Published online 22 June 2007

Nucleic Acids Research, 2007, Vol. 35, No. 14 ***4619–4628***
doi:10.1093/nar/gkm474

Combined experimental and computational approach to identify non-protein-coding RNAs in the deep-branching eukaryote *Giardia intestinalis*

Xiaowei (Sylvia) Chen[1], Timofey S. Rozhdestvensky[2], Lesley J. Collins[1,*], Jürgen Schmitz[2] and David Penny[1]

[1]Allan Wilson Centre, IMBS, Massey University, Palmerston North, New Zealand and [2]Institute of Experimental Pathology (ZMBE), University of Münster, Münster, Germany

Received February 8, 2007; Revised and Accepted May 29, 2007

ABSTRACT

Non-protein-coding RNAs represent a large proportion of transcribed sequences in eukaryotes. These RNAs often function in large RNA–protein complexes, which are catalysts in various RNA-processing pathways. As RNA processing has become an increasingly important area of research, numerous non-messenger RNAs have been uncovered in all the model eukaryotic organisms. However, knowledge on RNA processing in deep-branching eukaryotes is still limited. This study focuses on the identification of non-protein-coding RNAs from the diplomonad parasite *Giardia intestinalis*, showing that a combined experimental and computational search strategy is a fast method of screening reduced or compact genomes. The analysis of our *Giardia* cDNA library has uncovered 31 novel candidates, including C/D-box and H/ACA box snoRNAs, as well as an unusual transcript of RNase P, and double-stranded RNAs. Subsequent computational analysis has revealed additional putative C/D-box snoRNAs. Our results will lead towards a future understanding of RNA metabolism in the deep-branching eukaryote *Giardia*, as more ncRNAs are characterized.

INTRODUCTION

In the recent past, experimental and computational approaches have identified a vast variety of non-protein-coding RNAs (1), generally abbreviated as non-coding RNAs (ncRNAs), from both unicellular and multicellular eukaryotes. Many ncRNAs in modern eukaryotes function in RNA–protein complexes within which the RNAs may have direct regulatory roles at the reaction centres (1). The size of many ncRNAs is small compared with protein-coding RNAs, and lack of sequence homology often results in difficulties of identifying ncRNAs in distant eukaryotes through purely biological or computational approaches. In this study, our combined experimental and computational approach has been successful in finding novel ncRNAs in the distant eukaryote *Giardia intestinalis*.

Eukaryotic genomes are rich in non-protein-coding sequences. Large-scale cDNA cloning studies have shown that a large proportion of mammalian RNA transcripts do not appear to encode proteins (2), and an increasing number of ncRNAs have been shown to be functional (1). The origin of ncRNA is likely to date back to the earliest events when life emerged on earth. The theory of the 'RNA-World' (3,4) suggests that self-replicating RNAs are older than protein or DNA. The versatile features of RNA molecules support this hypothesis: first, RNA stores information in the same way as DNA; second, single-stranded RNA molecules are highly flexible to form secondary or tertiary structures, like peptides, they can form enclosed reactive centres and catalyze biological reactions in liquid environment. However, modern natural ribozymes have limited catalytic abilities, as natural ribozymes only perform ligation and/or nucleic acid cleavage reactions. These reactions are normally not limited by the rate of the catalytic reaction (5). Therefore, it is assumed that most ancient ribozymes have gradually been replaced by protein enzymes (5).

On the other hand, the evolution of ncRNAs has been continuous, and functions of ncRNAs have been diversifying throughout the evolution of eukaryotes. Based on structural and functional definition, eukaryotes have several distinct classes of ncRNAs, which form complex RNA-processing networks. Table 1 shows that each type of RNA often participates in the modification of another type of RNA, and the whole network fits into the general RNA-processing cascade (6). It is necessary to provide some brief background on the types of ncRNAs here,

*To whom Correspondence should be addressed. Tel: +64 6 350 9099-7345; Fax: +64 6 350 5626; Email: l.j.collins@massey.ac.nz

Table 1. A brief summary of ncRNAs in the RNA processing network of eukaryotes

Role	Type of ncRNA	Function
Transcriptional initiation	7SK snRNA (in mammals)	Inhibits transcription by binding to CDK/cyclin kinase complex
Intron splicing	U snRNAs	Function in the catalytic cores of major and minor spliceosomes involving in excision of introns
mRNA degradation	Micro RNAs	Guide the RNAi machinery to homologous mRNAs and trigger mRNA degradation
tRNA processing	RNase P	Involves in 5′ end nuclease activity in pre-tRNA processing
rRNA processing	MRP RNA	Involves in the endonuclease activity in pre-rRNA processing
	C/D box snoRNAs	2′-O-methylation guide
	H/ACA box snoRNAs	Pseudouridylation guide

because in this study, we have characterized a number of different types of ncRNAs from *Giardia*.

Probably the best studied ncRNAs are uridine-rich spliceosomal snRNAs (U-snRNAs). They function in the catalytic centre of major and minor spliceosomes. The major spliceosome that splices the majority of eukaryotic introns, consists of 5 U-snRNAs (U1, U2, U4, U5 and U6) and over 200 proteins (7). The minor spliceosome is low-abundant machinery containing U11 and U12 snRNAs instead of U1 and U2, and splices a 'minor' (less frequent) class of introns (8). Both major and minor spliceosomes may be ancestral to eukaryotes because they have now been identified in animals, plants, fungi and recently some distantly related protists (9,10).

The small nucleolar RNAs (snoRNAs) are involved in rRNA biogenesis. An increasing number of novel snoRNAs have been widely identified and have been reviewed in detail (11–15). Based on their structural motifs, snoRNAs are divided into two classes: C/D-box 2′-O-methylation snoRNAs and H/ACA-box pseudouridylation snoRNAs. The snoRNAs bind near the sites of modification through antisense recognition, and guide protein enzymes to the sites of editing. In addition, the functions of snoRNAs can be extended to acting as general chaperones targeting other nuclear or cellular RNAs (16–18).

There are a number of larger ncRNAs (>300 nt) such as the RNase P and RNase MRP RNAs. To date, besides the ribosome, RNase P is the only ribozyme required in both eukaryotes and prokaryotes (19). Eukaryotes have another related ribonuclease, RNase MRP, which processes a specific site in the pre-rRNA which is not found in prokaryotes, however, it seems likely that it is present in all eukaryotic lineages (6). Structural analysis of RNase P RNAs from phylogenetically diverse eukaryotes reveal a very similar minimum core (20). The overall structure of the RNA subunit in RNase MRP is similar to that of RNase P (21), and also the MRP enzyme shares a number of proteins with RNase P (22).

The smallest ncRNAs are micro RNAs (miRNAs) with length ranging from 21–25 nt and function in a variety of gene silencing pathways (23). About 800 miRNAs from different animals and plants have been reported (24). miRNAs from animals are usually transcribed as long and often polycistronic precursors, and then processed into small hairpin intermediates, which are then cleaved by a conserved protein Dicer (25) into mature miRNAs. The Dicer protein has been well studied for *Giardia* (26).

