whether the removed sequence could be classified as H/ACA. We performed this for all sequences in the RNA Training Set. The results are shown in Supplemental Table S3 and Supplemental Figure S3.

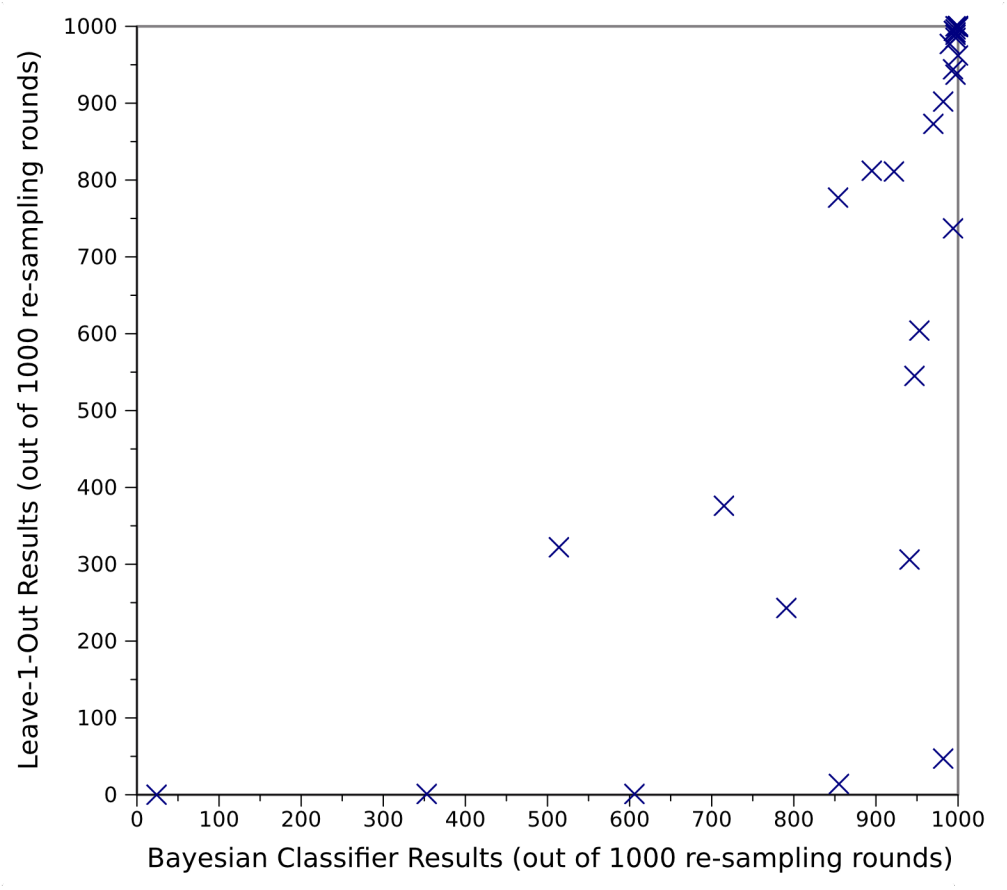
Supplemental Figures



Supplemental Figure S1. Secondary structure of a typical *C. elegans* H/ACA snoRNA. The snoRNAs are folded into two tandem hairpins of approximately equal lengths (50 – 80 nt), which are separated by the H-box (AnAnnA). The diagram also shows the topology of the snoRNA base-paired to its target sequence for pseudouridylation to be carried out on the target site (ψ).



Supplemental Figure S2. Sequence logo of *D. melanogaster* ACA-box. *Drosophila melanogaster* H/ACA sequences were obtained from FlyBase (flybase.org). The logo was generated from the ACA-box plus 3 downstream nucleotides from the 103 annotated H/ACA snoRNAs.



Supplemental Figure S3. Leave-1-out cross validation of the Bayesian Classifier. The x-axis represents the number of re-sampling rounds in which a given H/ACA snoRNA was classified as such by the default Bayesian Classifier. The y-axis represents the number of times an H/ACA snoRNA was recovered when training was carried out without this sequence being present in the training set. Note that most sequences, excluding the seven in the lower right quadrant, were recovered despite not being present in the training set. The remaining two in the lower left quadrant could not be recovered by either method.

Supplemental Tables

Supplemental Table S1. All currently annotated *C. elegans* snoRNAs.