Recently, new experimental and bioinformatic approaches have identified a great number of novel ncRNA candidates from many organisms including: bacteria (27), animals (28–31) plants (32) and protists (33). The most widely used experimental method for identifying novel RNA candidates is based on size-selected cDNA libraries. Since most mRNAs have lengths greater than 500 nt, it is possible to isolate the majority of ncRNAs by size fractionation on a denaturing PAGE gel. Several methods are available to generate cDNAs from purified RNAs including the addition of poly(C)/poly(A) tail, and adaptor ligation at 5′-end and/or 3′-end, followed by reverse transcription, cloning and cycle sequencing (34). Here, we have constructed a cDNA library for ncRNAs from the deep-branching eukaryote *G. intestinalis*. *Giardia*, a parasitic diplomonad, is phylogenetically distant to all model eukaryotes (35,36). This unicellular organism has reduced mitochondria (mitosomes) and lacks hydrogenosomes (37). Two spliceosomal introns have been found (38,39), as well as several spliceosomal proteins (9) which strongly suggests that *Giardia* has a functional spliceosome. To date, several studies have identified 24 sno-like RNAs and the RNaseP of *Giardia* (40–42). However, there is little systematic research reported for the RNomics of *Giardia*.

We have screened our *Giardia* cDNA library, resulting in 31 novel candidates, within which, three are possibly C/D-box snoRNAs, one is possibly an H/ACA box snoRNA, and one is a fragment of the RNase P RNA. A computational study using known *Giardia*'s C/D box snoRNAs has resulted in new putative snoRNAs. In addition, an extended transcript has been found for the RNase P RNA, and two unusual self-cleaving dsRNA candidates have been studied. Given its proposed basal position on the eukaryotic tree (36), *Giardia* is evolutionarily distant to all the eukaryotic species, and probably highly reduced. It is not surprising to see that there may be some different RNA processing components in this organism. Future comparison of RNA-processing between *Giardia* and other eukaryotes is very necessary in understanding the evolution of RNA metabolism in reduced organisms (43). RNA processing in *Giardia* is expected to have changed in both the RNA and protein components as a result of genome reduction (43) due to

the parasitic nature of this organism. Our study moves towards understanding differences in *Giardia* RNA-processing machinery from that other eukaryotes which to date is largely confined to model, well-studied eukaryotes.

MATERIALS AND METHODS

Preparation of total RNA from *G. intestinalis* WB strain Trophozoites

Cells were collected from TY1-S-33 growth media at a concentration of 1.4×10^7 cells/ml by centrifugation (10 min, 2500 r.p.m., 4°C). Total RNA was prepared using Trizol reagent (Invitrogen) according to the protocol provided by the manufacturer.

cDNA library construction

Total RNA (10 μg) was run on an 8% denaturing PAGE gel (7 M urea, 1 × TBE buffer). RNA in the range of 70–600 nt was excised and eluted in 0.3 M NaOAc overnight at 4°C. Subsequently, 10 μg RNA was treated with tobacco acid pyrophosphatase (Epicenter) for 1 h at 37°C, then C-tailed by poly-A polymerase (Invitrogen) for 2 h at 37°C.

A 5′ DNA Sal-1 adaptor (5′-CAACGCGTCGACTAC GTGAGATTTGAGGTTC-3′) was then ligated to the RNA using T4 RNA ligase at 4°C overnight. First-strand cDNA synthesis was performed using the Thermoscript cDNA synthesis kit (Invitrogen) with Not-1 primer (5′-GACTAGTTCTAGATCGCGAGCGGCCGCCCG GGGGGGGGGGGGGG-3′).

The RNA-cDNA mix was treated with RNase A and PCR amplified using Sal-1 and Not-1 primers using a Biometra thermocycler. The PCR product was then double digested by Sal-1 and Not-1 restriction enzymes and ligated into the pSPORT1 vector (Invitrogen), followed by transformation into *Eschericia coli* Top10 cells (Invitrogen).

Sequencing

E. coli cells were grown on LB agar plates (100 μg/ml Ampicilin) at 37°C overnight. Colonies were PCR amplified using the M13for and M13rev primers (Roche Taq polymerase):

M13for: 5′-CGCCAGGGTTTTCCCAGTCACGAC-3′
M13rev: 5′-AGCGGATAACAATTTCACACAGG -3′

PCR products were cleaned by SAP/EXO-1 (GE Healthcare) treatment and cycle sequenced using BigDye Terminator version 3.1 and M13rev primer. The sequencing products were cleaned using CleanSeq (Agencourt) magnetic beads, and capillary sequenced on a capillary ABI3730 Genetic Analyzer (Applied Biosystems Inc.).

Computational analysis

The sequences were assembled using DNAMAN 5.2 and DNASTAR 5.0 packages, and were then blasted against the *Giardia* genomic database (http://www.mbl.edu/Giardia) as well as the NCBI databases (http://www.ncbi.nlm.nih.gov). Putative snoRNA prediction used the modified Snoscan program (Snoscan-G) in C for Windows (the original source code is available at http://lowelab.ucsc.edu/snoscan/). However, the C-box scoring function was modified so that it read user-specified input of the C-box scoring matrix.

RNA structures were generated using the RNAfold program from the Vienna-RNA-1.4 Package (http://www.tbi.univie.ac.at/~ivo/RNA/windoze/), structural alignment was done using RSmatch1.0 converted for Windows (original program is available at http://exon.umdnj.edu/software/RSmatch/) and FoldalignM (http://foldalign.ku.dk/software/index.html).rRNA sequence alignments for preliminary methylation site analysis were generated using ClustalW (44).

RT-PCR and PCR. RT-PCR reactions used Invitrogen Thermoscript first strand cDNA synthesis kit and subsequent PCR reactions used Roche Taq polymerase. Primers:

U5For: 5′- CATTCATCTCTGCGGTGGATG -3′
U5Rev:5′-ACCCCAAAAAATGCAACTGTCTGCC-3′
U6For: 5′- CAAATTGAAACGATACAGAG -3′
U6Rev: 5′- TCATCCTTGTGCAGGGGCCA -3′
testP/GlsR15_For: 5′- GGGGAAGGTCTGAGGTC ATT -3′
testP/GlsR15_Rev: 5′- AGCTCATAGTCGTGCTTG CTC -3′

In vitro transcription and RNA self-cleavage assays. In vitro transcription reactions used the Invitrogen T7 RNA polymerase kit to add T7 promoter sequences to the 5′ and 3′ ends of the PCR products. The RNA products from *in vitro* transcription were heated to 80°C for 5 min and gradually cooled down to anneal. The dsRNAs were then purified using Roche PCR product purification kit. All the self-cleavage reactions were carried at 37°C for 2 h.

Primers used for generating templates for *in vitro* transcription:

Genie1_T7_For: 5′- TAATACGACTCACTATAGGG AGACGACCCTCTTCTCCAGCA -3′
Genie1_T7_Rev: 5′- TAATACGACTCACTATAGGG AGAGGAGCGCAAAGAGGATGA -3′
Girep1_T7_For: 5′- TAATACGACTCACTATAGGG AGATGCAGCCCTTCTTGTCC -3′
Girep1_T7_Rev: 5′- TAATACGACTCACTATAGGG AGAGATACCCGGCTGTGC -3′

RESULTS

Assembly of cDNA sequences from the RNA library

Assembly of the cDNA sequences resulted in 31 novel ncRNAs, 15 previously known snoRNAs (40–42) and 10 out of 48 characterized tRNAs (http://www.mbl.edu/Giardia). Candidates were obtained in the following manner. A total of 616 initial sequences were assembled into 166 contigs and each contig was blasted against the *Giardia* genome database and NCBI databases to screen for easily characterized RNAs. After discarding empty vector contaminants, sequences below the length of 20 nt

and *E. coli* contaminant sequences, the remaining 152 contigs (including repeats or duplicates) contained 33 mRNA fragments, 28 known tRNA sequences, 10 5.8S rRNA sequences, 7 LSU rRNA and SSU rRNA fragments, 29 known ncRNAs sequences and 45 unknown sequences. All the unknown sequences were further analysed so that any broken fragments of a single RNA could be reassembled into a complete sequence, leaving 31 novel RNA candidates. Details of candidate sequences and features are listed in Supplementary Data. In order to carry out further computational analysis, 5′-and 3′-extensions (200 nt from each end) were extracted from the genome database for each candidate.

New C/D box snoRNA candidates and putative snoRNAs from computational studies

Eukaryotic 2′-O-methylation C/D box snoRNAs are characterized by two short sequence motifs near their 5′-and 3′-termini: C-box (‘5′-AUGAUGA-3′’) and D-box (‘5′-CUGA-3′’), which are brought together by a short (4–8) terminal stem (45). There are one or two 10–20 nt antisense guide elements immediately upstream of the D-box or D’-box, and these elements bind to complementary sequences on rRNAs spanning the methylation sites (46). The position of the nucleotide which is methylated is usually the fifth position upstream of the D-box or D’-box (47).

Since the *Giardia* genome is fully sequenced (NCBI accession number: AACB00000000), it is possible to check our experimentally found RNAs for snoRNA features using potential interactions to rRNA sequences. Once we identify the conserved features of a *Giardia* snoRNA, we can identify more snoRNAs using a computational search. However, to date there are no full-length rRNA large subunit and small subunit rRNA sequences available for *Giardia*. Raw sequence reads from the GiardiaDB (http://www.mbl.edu/Giardia) were pulled out individually and assembled using SeqMan. Three contigs were generated, and correspond to the large subunit (LSU), small subunit (SSU) and 5.8S rRNAs, with lengths of 2908, 1449 and 138 nt respectively, and they arrange in the typical eukaryotic rRNA-gene order of SSU-5.8S-LSU, which reveals a site of cleavage by RNase MRP (6). The sequences are listed in Supplementary Data. Shortened lengths of the *Giardia* rRNAs are consistent with an earlier study (48) that *Giardia*’s rRNAs are much shorter than usual eukaryotic rRNAs, and unlike other eukaryotes, *Giardia* does not appear to have the 5S rRNA (48), which was also not found during our searches. The snoRNA search was done using modified source code of the Snoscan program, which was originally used to identify a large number of C/D-box snoRNAs from *Saccharomyces cerevisiae* (49).

We have predicted 3 C/D box snoRNA candidates from the 31 novel candidate sequences. Of the 15 known snoRNAs (40–42) that were found in our cDNA library, 14 are C/D box snoRNAs and 1 is an H/ACA box snoRNA. Comparing all the available C/D box, snoRNA sequences revealed that snoRNAs from *Giardia* share common sequence features within the

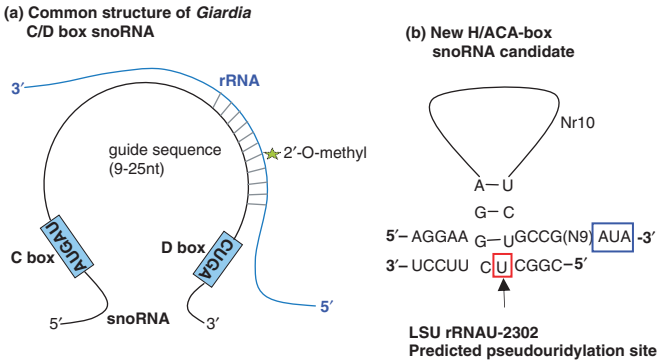

Figure 1. (**a**) Conserved structure of C/D box methylation snoRNAs in *Giardia*. (**b**) Structural prediction of the new H/ACA-box snoRNA candidate.

C boxes and D boxes. All but one of the confirmed C/D box snoRNAs has a perfect ‘CUGA’ D-box near the end of the 3′-end, and most *Giardia* C-boxes have a conserved sequence ‘5′-AUGAU-3′’ allowing one mismatch at either 5′-or 3′-end. The C-box sequences also appear more variable as their lengths range between 5 and 7 nt. The C-box scoring function of Snoscan was adjusted to use the *Giardia* consensus sequence. The C’-box is generally missing or poorly identifiable, and the existence of D’-box is not essential. The length between the C- and D-boxes is varying from 28 to 124 nt. In addition, few of the known *Giardia* C/D box snoRNAs have a terminal stem.

The general structure of *Giardia* C/D box snoRNAs during rRNA modification is shown in Figure 1a. Structural alignment was done on all the experimentally found *Giardia* C/D box snoRNAs using RNA structures generated from Vienna-RNAfold program, but the result did not indicate any additional consensus motifs. Therefore, no further structural features were incorporated into modifying the Snoscan program. Our modified Snoscan program, Snoscan-G, identified 13 out of 18 confirmed C/D-box snoRNAs with the following parameters: cutoff total score (10), C-box score (2.0) and the maximum distant between C and D boxes (150 bp). The others were not recovered due to poorly defined C-boxes or imperfect D-boxes. This testing indicated that it was possible to identify additional C/D-box snoRNAs from the *Giardia* genome with this computational method. Table 2 shows the range of scores obtained from experimentally identified snoRNAs. These are considered as standard scores for *Giardia*, thus used to compare with the scores generated for computationally predicted snoRNAs further on. We refer to these computationally predicted snoRNAs as ‘putative’ snoRNAs in order to distinguish them from the ‘candidate’ snoRNAs found experimentally.

Due to the short (5 nt) and less conserved *Giardia* C-box, large volume of output was expected. A whole genome search for C/D box snoRNAs using the same parameter settings yielded many (6280) non-repetitive putative candidates, which were subsequently analysed through a strict three-step post-scan filtering.

Table 2. Snoscan scores obtained for experimentally identified *Giardia* C/D box snoRNAs

Feature	Consensus	Best score	Average score	Worst score
C box	AUGAU(GA)	8.76	7.9	3.55
D box	CUGA	8.05	7.9	3.77
D' box	CUGA	7.34	4.8	0.59
rRNA complement	9–25 nt with 1 or 2 mismatches	33.93	22.7	15.92
Total score		21.05	12.4	10

Three features of the putative snoRNAs were looked for during the post-scan filtering:

(i) The sequences should locate in the non-coding regions.
(ii) The sequences should locate close to reading frames since *Giardia* appears not to have separate transcription start sites for snoRNAs.
(iii) The C-boxes of putative snoRNAs are more similar to the experimentally confirmed snoRNAs in *Giardia* (41).