A. Box H/ACA snoRNAs

WB	Zemann	Huang	Deng	Wachi	Chr	Host gene	Orient.	Operon
WBGene00023428	ComCe14	PsiCeLSU-1058		CeR-7	5	WBGene00010627	parallel	
WBGene00023429	Ce309	PsiCeSSU-513b	CeN41	CeR-9	5	WBGene00012347	parallel	CEOP5538
WBGene00023461	comCe6	PsiCeLSU-2966	CeN87	CeR-3	1	WBGene00004418	parallel	
WBGene00023466	Ce192	PsiCeLSU-2519b	CeN100	CeR-4	3	WBGene00004417	parallel	CEOP3836
WBGene00023468	Ce236		CeN90	CeR-6	3	WBGene00004498	parallel	CEOP3036
WBGene00023469	Ce280		CeN82	CeR-8	3	WBGene00004482	parallel	
WBGene00044928	CeN59	PsiCeSSU-1152	CeN59		1	WBGene00022121	parallel	CEOP1076
WBGene00044931	ComCe8	CeACAOrph1a	CeN95		1	WBGene00021350	parallel	CEOP1032
WBGene00044932	comCe9	CeACAOrph1b	CeN88		1	WBGene00021350	parallel	CEOP1032
WBGene00044935	Ce162	PsiCeLSU-1996	CeN43		1	WBGene00001497	parallel	CEOP1312
WBGene00044936	Ce48	PsiCeLSU-1573	CeN58		2	WBGene00020781	parallel	CEOP2238
WBGene00044937	Ce25	PsiCeSSU-971	CeN48		2	WBGene00007012	parallel	CEOP2144
WBGene00044939	comCe12	PsiCeSSU-513a	CeN110		3	WBGene00004498	parallel	CEOP3036
WBGene00044941	Ce345		CeN92		3	WBGene00004481	parallel	CEOP3316
WBGene00044942	Ce182		CeN99		3	WBGene00004481	parallel	CEOP3316
WBGene00044943	Ce170	PsiCeSSU-1523a	CeN80		3	WBGene00004417	parallel	CEOP3836
WBGene00044944	Ce283	PsiCeSSU-1523b	CeN79		3	WBGene00004417	parallel	CEOP3836
WBGene00044946	CeN91		CeN91		3	WBGene00003002	parallel	
WBGene00044948	Ce248	PsiCeLSU-2294a	CeN84		4	WBGene00004427	parallel	CEOP4032
WBGene00044952	Ce210		CeN96		4	WBGene00004419	parallel	
WBGene00044954	Ce58		CeN86		4	WBGene00017075	parallel	CEOP4288
WBGene00044955	ComCe22		CeN104		5	WBGene00000098	parallel	
WBGene00044956	Ce375		CeN94		5	WBGene00000098	parallel	
WBGene00044962	ComCe11	PsiCeLSU-2519a			4	WBGene00004492	parallel	
WBGene00044963	ComCe13	PsiCeSSU-1266a			5	WBGene00010627	parallel	
WBGene00044964	Ce286	PsiCeU6-26			1	WBGene00004186	parallel	
WBGene00045051	Ce352	PsiCeLSU-2294b			4	WBGene00004427	parallel	CEOP4032
WBGene00045094	Ce80				1	WBGene00004436	parallel	CEOP1188
WBGene00045113	Ce104.1				3	WBGene00002229	parallel	
WBGene00045114	Ce105				2			
WBGene00045118	Ce141		CeN38		3	WBGene00011407	*anti*	
WBGene00045120	Ce149	PsiCeSSU-1146	CeN125		5	WBGene00012829	parallel	CEOP5492
WBGene00045126	Ce166	PsiCeLSU-3220	CeN126		5	WBGene00007586	parallel	CEOP5360
WBGene00045130	Ce175	PsiCeLSU-2483	CeN93		2	WBGene00004440	parallel	CEOP2372
WBGene00045132	Ce176		CeN36-1		3			
WBGene00045134	Ce222	PsiCeSSU-1266b	CeN101		5	WBGene00010627	parallel	
WBGene00045136	Ce23	PsiCeLSU-1176	CeN55		4	WBGene00008362	parallel	CEOP4316
WBGene00045138	Ce244		CeN83		1	WBGene00009122	parallel	CEOP1624
WBGene00045142	Ce27		CeN49		2			
WBGene00045148	Ce273		CeN85		2	WBGene00007352	parallel	
WBGene00045151	Ce323	PsiCeLSU-2361	CeN81		5	WBGene00008920	parallel	
WBGene00045152	Ce356				2	WBGene00007352	parallel	
WBGene00045153	Ce36		CeN46		4	WBGene00008688	parallel	CEOP4488
WBGene00045154	Ce57	PsiCeLSU-2992	CeN39		4	WBGene00007514	parallel	CEOP4444
WBGene00045157	Ce87	PsiCeLSU-2836	CeN51		2			
WBGene00045171	BlCe176		CeN36-2		**X**			
*(WBGene00045373)**	CeN102		CeN102		2	WBGene00012897	parallel	CEOP2600
(WBGene00045374)	CeN105	PsiCeU5-48	CeN105		3	WBGene00007168	parallel	
(WBGene00045368)	CeN127		CeN127		3	WBGene00004391	parallel	
(WBGene00045376)	CeN45	CeACAOrph2	CeN45		1	WBGene00009126	parallel	
(WBGene00045378)	CeN67		CeN67		3	WBGene00010845	*anti*	CEOP3672
	CeN68		CeN68		**X**			
(WBGene00045367)	CeN97		CeN97		5	WBGene00008920	parallel	
	ComCe15				5	WBGene00010627	parallel	
	ComCe17				1	WBGene00004495	parallel	
	comCe5				5	WBGene00004413	parallel	
	comCe7		CeN78		1	WBGene00004436	parallel	CEOP1188
(WBGene00077515)		PsiCeSSU-499			1	WBGene00017088	parallel	CEOP1892
		PsiCeSSU-970			3	WBGene00015468	parallel	CEOP3312
(WBGene00077516)		PsiCeU2-16			3	WBGene00004187	parallel	
(WBGene00077517)		PsiCeU2-41			3	WBGene00004187	parallel	
(WBGene00077462)		PsiCeU4-62			4	WBGene00006466	parallel	CEOP4535