All the output sequences from Snoscan-G were compared against the database of *Giardia* open-reading frames (ORFs) downloaded from GiardiaDB (http://www.mbl.edu/Giardia) to exclude possible mRNA sequences. These ORF datasets have been expertly compiled using software such as GLIMMER and CRITICA with parameters adjusted for this unique eukaryote. Our search of this database implicitly filtered out putative candidates with obvious coding potential. The status of the *Giardia* genome is such that a large number of ORFs remain hypothetical. Any explicit assessment for coding potential could be on only a subset of highly conserved proteins and would not be representative of the entire *Giardia* proteome. Hence, the use of this database maximizes our exclusion of contaminant mRNAs.

Unlike other eukaryotes, *Giardia* has only two confirmed introns (38,39), and most ncRNAs characterized to date are located between protein-coding genes, with a small number (less than 10) of them located on the minus strand of protein-coding genes. To exclude any ambiguities, only sequences located between protein-coding genes were considered. Sequence searches showed that most of the Snoscan-G outputs (5857) had full-length 100% match to ORFs, leaving 423 potential putative snoRNAs. After excluding shorter partial sequences and repetitive sequences with different names, 357 sequences remained. To date, all 13 experimentally confirmed C/D-box snoRNAs that had been detected in the small-scale Snoscan-G testing were also found in this large-scale genome search.

It was noticeable that all the experimentally characterized snoRNAs were located in ORF-rich regions of the genome, which could due to the fact that these snoRNAs do not seem to possess their own promoters. Therefore, further screening was done based on genomic location. Only putative sequences that are located near ORFs were selected with those appearing in heterochromatic regions excluded because they are less likely to be transcribed. This screening left 101 putative snoRNA sequences. Strict post-scan filtering based on C-box and D-box sequences was then done so that only sequences with 'AUGAU' or 'GUGAU' in C-box and 'CUGA' in D-box were considered as highly likely putative snoRNAs. In the end there were 60 strong putative snoRNAs. All sequences had distinct C-box and D-box motifs and fulfill the criteria for *Giardia* snoRNAs (41,45). In addition, they had average Snoscan-G total score of 12.5, which was slightly above the average total score of experimentally identified snoRNAs. The details of candidates are shown in the Supplementary Data.

As a control, we generated a random database with its size equivalent to *Giardia* genome using a third-order Markov chain based on 4-mer frequencies (49) within the *Giardia* genome. A search of this random sequence database yielded 6721 false positives with an average score of 11.8 and a best score of 25.26. As downstream filtering based on genomic location was impossible to carry out on randomized data, only the last step of the three-stage filtering could be performed on this output. Therefore, a parallel comparison between the *Giardia* Snoscan-G outputs and the randomized data outputs was not entirely applicable since the first two steps of the post-scan filtering were the most important and based on *Giardia* genomic information. However, a strict scan was still performed on this output with more stringent parameter settings based on C-box and D-box motifs, as was done in the final stage of post-scan filtering described above, reduced the positives down to 89 non-overlapping ones. Although these outputs contain C-box and D-box motifs, they do not represent comprehensive data for comparisons. In all, the purpose of generating a randomized dataset was to show that post-scan using genomic information was necessary to improve the selection of putative snoRNAs in a distant organism such as *Giardia*.

To test if the large number of initial output from the random database was due to special features within the *Giardia* genome, another Snoscan-G was run on a partial yeast genome (with a size similar to *Giardia* genome) using the same parameter settings. There were 1756 non-repetitive outputs. This test showed that the *Giardia* genome has less regional variation in its sequence, and this may result in the observation of more false positives. This testing showed that it was necessary to carry out stringent downstream filtering as was done in our Snoscan-G of the *Giardia* genome to obtain acceptable putative snoRNAs.

As an additional analysis, human and yeast C/D box snoRNAs have been mapped onto *Giardia* rRNAs (alignments included in Supplementary Data). Since human and yeast are extremely evolutionarily distant from *Giardia*, most known methylation sites do not have homologues in *Giardia*, apart from two. ncRNA candidate-1 from our cDNA library is predicted to guide methylation of G_{1131} on SSU-rRNA, which corresponds to the site of modification by human U25 snoRNA. Snoscan-G predicted putative snoRNA U0025 is likely to

guide methylation of C_{1191} on LSU-rRNA, which corresponds to the site of modification by an undetected human snoRNA. However, as these alignments are between such diverse organisms, no extensive conclusions can be drawn at this time.

In all, our Snoscan-G in combination of the post-scan filtering has identified 60 C/D-box snoRNA putative snoRNAs based on information from previously experimentally characterized snoRNAs. This approach was tested against two negative controls and showed that the use of *Giardia*-specific information made it possible to screen for functional ncRNAs in this reduced genome.

A new H/ACA box snoRNA candidate

The pseudouridinylation guide H/ACA box snoRNAs have a common secondary structure consisting of two parallel hairpins linked by a hinge. Two conserved motifs box H (ANANNA) and box ACA are located at the hinge and the 3′ tail, respectively, together with the flanking helix, they play important roles in box H/ACA snoRNA accumulation (50). However, compared to the single continuous antisense elements in box C/D snoRNAs, the antisense elements of H/ACA box snoRNAs are very short and bipartite (51). Almost all the H/ACA box snoRNA adopt the two hairpin model, except one small H/ACA box snoRNA containing only one hairpin described in *Trypanosoma* (52). Based on hallmark sequences and structural features, one of the identified potential novel ncRNA (candidate 16, Supplementary Data), is likely to represent a novel H/ACA box snoRNA. It features a single, long stem positioned upstream from the ACA box motif as shown in Figure 1b. As such, it is strongly reminiscent of archaeal and Trypanosomal H/ACA box snoRNAs, that also feature a single hairpin (52–54). In agreement with the rules applying to eukaryotic H/ACA snoRNAs, the targeted uridine is separated from the H/ACA box by 9–16 nt. Therefore, according to structural modelling, we predict that candidate_16 may guide a pseudouridylation in LSU rRNA.

RNase P

The ribozyme RNase P cleaves the 5′-end of pre-tRNAs. The *Giardia* RNase P RNA was recently identified by sequence similarity search and the RNase P holoenzyme was purified (20), and showed that *Giardia* RNase P RNA has the conserved eukaryotic RNase P core structure, and shared extensive similarity with the RNase P RNA of the microsporidian *Encephalitozoon cuniculi*. Both RNAs lack the conserved P3 helix bulge loop, which has been found in all the other eukaryotes studied so far. The RNase P RNA has been found in our library (candidate 9), but surprisingly, the sequence was not terminated at the previously predicted 3′ end, and extended further into the GlsR15 snoRNA (41). These two known RNAs have a 24 nt overlap, which is shown in Figure 2. It is likely that candidate 9 is part of a full-length RNA transcript. To verify this idea, RT-PCR was done using an upstream primer (testP/GlsR15_For) that binds within the RNase P sequence (position 34–53 on the possible full length transcript) and a downstream primer (testP/GlsR15_Rev) which binds within the GlsR15 snoRNA sequence (position 269–289 on the possible full length transcript).

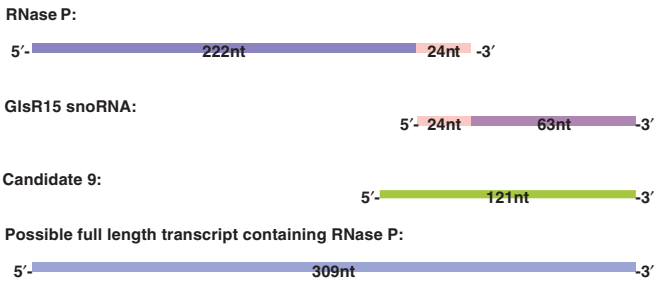

Figure 2. Comparison of RNase P, GlsR15 snoRNA and the new ncRNA candidate 9.

RT-PCR results (data not shown) indicate that the RNase P and GlsR15 are indeed transcribed as a single transcript. This rises to a question that whether this transcript is a single functional RNA molecule, or a precursor to give two different RNAs. Structural studies (20) indicate that the shorter transcript could fold with conserved eukaryotic RNase P motifs. Therefore, the second assumption is preferred. It is possible that an as yet unknown ribonuclease is involved in producing two different RNAs from one precursor. However, this leads to a result that only one of the two RNAs can be generated as a full-length molecule and the other one will be non-functional.

Transcribed intergenic repeats

A fragment of the variant surface protein (VSP) mRNA was found in the cDNA library. It has been suggested (55) that antisense regulation controls the expression of VSP genes, and the function of RNA-dependent RNA polymerase (RdRp) is involved to restrict the VSP gene repertoire to a single gene at any one time. Careful sequence mining within the *Giardia* genome observed that there were many tandem repeats sharing short sequence fragments, and these fragments are often complementary to repeated sequences in VSP genes and cysteine-rich protein genes. Blasting a VSP-fragment sequence found in our cDNA library against the *Giardia* genome yielded a potentially functional antisense element. This sequence is a long tandem repeat consisting of nine units, each containing one fragment complementary to the VSP ORF (Figure 3). RT-PCR was carried out targeting both the '+' and '−' strand of this sequence, and the results showed that both strands were transcribed, to give a double-stranded RNA product.

Unlike other tandem repeats of retrotransposons such as LINEs or SINEs, this tandem repeat shows no feature of any known retrotransposon. In comparison, there have been a few studies on unusual repeated sequences in *Giardia*: one study (56) showed a non-LTR element with site-specific tandem insertions in a chromosomal DNA repeat, and suggested that this element was unlikely to have evolved site specificity unless it did have a function. Another more recent study showed this element was transcribed into a dsRNA (57). In addition, there are

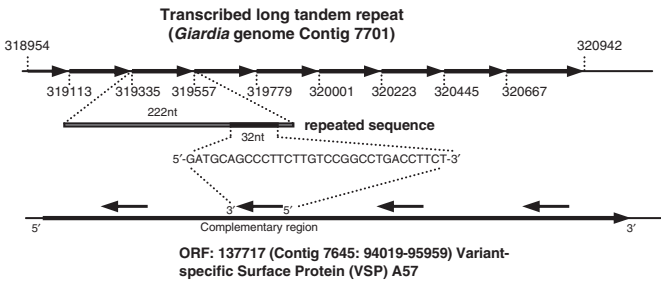

Figure 3. Tandem repeats of the Girep-1 RNA. Each fragment coloured in grey represents a repeating unit (222 nt in length, with the first unit lacking the 5′ 63 nt and the last unit extending 54 nt at 3′ end) on Girep-1 RNA. Each 32 nt fragment coloured in black represents the repeating Girep-1 sequence that is complementary to the various-surface-protein (VSP) gene.

22 antisense transcripts identified in the *Giardia* genome (www.mbl.edu/Giardia); however, there are no known functions of these transcripts.

Our study has revealed a surprising feature shared by two tandem repeats in *Giardia*: one repeat is the experimentally verified dsRNA with fragments complementary to the VSP (Rep-1); and the other is the non-LTR element Genie-1 (56). A partial sequence from each element was amplified by PCR with T7-promoter attached primers. The PCR products were transcribed by T7 RNA polymerase to produce dsRNAs. As a control, a single stranded Rep-1 RNA was also produced by elimination of T7 promoter sequence from the reverse primers. Both dsRNAs underwent one self-cleavage at roughly the middle of the sequence (under a basic assay condition with Mg^{2+} added to water or buffer) (Figure 4a). The single stranded Rep-1 control did not cleave (Figure 4b). Timing Mg^{2+} titration (Figure 4c) assay and divalent ion assays (Figure 4d) were performed with the Genie-1 dsRNA. Results showed that the self-cleavage did not happen when Mg^{2+} concentration was below 1 mM; and self-cleavage only happened at the present of Mg^{2+} or Co^{2+}, while Mn^{2+} and Ca^{2+} did not have any effect. In addition, addition of EDTA prevented Mg^{2+} induced cleavage. Further investigation will be necessary to analyse this unusual phenomenon.

DISCUSSION

Combined experimental and computational approach

The aim of this study was to explore the variety of ncRNAs in *Giardia* and obtain a view of ncRNA expression in this genomically reduced deep-branching eukaryote. The scale of this cDNA library is small compared with equivalent studies of ncRNAs in other organisms (28–32). However, studying on a relatively small scale can help getting a comprehensive view of the special features and conserved patterns within this organism, before any large scale studies are attempted. There were previously no systemic studies on the ncRNA composition of *Giardia*. As an extant group of eukaryotes, Diplomonads share very low sequence homology with other eukaryotes, which makes characterization of RNAs extremely difficult. From the 31 novel ncRNA candidates, only 3 can be identified by homology searching as C/D box snoRNAs, the rest have little similarities to known types of ncRNAs.

However, comparing the 18 characterized C/D box snoRNAs from *Giardia* has shown that these snoRNAs still share the basic conserved features seen with snoRNAs from other eukaryotes. This makes a computational screen possible. Within the computationally identified putative snoRNAs, we recovered 13 out of our control set of the 18 experimentally characterized snoRNAs. Snoscan-G used looser parameters than the original Snoscan program in order that the experimentally identified snoRNAs (13 in this study) were included in the results. This ensured the sensitivity of the algorithm which was then used for a whole-genome search. However, the large number of false positive hits obtained from the negative control search on a random database, indicated the requirement for other post-scan filtering of putative snoRNA sequences using data unable to be included in the Snoscan-G software. Also, a fairly large result obtained from scan of the yeast genome confirms that the parameter settings for Snoscan-G are less stringent than the original Snoscan program. Comparing putative snoRNA sequences against the ORF database excluded most of the first-round positive hits, and information from genomic locations of the sequences extended the reliability of the putative snoRNAs.