B. Box C/D snoRNAs

WB	Zemann	Huang	Deng	Wachi	Higa	Chr	Host gene	Orient.	Operon
WBGene00007539	Ce61	MeCeLSU-A1185	CeN118		sn1185	2	WBGene00006660	anti	
WBGene00014290	Ce372	MeCeLSU-C3071	CeN123		sn3071	5	WBGene00001493	parallel	
WBGene00014397	Ce209	MeCeLSU-G2903	CeN124		sn2903	5	WBGene00010079	parallel	
WBGene00014425	Ce100	MeCeLSU-U2762			sn2762	2			
WBGene00014447	Ce239	MeCeLSU-A2317	CeN119		sn2317	2	WBGene00003367	parallel	
WBGene00014506	Ce325	MeCeLSU-A3159	CeN120		sn3159	2	WBGene00011524	parallel	
WBGene00014541		MeCeLSU-C3060			sn3060	5	WBGene00002133	parallel	
WBGene00022895					sn2524	5	WBGene00015043	anti	
WBGene00022932	Ce252	MeCeLSU-A2429	CeN122		sn2429	5			
WBGene00022968					sn3087	4	WBGene00016957	anti	
WBGene00023020	Ce230	MeCeLSU-C1502	CeN15		sn1502	3			
WBGene00023027					sn2991	4	WBGene00018073	anti	
WBGene00023071	Ce297	MeCeLSU-A2384c	CeN17		U15	3	WBGene00001748	parallel	
WBGene00023083	Ce354	MeCeLSU-U2841	CeN14		sn2841	4			
WBGene00023183	Ce39	MeCeLSU-G2343	CeN121		sn2343	4	WBGene00021939	anti	
WBGene00023189	CeN13	MeCeLSU-G668	CeN13		sn668	3	WBGene00022104	parallel	
WBGene00023193	Ce81	MeCeLSU-U2417	CeN117		sn2417	1			
WBGene00023430	CeR19	MeCeU2-G12	CeN69	CeR-19		5	WBGene00008371	parallel	CEOP5260
WBGene00023467	CeR5			CeR-5		4	WBGene00021430	parallel	
WBGene00044704	Ce211	MeCeLSU-A678	CeN5		U18	2			
WBGene00044925	Ce177	MeCeSSU-C400	CeN33			1	WBGene00021352	parallel	
WBGene00044926	CeN65	MeCeSSU-A90	CeN65			1			
WBGene00044929	CeN54	MeCeLSU-C2300	CeN54			1	WBGene00004464	parallel	
WBGene00044930	CeN53		CeN53			1			
WBGene00044933	Ce223		CeN60			1	WBGene00001819	anti	
WBGene00044934	Ce55		CeN61			1			
WBGene00044938	Ce75		CeN63			2	WBGene00003966	anti	
WBGene00044940	Ce285	MeCeSSU-C1180	CeN106			3	WBGene00006725	parallel	CEOP3088
WBGene00044945	CeN89	MeCeU2-C42	CeN89			3	WBGene00000479	parallel	
WBGene00044947	Ce243		CeN111			4			
WBGene00044957	Ce63	MeCeSSU-C980	CeN27			5	WBGene00015592	parallel	
WBGene00044958	Ce139		CeN109			5	WBGene00004256	parallel	
WBGene00044970	Ce238	MeCeLSU-U1566				4	WBGene00000563	parallel	
WBGene00044974	Ce67					5	WBGene00000167	parallel	
WBGene00044975	Ce59					1			
WBGene00045055	Ce86					5	WBGene00004395	parallel	
WBGene00045071	Ce18	MeCeLSU-C2407				5	WBGene00009830	anti	
WBGene00045072	Ce104.2					3	WBGene00002229	parallel	
WBGene00045074	Ce110					2	WBGene00011304	parallel	CEOP2492
WBGene00045082	Ce234		CeN47			1	WBGene00018512	parallel	CEOP1264
WBGene00045083	Ce169					2	WBGene00003402	anti	
WBGene00045084	Ce151					5	WBGene00006803	parallel	
WBGene00045085	ComCe19					4	WBGene00004419	parallel	
WBGene00045086	ComCe18					4	WBGene00004473	parallel	
WBGene00045087	ComCe16					1	WBGene00004415	parallel	
WBGene00045088	Ce251					2	WBGene00018586	parallel	
WBGene00045089	Ce254b	MeCeLSU-A2384b				3	WBGene00000518	anti	
WBGene00045090	ComCe10					2	WBGene00003966	anti	
WBGene00045092	ComCe20					3	WBGene00006725	parallel	CEOP3088
WBGene00045093	Ce282					1			
WBGene00045109	Ce135					4	WBGene00007882	parallel	
WBGene00045112						2	WBGene00010785	parallel	
WBGene00045115	Ce118	MeCeSSU-A601	CeN24			2	WBGene00003089	anti	
WBGene00045121	Ce138a	MeCeLSU-C2440				4			
WBGene00045124	Ce160	MeCeLSU-A3023	CeN30			3	WBGene00003572	anti	
WBGene00045125	Ce161	MeCeSSU-A422	CeN108			2	WBGene00002065	parallel	
WBGene00045127	Ce167					2			
WBGene00045129	Ce171		CeN40			5			
WBGene00045137	Ce240					5			
WBGene00045139	Ce245		CeN28			2	WBGene00008500	parallel	
WBGene00045140	Ce246	MeCeLSU-G860	CeN70			5			
WBGene00045141	Ce254a	MeCeLSU-A2384a				2			
WBGene00045147	Ce271		CeN44			2			
WBGene00045149	Ce298	MeCeU6-A48				3	WBGene00012272	anti	
WBGene00045150	Ce304		CeN114			1			
WBGene00045155	Ce62	MeCeU6-C55	CeN98			4	WBGene00006698	parallel	CEOP4596
WBGene00045156	Ce83		CeN57			1			

WBGene00045158	Ce94					X			
WBGene00045159	Ce96					5			
WBGene00045160	Ce98	MeCeSSU-C1196				2			
WBGene00045172	BlCe298					3	WBGene00000875	parallel	
WBGene00045176	Ce138b					X			
WBGene00045200						5	WBGene00008395	anti	
	BlCe173.2					2			
	Ce97					2	WBGene00016116	parallel	CEOP2208
	CeN103		CeN103			1	WBGene00000371	parallel	CEOP1930
(WBGene00045375)	CeN113		CeN113			1	WBGene00007640	anti	
	CeN22		CeN22			1			
(WBGene00045377)	CeN62		CeN62			2	WBGene00008964	parallel	
	ComCe1					3	WBGene00004482	parallel	
	ComCe2					3	WBGene00004414	parallel	
	ComCe21					5	WBGene00000167	parallel	
	ComCe23					5	WBGene00015592	parallel	
	ComCe3					2	WBGene00004434	parallel	
	ComCe4					5	WBGene00010627	parallel	
		MeCeSSU-U1342				4			
		MeCeU2-U49				2	WBGene00007013	parallel	CEOP2552

C. Sequences with ambiguous or unknown classification

WB	Zemann	Huang	Deng	Chr	Host gene	Orient.	Operon
WBGene00045110	Ce173.1		CeN128	2	WBGene00010785	parallel	
WBGene00045111		PsiCeU5-45		2	WBGene00010785	parallel	
WBGene00045133				X			

Supplemental Table S1. All currently annotated snoRNAs. **A.** Box H/ACA snoRNAs. **B.** Box C/D snoRNAs. **C.** snoRNAs with unknown or ambiguous classification. Abbreviations: WB = Wormbase Gene ID, Zemann = gene identification given in Zemann et al. 2006, Huang = gene ID by Huang et al. 2007, Deng = gene ID by Deng et al. 2006, Wachi = gene ID by Wachi et al. 2004, Higa = gene ID by Higa et al. 2002, Chr = chromosome, Orient. = relative orientation between host gene and nested snoRNA, Operon = Wormbase operon ID. Rows shaded in gray represent snoRNAs that are not nested, those shaded in blue represent genes in anti-parallel orientation. (*) Italicised and bracketed Wormbase gene ID's denote annotation added into Wormbase after WS170 release.

Supplemental Table S2. Genomic organization of *C. elegans* small RNA and their host genes.

A. Nesting properties of *C. elegans* small RNAs.

RNA type	unnested	antiparallel	parallel	TOTAL
tRNA*	348	116	144 (24%)	608
snRNA*	43	13	25 (31%)	81
C/D*	25	15	33 (45%)	73
C/D**	28	16	43 (49%)	87
H/ACA*	5	2	45 (87%)	52
H/ACA**	6	2	54 (87%)	62

B. Host genes and operons

	tRNA*	snRNA*	C/D**	H/ACA**
Number of parallel host genes	129	22	40	40
Number of parallel host genes in operons	11	3	8	23
Number of antiparallel host genes	101	10	15	2
Number of antiparallel host genes in operons	10	0	0	1
Percentage (%) of parallel host genes in operon	8.5 %	13.6 %	20.0 %	57.5 %
Percentage (%) of all host genes in operons	9.1 %	9.4 %	14.5 %	57.1 %

Protein-coding genes: 2871/20084 = 14.3%
Compact genes: 783/4699 = 16.7%

C. Co-expression and function of H/ACA host genes

Mount	# host genes	protein functions
2	6	ATPase, Vitamin-D receptor, chaperone
5	4	zinc-finger protein, U5 snRNP, mRNA cleavage factor
7	1	Ser/Thr-kinase
11	4	kinesin-like protein, Ser/Thr-kinase, splicing factor, ribonucleotide reductase, protein-binding
20	2	tRNA synthase, ribosome genesis
23	12	mostly rpl and rps genes, glutathione S-transferase, nucleosome assembly
30	1	
-	10	ribosomal proteins, RNA-binding, pyruvate kinase, phosphatidylinositol synthase, Golgi complex

Supplemental Table S2. A. Nesting properties of small RNAs in *C. elegans*. The RNAs were divided into three groups according to nesting and relative orientation with respect to the host gene. **B.** Small RNA host genes and operons. 2,871 out of 20,084 (14.3%) protein coding genes in *C. elegans* reside in 1,117 operons. **C.** Co-expression Mounts (Kim et al. 2001) of box H/ACA snoRNA host genes. (*) These data were calculated from Wormbase WS170 only. (**) These data include additional snoRNA annotated in literature (all listed in Supplemental Table S1).