Possibly due to its reduced genome, *Giardia*'s snoRNAs are less conserved than those of other eukaryotic organisms; therefore it was necessary to apply less stringent searching criteria. This is because there are as yet no additional *Giardia*-specific sequence features, which can be incorporated into the algorithm. This explains the increase in false positives when large databases are screened. However, combining several filtering steps dramatically reduced the number of positive hits, and at the same time did not result in the loss of any true positives. The remaining putative snoRNAs showed greater similarities to the experimentally identified snoRNAs than the first-round Snoscan-G results before post-scan filtering. Therefore, our computational approach is reliable when used in parallel with an experimental approach speeding up the discovery of novel putative ncRNAs.

Encoding patterns of ncRNAs in *Giardia*

Blasting the novel RNA candidates against the *Giardia* genome revealed three types of encoding patterns.

(i) Most ncRNAs in *Giardia* are encoded as single copies between protein-coding genes. According to current knowledge of *Giardia*, almost all the protein-coding genes are intronless (38,39) and it becomes natural that ncRNAs find their places in intergenic regions. The genome of *Giardia* is compact; and the genes generally have very short gaps (often <200 nt) between one another. Almost all the ncRNAs observed so far are located in ORF-rich regions, but do not appear to possess their own promoters, although this may be due to the fact that *Giardia* does not appear to have well characterized

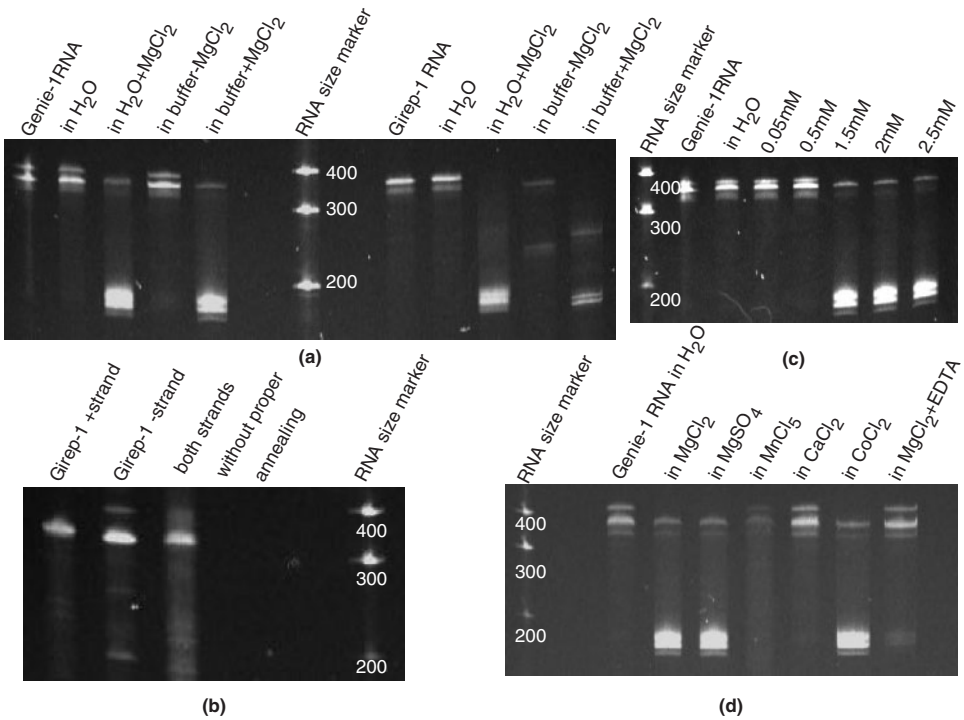

Figure 4. Self-cleavage reactions of the Genie-1 dsRNA and Girep-1 dsRNA. All the reactions were incubated at 37°C for 2 h, and run on 8% denaturing polyacrylamide gel containing 7 M urea at 350 V. (**a**) Self-cleavage reactions of dsGenie-1 RNA (left to the size marker) and dsGirep-1 RNA (right to the size marker); buffer: 20 mM HEPES, 150 mM NaCl, with or without 2.5 mM $MgCl_2$; (**b**) The test of ssRNA of Girep-1 in water with 2.5 mM $MgCl_2$; (**c**) Mg^{2+} titration assay of Genie-1 dsRNA and (**d**) the test of different ions with dsGenie-1 RNA in water with 2.5 mM of each divalent ion, and EDTA was added to 50 mM on the last lane.

promoter sequences as there is a lack of conserved sequence in the promoter region. One possibility is that these ncRNAs may co-transcribe with their adjacent ORFs, and the pre-transcripts are later processed to give mRNAs and ncRNAs. If this is the case, there must be specific RNA-processing machinery to carry out the task. One possible candidate is the spliceosome, as it is highly unlikely for a whole spliceosome to remain just for processing two introns (38,39).

(ii) Three novel candidates from the cDNA library show polymorphic variations in having several nearly identical copies in the genome with most of the polymorphic copies not located near predicted ORFs. It is not known if all the polymorphic copies of these RNAs are transcribed, because for each of the three candidates only one form has been seen in our cDNA library. Some of these polymorphic copies are encoded in tandem repeats, but the rest are located in a distant part of the genome. It has been known that some ncRNAs such as U2 snRNAs in *Xenopus* do have this feature in developmental regulation (58); however, the polymorphic forms of our candidates do not have any sequence similarity to known spliceosomal snRNAs.

(iii) Long retrotransposon-like tandem repeats of ncRNAs are described in the Results section. The experimentally confirmed tandem repeat is located in an ORF-rich area of the genome with both '+' and '−' strands adjacent to neighbouring protein-coding genes. We suggest that it is likely that they are co-transcribed with mRNAs and are subsequently cleaved by a specific but yet unknown mechanism. The novel self-cleaving feature of the dsRNAs derived from the two retrotransposon-like elements will require further investigation.

The puzzle of spliceosomal snRNAs in *Giardia*

There is very little known about splicing in *Giardia*. Sequence mining from the genome shows that most of the eukaryotic specific spliceosomal proteins (9) are present in *Giardia*, as well as the important U5 snRNA (59), which functions at the centre of both major and minor spliceosomes. It is common in eukaryotes that the spliceosomal snRNAs are expressed at a high level (60), since intron splicing generally occurs at a high rate. However, it seems not the case in *Giardia*. We did not find any sequence in our cDNA library with similarity to any known spliceosomal snRNA. To determine the possible presence of any spliceosomal snRNAs in the library, PCR reaction using the U5 primers (Materials and Methods section) was done on the cDNAs. Results show that U5 snRNA is expressed and present, but in very low quantities. Another puzzling question concerns the U6 snRNA. U6 snRNA is the most conserved spliceosomal snRNA across all the eukaryotes studied to date. U6 snRNAs take part in the actual catalysis during

splicing (61), and share extensive sequence similarities across eukaryotes. In an early study (62), it has been shown that a single pair of PCR primers could detect U6 snRNAs from 17 different species of eukaryotes. As a trial, the same pair of primers was tried on *Giardia* in both genomic PCR and RT-PCR reactions. Despite extensive effort, there is as yet no detectable candidates for a *Giardia* U6 snRNA. It is therefore concluded that our current approach is not powerful enough to solve the puzzle of *Giardia*'s spliceosomal snRNAs.

Novel ncRNA candidates

Total 26 out of our 31 novel RNA candidates cannot yet be extensively characterized as belonging to any known class of ncRNA; a feature seen in other species-specific studies (29). Structural studies and motif analysis of these RNAs did not show distinct features found in known ncRNAs. A number of these RNAs are GC rich, providing a basis for strong helical structures. Lack of characterization is possibly due to the highly divergent sequences of *Giardia* compared to those of the major eukaryotic groups, and because most computer programs developed for identifying ncRNAs are based on human and yeast. One way to further approach the identification of ncRNA is through more computational studies by incorporating more *Giardia*-specific information into the existing programs, followed by experimental verification of our proposed candidates. Another way is through biochemical studies of central protein components of various RNA processing pathways. These are to be investigated in the future.

In conclusion, our cDNA library successfully uncovered 31 novel ncRNAs from *Giardia*, and our computational approach was shown to be a useful method that worked well in parallel with an experimental approach to aid discovery of 60 potential putative snoRNAs in a deep-branching eukaryote. Although it is hard to characterize each candidate ncRNAs found from the cDNA library due to sequence divergence, as far as we can tell, *Giardia* has quite typical eukaryotic RNA processing despite being reduced and with many introns lost. The transcriptional patterns seen in these ncRNAs may help in understanding the mechanism of RNA processing. Future work will continue to be done in investigating the unusual properties of ncRNAs by combined biochemical and computational methods.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank George Ionas and Errol Kuan from Microaquatech for kind supply of *G. intestinalis* culture, Anja Zemann and Claudia Marker (ZMBE) for great help with the cDNA library and Timothy White (Allan Wilson Centre) for time and effort on computer programming. This work is supported by the New Zealand Marsden Fund; the Allan Wilson Centre for Molecular Ecology and Evolution; European Union (EU; LSHG-CT-2003-503022) and the Nationales Genomforschungsnetz (NGFN; 0313358A). Funding to pay the Open Access publication charges for this article was provided by the New Zealand Marsden Fund.

Conflict of interest statement. None declared.

REFERENCES

1. Mattick,J.S. and Makunin,I.V. (2006) Non-coding RNA. *Hum. Mol. Genet.*, **15 Spec No 1**, R17–R29.
2. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
3. Gilbert,W. (1986) The RNA world. *Nature*, **319**, 618.
4. Brosius,J. (2005) Echoes from the past – are we still in an RNP world? *Cytogenet. – Genome Res.*, **110**, 8–24.
5. Jeffares,D.C., Poole,A.M. and Penny,D. (1998) Relics from the RNA world. *J. Mol. Evol.*, **46**, 18–36.
6. Woodhams,M.D., Stadler,P.F., Penny,D. and Collins,L.J. (2007) RNase MRP and the RNA processing cascade in the eukaryotic ancestor. *BMC. Evol. Biol.*, **7**(Suppl. 1), S13.
7. Kramer,A. (1996) The structure and function of proteins involved in mammalian pre-mRNA splicing. *Annu. Rev. Biochem.*, **65**, 367–409.
8. Patel,A.A. and Steitz,J.A. (2003) Splicing double: insights from the second spliceosome. *Nat. Rev. Mol. Cell Biol.*, **4**, 960–970.
9. Collins,L. and Penny,D. (2005) Complex spliceosomal organization ancestral to extant eukaryotes. *Mol. Biol. Evol.*, **22**, 1053–1066.
10. Russell,A.G., Charette,J.M., Spencer,D.F. and Gray,M.W. (2006) An early evolutionary origin for the minor spliceosome. *Nature*, **443**, 863–866.
11. Bachellerie,J.P., Cavaille,J. and Huttenhofer,A. (2002) The expanding snoRNA world. *Biochimie*, **84**, 775–790.
12. Brown,J.W., Echeverria,M. and Qu,L.H. (2003) Plant snoRNAs: functional evolution and new modes of gene expression. *Trends Plant Sci.*, **8**, 42–49.
13. Uliel,S., Liang,X.H., Unger,R. and Michaeli,S. (2003) Small nucleolar RNAs that guide modification in trypanosomatids: repertoire, targets, genome organisation, and unique functions. *Int. J. Parasitol.*, **33**, 235–255.
14. Dennis,P.P. and Omer,A. (2005) Small non-coding RNAs in Archaea. *Curr. Opin. Microbiol.*, **8**, 685–694.
15. Mehler,M.F. and Mattick,J.S. (2006) Non-coding RNAs in the nervous system. *J. Physiol.*, **575**, 333–341.
16. Newby,M.I. and Greenbaum,N.L. (2001) A conserved pseudouridine modification in eukaryotic U2 snRNA induces a change in branch-site architecture. *RNA*, **7**, 833–845.
17. Watkins,N.J., Segault,V., Charpentier,B., Nottrott,S., Fabrizio,P., Bachi,A., Wilm,M., Rosbash,M., Branlant,C. *et al.* (2000) A common core RNP structure shared between the small nucleoar box C/D RNPs and the spliceosomal U4 snRNP. *Cell*, **103**, 457–466.
18. Staley,J.P. and Guthrie,C. (1998) Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell*, **92**, 315–326.
19. Frank,D.N. and Pace,N.R. (1998) Ribonuclease P: unity and diversity in a tRNA processing ribozyme. *Annu. Rev. Biochem.*, **67**, 153–180.
20. Marquez,S.M., Harris,J.K., Kelley,S.T., Brown,J.W., Dawson,S.C., Roberts,E.C. and Pace,N.R. (2005) Structural implications of novel diversity in eucaryal RNase P RNA. *RNA*, **11**, 739–751.
21. Collins,L.J., Moulton,V. and Penny,D. (2000) Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP. *J. Mol. Evol.*, **51**, 194–204.
22. Chamberlain,J.R., Lee,Y., Lane,W.S. and Engelke,D.R. (1998) Purification and characterization of the nuclear RNase P holoenzyme complex reveals extensive subunit overlap with RNase MRP. *Genes Dev.*, **12**, 1678–1690.