Supplemental Table S3. Parallel-nested H/ACA snoRNAs in *C. elegans*

			Host Gene Identity			Classifier Results			Folding		Expression		comments
ceidn	Compact Host	Training Set	WB_ID	Transcript	Rank	BC		L1O	HP1	HP2	NPA	SNPA	
18	y	y	10627	K07C5.4	6	997		985	53	63	-	-	
19	y	y	12347	W08G11.3a	6	997		1000	59	59	-	78993	
21	y	y	4418	F53G12.10.1	2	982		47	61	56	4	138	
22	y	y	4417	R151.3.1	1	997		989	55	60	2170	41351	
24	y	y	4498	B0412.4.1	1	941		306	63	56	1807	-	
25	y	y	4482	C16A3.9.1	2	970		873	60	57	2122	-	
29	**n**	**n**	22121	Y71F9AM.4a	5	978		N/A	61	56	164	-	
32	y	y	21350	Y37E3.8a.1	1	995		994	51	59	109	-	
33	y	y	21350	Y37E3.8a.1	2	1000		1000	53	59	107	2232	
36	y	y	1497	T08B2.9a	7	1000		1000	53	59	-	-	
37	y	y	20781	T24H7.2.1	3	1000		999	63	56	1258	-	
38	y	y	7012	ZK546.13	4	994		737	56	51	1170	-	
40	y	y	4498	B0412.4.1	2	922		811	62	56	1808	33828	
42	y	y	4481	F54E7.2.1	1	1000		962	59	57	2074	39318	
43	y	y	4481	F54E7.2.1	2	982		902	55	59	2075	39320	
44	y	y	4417	R151.3.1	2	1000		999	53	57	2171	41352	
45	y	y	4417	R151.3.1	3	1000		1000	57	57	2172	-	
47	**n**	**n**	3002	C03B8.4	6	0		N/A	-	-	2228	-	
49	y	y	4427	K11H12.2.1	2	947		545	60	51	2577	-	
50	y	**n**	4419	Y24D9A.4a.1	2	0		N/A	-	-	2778	53314	HP2 = 112nt
51	y	y	17075	D2096.8	2	1000		1000	52	56	3018	-	
52	y	y	98	K07C11.2.1	5	1000		1000	64	58	-	-	
53	y	y	98	K07C11.2.1	6	1000		1000	42	57	-	-	
56	y	y	4492	F28D1.7.1	2	994		944	55	58	3249	61795	
57	y	y	10627	K07C5.4	1	997		992	61	50	-	-	
58	y	y	4186	F22D6.5	7	997		996	57	58	-	8280	
62	y	y	4427	K11H12.2.1	1	854		777	48	54	2578	-	
78	y	y	4436	D1007.12.1	1	990		977	56	76	334	-	
83	y	y	2229	Y43F4B.6.1	6	997		937	54	58	-	-	
87	**n**	**n**	12829	Y43F8C.7.2	3	74		N/A	60	55	-	-	
91	y	y	7586	C14C10.3a.1	6	514		322	58	51	-	76036	
94	y	y	4440	F28C6.7a.1	1	855		14	55	57	1365	-	
97	y	y	10627	K07C5.4	5	953		604	70	50	-	-	
98	y	**n**	8362	D1046.1.1	5	0		N/A	-	-	3041	57941	non-canonical H-box
100	y	**n**	9122	F25H2.11.1	1	0		N/A	-	-	-	12632	
106	y	**n**	7352	C06A1.1.1	3	997		N/A	56	63	1445	-	
109	y	y	8920	F17C11.9a.1	5	606		1	53	57	-	74408	
110	y	**n**	7352	C06A1.1.1	1	1000		N/A	63	78	1447	-	
111	y	y	8688	F11A10.7	2	715		376	52	55	3239	61452	
112	y	**n**	7514	C10C6.6	7	0		N/A	-	-	3208	-	non-canonical H-box
125	y	y	12897	Y46G5A.5	1	24		0	64	57	1590	-	
127	y	y	7168	B0393.3	2	353		1	57	64	2031	-	
129	y	y	4391	T23G5.1.3	3	791		243	58	73	2292	-	
131	y	y	9126	F25H5.3b	5	895		812	72	56	-	10960	
135	y	**n**	8920	F17C11.9a.1	4	0		N/A	-	-	-	74405	HP2 = 150nt
137	y	y	10627	K07C5.4	7	95		1	53	62	-	-	
138	y	y	4495	F39B2.6.1	1	697		383	69	55	-	-	
144	y	y	4413	B0250.1.1	1	941		487	53	50	-	-	
145	y	y	4436	D1007.12.1	3	999		991	65	84	336	-	
148	y	y	17088	E01A2.6	1	919		919	58	56	226	-	
149	**n**	**n**	15468	C05D2.6a.1	4	999		N/A	65	56	-	-	
150	y	y	4187	C50C3.6	2	855		521	60	35	2250	-	
151	y	y	4187	C50C3.6	5	277		21	54	56	2251	42602	

152	y	y	6466	JC8.13	1	990	951	63	61	3296	-	

Supplemental Table S3. All parallel-nested box H/ACA snoRNAs in *C. elegans*. "ceidn": unique identification number developed for all *C. elegans* snoRNAs. This had to be done because there currently is no single identification number that can be consistently used to describe every snoRNA. "Compact Host": whether the H/ACA host gene is compact. "Training Set": whether the H/ACA was included in the RNA Training Set. "WB_ID": Wormbase Gene ID, which is of the form "WBGene12345678"; here it is truncated such that 17125 represents WBGene00017125. "Transcript": transcript "Sequence Name" as given in Wormbase. "Rank": intron rank. "BC": Bayesian Classifier result, number of re-sampling rounds out of 1,000 the sequence was classified as H/ACA. "L1O": Leave-1-Out cross-validation result, see Supplemental Text for details. "Folding": predicted hairpin length for HP1 and HP2. "Expression": number of a sequence tag with which a given snoRNA was detected (He et al. 2007). NPA and SNPA refer to non-polyadenylated and small (< 500 nt) non-polyadenylated RNA fractions. The numbers are truncated such that, for example, 42602 in the SNPA column represents SNPA42602. "Comments" note the H/ACAs with irregular sequence features.

Supplemental Table S4. Expressed and Conserved Candidate H/ACA sequences in *C. elegans*

Cel Host Intron			BC	Folding		Expression		Conservation			Comments
WB_ID	Transcript	Rank	passes	HP1	HP2	NPA	SNPA	E_BLAST	alignment	*Cbr* match	
17125	E04F6.5b.1	5	981	52	62	0	0	2.5E-36	294 | 239	intergenic	
19667	K12B6.2	11	697	68	64	0	0	5.5E-25	215 | 174	intergenic	
19384	K04F1.1	4	895	74	50	0	0	1.2E-17	263 | 187	intergenic	
21804	Y53G8AM.4	3	697	75	66	0	0	2.4E-15	270 | 183	multiple (33234 | 2)	
5488	C47A10.2	2	606	62	63	0	0	1.3E-15	148 | 99	multiple (36641 | 5)	snoReport = 0.935
20105	R148.7	5	510	71	69	0	0	8E-13	281 | 177	multiple (42349 | 2)	snoReport = 0.823
19545	K09B3.1	2	970	58	66	0	0	1E-09	145 | 92	multiple (27535 | 2)	snoReport = 0.528
18519	F46H5.3b.1	2	571	53	52	0	0	4.9E-09	198 | 134	32214 | 2	Orthologous *Cbr* intron is conserved
19306	K02E7.1a	2	982	54	56	0	0	1.5E-10	175 | 123	35967 | 4	host is lost in *Cbr*, but exists in other nematodes
10072	F54F12.1	6	941	72	52	0	0	1.2E-07	107 | 85	intergenic	
2007	C15H9.6.1	2	676	52	61	0	0	2.5E-10	138 | 107	35220 | 2	Orthologous *Cbr* intron is conserved
10976	R02D5.1	1	676	59	57	0	0	5E-08	115 | 90	27237 | 1	Orthologous *Cbr* intron is conserved
12638	Y38H8A.4	2	854	62	59	0	0	1.9E-07	141 | 89	multiple (39963 | 1)	
16081	C25A6.1	1	842	72	78	0	0	8.7E-08	156 | 111	30987 | 1	Orthologous *Cbr* intron is conserved
1616	ZC196.7	5	551	63	75	0	0	2.1E-07	268 | 167	multiple (41003 | 2)	
11116	R07E5.10	4	606	69	73	0	0	2.2E-06	295 | 177	multiple (42179 | 7)	
19730	M02D8.4a	2	783	64	55	0	0	9.1E-06	154 | 102	35706 | 2	Orthologous *Cbr* intron is conserved
18805	F54D10.4	4	855	61	65	0	0	4.1E-07	153 | 97	36121 | 6	
16378	C33H5.11.1	8	551	57	72	0	0	4.7E-06	247 | 149	32895 | 14	
10785	K12D12.1	3	718	52	58	0	1547	6.1E-15	301 | 193	25743 | 3	C/D, orthologous *Cbr* intron is conserved
7603	C15C6.4	4	697	62	58	14452	0	10	-	-	
660	T21B4.2	2	510	57	57	30069	1576	2.3E-16	136 | 113	31690 | 2	tRNA, orthologous *Cbr* intron is conserved
137	R13G10.2	4	697	75	51	36837	0	1.5	-	-	
567	ZK632.6	4	947	54	57	44516	2317	4.9E-06	120 | 78	27821 | 6	
16763	C49A9.9a.1	2	551	51	56	55032	0	6.6E-06	142 | 88	30569 | 9	
13657	Y105C5B.18	2	990	62	55	64707	0	1.8	-	-	
2005	F26D10.3.1	2	1000	66	61	65349	3472	5.9E-13	170 | 126	23842 | 2	Orthologous Cbr intron is conserved
16822	C50E3.11	2	994	55	67	71253	0	0.065	-	-	
9770	F46B6.5b	5	783	70	60	73261	0	0.00096	149 | 97	39128 | 2	
11194	R10D12.13a	9	994	53	75	77308	0	5.8E-10	196 | 136	multiple intergenic	
11195	R10D12.14a	4	994	53	75	77308	0	10	-	-	

Supplemental Table S4. List of host introns containing candidate box H/ACA snoRNA. “WB_ID”: truncated form of Wormbase Gene ID, e.g. 17125 represents WBGene00017125. “Rank”: intron rank. “BC”: Bayesian Classifier, the values are the number of re-sampling rounds out of 1000 the sequence was classified as H/ACA. “Folding”: the length of the folded hairpin for HP1 and HP2. “Expression”: number of a sequence tag with which a given snoRNA was detected (He et al. 2007). NPA and SNPA refer to non-polyadenylated and small (< 500 nt) non-polyadenylated RNA fractions. “Conservation”: results of BLAST searches against *C. briggsae* genome. E-score, alignment (number of identical nucleotides / length of aligned region), and the matching *C. briggsae* intron, if the BLAST match was found within an intronic region (WB_ID | intron rank). Rows highlighted in blue are genes which have also been predicted by snoReport; the score given is the the highest score given to a candidate found within the same intron by snoReport (Hofacker et al. 1994). Rows in grey are genes of particular

interest (they are both conserved and expressed, or they are conserved in orthologous introns), their predicted secondary structures are given below.

RNA Secondary Structure and Sequence Homology of Eight Predicted *C. elegans* H/ACA snoRNAs from Supplemental Table S4

Here we provide the predicted H- and ACA-boxes with the RNA secondary structure (as predicted by RNAfold), as well as the BLASTN results (against *C. briggsae*), for eight cases highlighted in yellow in the table above. The identity of the matching *C. briggsae* introns for each predicted *C. elegans* H/ACA snoRNA are given in the table above.

1) Cel: WBGene00018519 (F46H5.3b.1) Intron 2
Cbr: WBGene00032214 – Intron 2

Folding of the *C. elegans* sequence

```
 [HP1]                                                  [H-box]
aggaaaaaagacgcggcgcaatttcaaatgggattactagcgtattgagtgaccgAAACTAgcACATAA
.((.......((((...(.(((..(....)..))).)..)))).........)).

[HP2]                                                   [ACA-box]
acgcaatggcctcgctcccaaagtaactaactctatttgcggttcacgtcgcgccaACAGTT
.(((.(((((..((((......(((.....)))......)))))..).))).)))...
```

Alignment of *C. elegans* and *C. briggsae* sequence

```
Cel  aatgatgaataagagaaattccaa-----------------------------------------------tgct
                                                                            || |
Cbr  gtttctcttttctttttccgtggtgagttgatgcaaatgcccggcgcgtcagcatgcacaaaaatgttgtt

Cel  atagtag--ggaaatagactagaaaacaaggaaaaaagacgcggcgcaatttcaaatgggattactagcgta
      |||||   |||||| ||   ||||  | ||| |||||  ||| ||  |||||||||||||||||||||| |
Cbr  gtagtacaaggaaatggaagcgaaat-agggagaaaag--gcg-cg--atttcaaatgggattactagcgca

Cel  ttgagtgaccgAAACTAgcACATAAacgcaatggcctc---gctcccaaagtaactaactctatttgcggtt
     ||| |||  | || | || | |  |||| |||||||||   | |       ||||||||| | ||||
Cbr  ttgggtgggc-aA-C-AGTAga--aacgaaatggcctcaaagttgttggcctaaccaactatgtttgtctcc

Cel  cacgtcgcgccaACAGTTatta--aaaatgaccatctgccgattcgcaaACACTT------------tgacg
     |  ||| | | | || ||||||  |||||| ||| ||| || |  |||||    |
Cbr  cgtgtccctctAGCAATtattatgaaaatggccacctggcgctcggcaaatgtATAATC--ggcgtcgcgcc
```

2) Cel: WBGene00002007 (C15H9.6.1) – Intron 2
Cbr: WBGene00035220 – Intron 2

Folding of the *C. elegans* sequence

```
[HP1]                                                                                [H-box]
tacccttttgaccgttttgtgtaaacaatagattttgggtcagtgactggtacaggttcctctctcgttaggaatgaggaATAGGA
((((...((((((..(((((.......)))))).....))))))).....))))).....((((((..((.....))..))))))
  [HP2]                                                                  [ACA-box]
atgtttgctcaggtccgaagctgtaccaaatcacagattaagtatagagggttgactgcagatttgaaACAAAA
..((((((...(..((....((((((..((((...))))).))))))).))..)....))))))).....
```

Alignment of *C. elegans* and *C. briggsae* sequence

```
Cel     gtgagtttatgaga---------aaagtgctctctaatatttgtctctggactacccttt
                              ||||  | |||||||||||||   ||||||||||||
Cbr     gtgagttagtagagagaaaacagaaaga-c-ctctaatatttgtacgtggactaccttta

Cel     ----tgaccgttttgtgtaaacaatagattttgggtcagtgactggtacaggttc-ctct
            |||||||||||||| |||||||||||||| |||||||||||||||||||||| |||
Cbr     ccaatgaccgttttgtgcaaacaatagactttcggtcagtgactggtacaggttcgctca

Cel     ctcg-tt-agg-aatgaggaATAGGAatgtttgctcaggtccgaagctgtaccaaatcac
         ||| || ||| |||||          |||||| ||  ||   || |   |||||||
Cbr     ttcgattgaggcaatgacagtgc-tgt-gtttgcacaactct--AGAT---CCAaatcta

Cel     agattaagtatagagggttgactgcagatttgaaACAAAAtaatt---------------
           ||    | |     ||
Cbr     gtgtttcgaaaactagatttgctcattaatagggaattgagggactagACAATG
```

3) Cel: WBGene00016081 (C25A6.1) – Intron 1
Cbr: WBGene00030987 – Intron 1

Folding of the *C. elegans* sequence

```
    [HP1]                                                                        [H-box]
tctgcaaaaaagtgcgcatctctcaagtatcaaagtgacagaaagttgatgctgctccgatgagctgtcaaatttgaaAAATGA
....((((.....((.((((.....(((((((((..(......)..)))))))).....)))).))).......)))))..
[HP2]                                                                          [ACA-box]
cgtcaactggagtgatggatcttttggacgtttatttctctctcccgatcttccaaatcccatggcgtgaacttctttgaaACATTT
((((((..(((..(..((((....((((.................))))...))))).)..)))))))).............
```

Alignment of *C. elegans* and *C. briggsae* sequence

```
16081/1-1544    gtgttttcaaaaagaaacat-atttattgtccatca-catattgctctgcaaaaa-agtg
                 | |     | |  |      |||| |||  | ||  || | |  | ||||||| |||
30987/1-1203    atttacgctagattactcga-attt-ttgatcttcttcaaaatt-t-tgcaaaaatagtt

16081/1-1544    cgcatctctcaagtatcaaagtgacagaaagttgatgctgctccgatgagctgtcaaat-
                   ||||||||||   || ||||||||     ||||||   | || |||||||| ||
30987/1-1203    tctatctctcaagtcaaaatgtgacagaT----gatgct---cggaggagctgtcatatg

16081/1-1544    ttgaaAAATGAcgtcaactggagtgatggatcttt-tggacgttt-atttctctctcccg
                || |||||||||||||    | ||||||||||   || || ||| || | |||| |
30987/1-1203    ttcaaAAATGAcgtcaaaaataatgatggatctatctgcacttttgatatttctccctta

16081/1-1544    atcttccaaatcccatggcgtgaacttctttgaaACATTTttcaagtcatgattccag--
                ||       |     |
30987/1-1203    ataacgttcaatgtttctaggtcag-----------------------------------
```

4) Cel: WBGene00019730 (M02D8.4a) – Intron 2
Cbr: WBGene00035706 – Intron 2

Folding of the *C. elegans* sequence

```
[HP1]                                                              [H-box]
tcttatcgcttgttatcattagtgtatcacttctcctcctcctaagataagcgaaaagtgtaACACAA
.(((.((((((((((......((((...))))).............)))))))))).))).....
    [HP2]                                                        [ACA-box]
cacacgccgtgttctttgtgctgatattaatgacccccagacacaccttagcacctaggcACAAAG
.....(((((((((...(((((((.............))).)))).....)))))...)))
```

Alignment of *C. elegans* and *C. briggsae* sequence

```
Cel     tagtcaaagtacaa-ttagttcttatcgct-tgttatca-ttagtgtatcact-tctcct
        |  ||  |  ||   |||   ||||||  | | ||   | ||  |  | |||  ||||||
Cbr     tt-tcttatcacttgttatcacttatctatctcttcgaacttcctccaccaccctctcct

Cel     cctcctaa-gataagcgaaaagtgtaACACAAcacacgccgtgttctttgtgctgatatt
        | | ||   || | | | ||  |  ||||||   | || ||||||||||||||||||||
Cbr     cttgcttttgaca-gtgtaACAT--AAcacacgtcgcgtcgtgttctttgtgctgatatt

Cel     aatgacccccagacacaccttagcacctaggcACAAAGattttgttgcacattggttat
        ||||||| | ||||| || || |||    |
Cbr     aatgaccgctagacaaac-tt-gcaaaatgtggctcttgACACCTgcctgcggtcaaaga
```

5) Cel WBGene00000567 (ZK632.6) – Intron 4
Cbr WBGene00027821 – Intron 6

Folding of the *C. elegans* sequence

```
[HP1]                                                              [H-box]
gaagaaattcttaaattataacaaaaaatagagttcaatactcttgatttcttcaccgaatgcctATAAAA
(((((((((.....................(((((.....)))))).)))))))))...........
[HP2]                                                            [ACA-box]
cctagatcgacttcttcacccttttcgtgctcatccacgggcgacgaggatttaaagttgatACAATT
.....(((((((((......((((.((((.(........).)))).)))).....)))))))))
```

Alignment of *C. elegans* and *C. briggsae* sequence

```
Cel     --------------------------gaagaaattcttaaattataacaaaaaatagagt
Cbr     ggcgacaggaggggagggagacgcagagaagcgagagtttctgcgtctctctccgccaac

Cel     tcaatactcttgatttcttcaccgaatgcctATAAAAcctagatcgacttcttcaccctt
Cbr     tttaaaagcgaataactcgcttcaatgttgagctAGAAAAatgaaaattagctagttttg

Cel     ttcgtgctcatccacgggcgacgaggatttaaagttgatA-CAATTatacaattctagaa
              |               | ||| ||||||| |||  |||| | | |    | ||
Cbr     caagaaatttcaagagctttcgaatgatataaagtttATAAGAattttcgaaaaaACAA

Cel     atttcagacaaacttcgg
         |||  || ||| ||
Cbr     TTttttgaaaaattt---
```

6) Cel WBGene00016763 (C49A9.9a.1) – Intron 2
Cbr WBGene00027821 – Intron 6 (Reverse complement)

Folding of the *C. elegans* sequence

```
attgaaatgcaaaatacaataacatgaaaggtgaattgttggttttacaatttgttatcagttcgAGAGTA
...(((.((.((((..((((((((((.....)))..))))))..)))))..)).)))............
tatttatgcgtcatatgcttcttgtttgcaccgaaaatgaaagactgatataaatgaatcACATTT
(((((((..((((....(((.(..(((.......)))..)))))..)))))))))))....
```

Alignment of *C. elegans* and *C. briggsae* sequence

```
Cel      ----------------------------------attgaaatgcaaaatacaataacatg-
                                              ||| | || ||  || |   ||
Cbr      aacgactcctgagaatagagtcagaattcagtacaaaaaacggaagat-taaaattgtga

Cel      aaa-ggtgaattgttggttttacaatttgttatcagttcgAGAGTAtatttatgcgtca
         |||  | ||||| | || ||||  || || ||   | ||||||   ||  | |||  ||
Cbr      aaattgagaatttgggattttggaagttttgaaaattcgAG-TAAttctgatggatc-

Cel      tatgcttcttg--tttgcaccgaaaatgaaagactg-atataa--atgaatcACATTTtt
           || || |||  ||| |    |||||| || | || | ||||  | ||| |||| |||
Cbr      gctgatt-ttgatttttctggaaaaatggaataatgaagataatgaagaaacACAATTtg
```

7) Cel WBGene00002005 (F26D10.3.1) – Intron 2
Cbr WBGene00023842 – Intron 2

Folding of the *C. elegans* sequence

```
atcctcttctgttccggcttcgattcccgtagttttgccgatataaaatgaggaaATATAA
.(((((((.(((..((((...((((.....)))).)))..))))..))))).)).)))))).
ctctggcataaccctactgttgcgcttttatgtgttgtctgtttcttagtcgcctgtACAATT
....(((.(((....((.(..((((......)))).)))...).))...)))...)))...
```

Alignment of *C. elegans* and *C. briggsae* sequence

```
Cel      gtaagtttcaagaaaatttgtgttgttagtctgcttatcctcttctgttccggcttcga
                                      ||| | | || | ||||| ||||||||||||
Cbr      gtaagtagctgaattaatagc-----ttatttt-ctcaacctctactgttccggcttcag

Cel      ttcccgtagtttttgccgATATAAAATGAggaaATATAActctgg-cataaccctactgt
         || | ||||| | |||||||||||||||||||| |||  | |     |||||||||| |||
Cbr      tttctgtagtgtatgccgATATAATATGAggatATACTA-taaggacataaccctagtgt

Cel      tgcgctttta-tgtgttgtctgtttcttagtcgcctgtACAATT-aatgcaataaaatct
         | ||||||   || ||||||  |||||||||| | | ||||||
Cbr      t-cgctttcggtgcgttgtcactttcttagtcgtccgaACAATT----tcttcaaaa-cg
```

8) Cel WBGene00010976 (R02D5.1) – Intron 1
Cbr WBGene00027237 – Intron 1

Folding of the *C. elegans* sequence

```
ccaattaatcgtttatgcgcgacgagggcatactatgtacaacaagtgcgtcccctccgtaaagggaaagtccttgAAAAGA
........((.(((.((((.((.(.((((.((((.((.....)))))).)))))).)))))))))..)).....  ...
acatctgtccctttcaggccctggtttcatgtctcgaaagacgctcatgaatactatgcatactatgaattaaaaACAaaa
............((((.((..((((((((((.(((....)).)..)))))).))))).)).....)))).....
```

Alignment of *C. elegans* and *C. briggsae* sequence

```
Cel    acaatagccaattaatcgtttatgcgcgacgagggcatac-tatg-tacaacaagtgcgt
       ||||||||| |||||||||||||||||| ||||| |  | |  | |||||||||||| |
Cbr    acaatagcccattaatcgtttatgcgcgatgagggtagtcgtcgggtacaacaagtgtga

Cel    cccctccgtaaagggaaagtccttgAAAAGAAcatctgtccctttcaggccctggtttc
       ||| |  | |||      |||||||||||||||||||||||||||| | |     |
Cbr    cccgt--gaaaa------gtccttgACAAGAAcatctgtccctttaaagaagaagaaga

Cel    atgtctcgaaagacgctcatgaatactatgcatactatgaattaaaaACAAAaaaaatct
       || |  |   |        | |  |
Cbr    atctaacctcatctttatttcagaagACAACGatcacaaag-------------------
```

Supplemental Table S5. Conserved nested H/ACAs in *C. briggsae*

***Cbr* locus**	**ZM_ID**	**ceidn**	**Host Gene WB_ID**	**Rank**	**snoReport**	**BC**	**HP1**	**HP2**	
Sequences homologous to *C. elegans* RNA Training Set sequences									
29	Cb175	94	WBGene00023923	1	0.735	280	60	62	
57	CbN127	129	WBGene00031617	3	0.883	0	0	0	
103	-	148	WBGene00033070	1	0.694	23	35	56	
63	-	150	WBGene00036550	2	0.853	0	0	0	
62	-	151	WBGene00036550	6	0.680	147	53	72	
64	CbN105	127	WBGene00037490	2	0.796	42	57	77	
112*	-	94	WBGene00040082	6	0	0	0	0	
94	-	152	WBGene00024993	1	0.572	970	63	59	
43	Cb25	38	WBGene00025360	4	0.972	994	56	51	
37	Cb48	37	WBGene00025626	2	0.714	1000	63	57	
91	ComCb11	22	WBGene00026205	2	0.869	1000	55	58	
8	comCb6	21	WBGene00026668	1	0	847	61	54	
145	comCb5	144	WBGene00028008	1	0.945	992	62	54	
86	Cb36	111	WBGene00028531	2	0	575	52	47	
48	CbN102	125	WBGene00029804	1	0.534	847	62	49	
53	Cb345	42	WBGene00030667	1	0.766	999	51	57	
52	Cb182	43	WBGene00030667	2	0	999	68	58	
133	ComCb13	97	WBGene00031151	1	0	502	76	50	
134	Cb222	97	WBGene00031151	5	0.944	997	64	50	
135	ComCb14	18	WBGene00031151	6	0.949	818	53	61	
116	Cb375	53	WBGene00032455	6	0.997	1000	67	57	
16	Cb162	36	WBGene00033495	7	0.997	568	76	55	
59	Cb192	22	WBGene00036509	1	0.792	692	57	58	
60	Cb170	45	WBGene00036509	2	0.994	792	53	57	
61	Cb283	45	WBGene00036509	3	0.994	1000	53	57	
111	CbN45	131	WBGene00036739	4	0.827	1000	63	55	
100	Cb280	21	WBGene00038862	2	0.955	992	60	57	
107	Cb286	58	WBGene00039451	4	0.906	999	62	60	
114	Cb352	62	WBGene00040086	1	0	709	59	55	
113	Cb248	62	WBGene00040086	2	0.948	868	64	52	
115	Cb58	51	WBGene00040096	2	0.935	1000	52	55	
2	comCb8	33	WBGene00040684	1	0.960	998	55	50	
3	comCb9	33	WBGene00040684	2	0.990	998	73	57	
4	comCb7	145	WBGene00040874	3	0	992	58	73	
141	Cb309	19	WBGene00041282	3	0.974	994	67	59	
130	Cb323	109	WBGene00041491	5	0.797	847	63	57	
125	Cb166	91	WBGene00041875	9	0.939	997	59	51	
108	Cb375	53	WBGene00042822	6	0.997	1000	67	57	
Sequences not homologous to *C. elegans* RNA Training Set sequences									
95	Cb210	50	WBGene00024877	1	0	0	0	0	
15	-	100	WBGene00025340	1	0	0	0	0	
31	Cb356	110	WBGene00026009	1	0	280	40	74	
69*	-	29	WBGene00029132	10	0	0	0	0	
139	-	47	WBGene00032704	5	0	0	0	0	
17*	-	47	WBGene00033628	6	0	0	0	0	
65	Cb141	86	WBGene00037511	1	0	0	0	0	
0	CbN67	10	WBGene00037824	8	0	0	0	0	
21*	-	100	WBGene00038916	1	0	0	0	0	
84	Cb23	98	WBGene00040424	5	0	280	59	60	
71	CbN91	47	WBGene00023533	5	0	925	66	58	
30	Cb273	106	WBGene00026009	4	0	997	62	75	
146	Cb149	87	WBGene00028097	3	0.985	994	62	55	
87	Cb57	112	WBGene00028466	8	0.989	1000	53	59	
54	-	149	WBGene00030651	3	0.959	994	54	56	

131	CbN97	135	WBGene00041491	4	0	1000	40	13	

Supplemental Table S5. The list of *C. briggsae* loci homologous to *C. elegans* H/ACA snoRNAs, including several which have been previously predicted by Zemann et al. 2006 ("ZM_ID"). "ceidn": denotes the homologous *C. elegans* H/ACA sequence (see Supplemental Table S3). "Host Gene WB_ID": Wormbase Gene ID for the host gene, and "Rank" is the rank of the intron containing the homologous H/ACA. The score from applying snoReport ("snoReport") and the Bayesian Classifier ("BC") are also included, only the highest score for each intron is given. HP1 and HP2 are the hairpin lengths of the candidate sequence predicted by the Bayesian Classifier, calculated using RNAfold. The table is divided into two parts, first part includes sequences homologous to the *C. elegans* RNA Training Set, and the second part is homologous to sequences not within the RNA Training Set.

References

Bachellerie J, Cavaillé J, Hüttenhofer A. 2002. The expanding snoRNA world. *Biochimie* **84:** 775-790.

Deng W, Zhu X, Skogerbø G, Zhao Y, Fu Z, Wang Y, He H, Cai L, Sun H, Liu C, et al. 2006. Organization of the *Caenorhabditis elegans* small non-coding transcriptome: Genomic features, biogenesis, and expression. *Genome Research* **16:** 20-29.

Henras AK, Dez C, Henry, Y. 2004. RNA structure and function in C/D and H/ACA s(no)RNPs. *Current Opinion in Structural Biology* **14:** 335-343.

Higa S, Maeda N, Kenmochi N, Tanaka, T. 2002. Location of 2-O-methyl nucleotides in 26S rRNA and methylation guide snoRNAs in *Caenorhabditis elegans*. *Biochemical and Biophysical Research Communications* **297:** 1344-1349.

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie* (*Chemical Monthly*) **125:** 167-188.

Huang Z, Chen C, Zhou H, Li B, Qu L. 2007. A combined computational and experimental analysis of two families of snoRNA genes from *Caenorhabditis elegans*, revealing the expression and evolution pattern of snoRNAs in nematodes. *Genomics* **89:** 490-501.

Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson, GS. 2001. A Gene Expression Map for *Caenorhabditis elegans*. *Science* **293:** 2087-2092.

Wachi M, Ogawa T, Yokoyama K, Hokii Y, Shimoyama M, Muto A, Ushida, C. 2004. Isolation of eight novel *Caenorhabditis elegans* small RNAs. *Gene* **335:** 47-56.

Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naïve Bayesian Classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* **73:** 5261-5267.

Zemann A, op de Bekke A, Kiefmann M, Brosius J, Schmitz J. 2006. Evolution of small nucleolar RNAs in nematodes. *Nucleic Acids Research* **34:** 2676-2685.