23. Hammond,S.M., Caudy,A.A. and Hannon,G.J. (2001) Post-transcriptional gene silencing by double-stranded RNA. *Nat. Rev. Genet.*, **2**, 110–119.
24. Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.
25. Bernstein,E., Caudy,A.A., Hammond,S.M. and Hannon,G.J. (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, **409**, 363–366.
26. Macrae,I.J., Zhou,K., Li,F., Repic,A., Brooks,A.N., Cande,W.Z., Adams,P.D. and Doudna,J.A. (2006) Structural basis for double-stranded RNA processing by Dicer. *Science*, **311**, 195–198.
27. Argaman,L., Hershberg,R., Vogel,J., Bejerano,G., Wagner,E.G., Margalit,H. and Altuvia,S. (2001) Novel small RNA-encoding genes in the intergenic regions of Escherichia coli. *Curr. Biol.*, **11**, 941–950.
28. Yuan,G., Klambt,C., Bachellerie,J.P., Brosius,J. and Huttenhofer,A. (2003) RNomics in Drosophila melanogaster: identification of 66 candidates for novel non-messenger RNAs. *Nucleic Acids Res.*, **31**, 2495–2507.
29. Zemann,A., op de Bekke,A., Kiefmann,M., Brosius,J. and Schmitz,J. (2006) Evolution of small nucleolar RNAs in nematodes. *Nucleic Acids Res.*, **34**, 2676–2685.
30. 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.
31. Ruby,J.G., Jan,C., Player,C., Axtell,M.J., Lee,W., Nusbaum,C., Ge,H. and Bartel,D.P. (2006) Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell*, **127**, 1193–1207.
32. Marker,C., Zemann,A., Terhorst,T., Kiefmann,M., Kastenmayer,J.P., Green,P., Bachellerie,J.P., Brosius,J. and Huttenhofer,A. (2002) Experimental RNomics: identification of 140 candidates for small non-messenger RNAs in the plant Arabidopsis thaliana. *Curr. Biol.*, **12**, 2002–2013.
33. Aspegren,A., Hinas,A., Larsson,P., Larsson,A. and Soderbom,F. (2004) Novel non-coding RNAs in Dictyostelium discoideum and their expression during development. *Nucleic Acids Res.*, **32**, 4646–4656.
34. Huttenhofer,A. and Vogel,J. (2006) Experimental approaches to identify non-coding RNAs. *Nucleic Acids Res.*, **34**, 635–646.
35. Vanacova,S., Liston,D.R., Tachezy,J. and Johnson,P.J. (2003) Molecular biology of the amitochondriate parasites, Giardia intestinalis, Entamoeba histolytica and Trichomonas vaginalis. *Int. J. Parasitol.*, **33**, 235–255.
36. Keeling,P.J., Burger,G., Durnford,D.G., Lang,B.F., Lee,R.W., Pearlman,R.E., Roger,A.J. and Gray,M.W. (2005) The tree of eukaryotes. *Trends Ecol. Evol.*, **20**, 670–676.
37. Nixon,J.E., Wang,A., Field,J., Morrison,H.G., McArthur,A.G., Sogin,M.L., Loftus,B.J. and Samuelson,J. (2002) Evidence for lateral transfer of genes encoding ferredoxins, nitroreductases, NADH oxidase, and alcohol dehydrogenase 3 from anaerobic prokaryotes to Giardia lamblia and Entamoeba histolytica. *Eukaryot. Cell*, **1**, 181–190.
38. Nixon,J.E., Wang,A., Morrison,H.G., McArthur,A.G., Sogin,M.L., Loftus,B.J. and Samuelson,J. (2002) A spliceosomal intron in Giardia lamblia. *Proc. Natl Acad. Sci. USA*, **99**, 3701–3705.
39. Russell,A.G., Shutt,T.E., Watkins,R.F. and Gray,M.W. (2005) An ancient spliceosomal intron in the ribosomal protein L7a gene (Rpl7a) of Giardia lamblia. *BMC Evol. Biol.*, **5**, 45.
40. Niu,X.H., Hartshorne,T., He,X.Y. and Agabian,N. (1994) Characterization of putative small nuclear RNAs from Giardia lamblia. *Mol. Biochem. Parasitol.*, **66**, 49–57.
41. Yang,C.Y., Zhou,H., Luo,J. and Qu,L.H. (2005) Identification of 20 snoRNA-like RNAs from the primitive eukaryote, Giardia lamblia. *Biochem. Biophys. Res. Commun.*, **328**, 1224–1231.
42. Luo,J., Zhou,H., Chen,C.H., Li,Y., Chen,Y. and Qu,L.H. (2006) Identification and evolutionary implication of four novel box H/ACA snoRNAs from Giardia lamblia. *Chin. Sci. Bull.*, **51**, 2451–2456.
43. Kurland,C.G., Collins,L.J. and Penny,D. (2006) Genomics and the irreducible nature of eukaryote cells. *Science*, **312**, 1011–1014.
44. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.
45. Samarsky,D.A., Fournier,M.J., Singer,R.H. and Bertrand,E. (1998) The snoRNA box C/D motif directs nucleolar targeting and also couples snoRNA synthesis and localization. *EMBO J.*, **17**, 3747–3757.
46. Cavaille,J., Nicoloso,M. and Bachellerie,J.P. (1996) Targeted ribose methylation of RNA in vivo directed by tailored antisense RNA guides. *Nature*, **383**, 732–735.
47. Kiss-Laszlo,Z., Henry,Y. and Kiss,T. (1998) Sequence and structural elements of methylation guide snoRNAs essential for site-specific ribose methylation of pre-rRNA. *EMBO J.*, **17**, 797–807.
48. Edlind,T.D. and Chakraborty,P.R. (1987) Unusual ribosomal RNA of the intestinal parasite Giardia lamblia. *Nucleic Acids Res.*, **15**, 7889–7901.
49. Lowe,T.M. and Eddy,S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, **283**, 1168–1171.
50. 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.
51. Ganot,P., Caizergues-Ferrer,M. and Kiss,T. (1997) The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. *Genes Dev.*, **11**, 941–956.
52. Liang,X.H., Liu,L. and Michaeli,S. (2001) Identification of the first trypanosome H/ACA RNA that guides pseudouridine formation on rRNA. *J. Biol. Chem.*, **276**, 40313–40318.
53. Tang,T.H., Polacek,N., Zywicki,M., Huber,H., Brugger,K., Garrett,R., Bachellerie,J.P. and Huttenhofer,A. (2005) Identification of novel non-coding RNAs as potential antisense regulators in the archaeon Sulfolobus solfataricus. *Mol. Microbiol.*, **55**, 469–481.
54. Rozhdestvensky,T.S., Tang,T.H., Tchirkova,I.V., Brosius,J., Bachellerie,J.P. and Huttenhofer,A. (2003) Binding of L7Ae protein to the K-turn of archaeal snoRNAs: a shared RNA binding motif for C/D and H/ACA box snoRNAs in Archaea. *Nucleic Acids Res.*, **31**, 869–877.
55. Ullu,E., Tschudi,C. and Chakraborty,T. (2004) RNA interference in protozoan parasites. *Cell Microbiol.*, **6**, 509–519.
56. Burke,W.D., Malik,H.S., Rich,S.M. and Eickbush,T.H. (2002) Ancient lineages of non-LTR retrotransposons in the primitive eukaryote, Giardia lamblia. *Mol. Biol. Evol.*, **19**, 619–630.
57. Ullu,E., Lujan,H.D. and Tschudi,C. (2005) Small sense and antisense RNAs derived from a telomeric retroposon family in Giardia intestinalis. *Eukaryot. Cell*, **4**, 1155–1157.
58. Mattaj,I.W. and Zeller,R. (1983) Xenopus laevis U2 snRNA genes: tandemly repeated transcription units sharing 5′ and 3′ flanking homology with other RNA polymerase II transcribed genes. *EMBO J.*, **2**, 1883–1891.
59. Collins,L.J., Macke,T.J. and Penny,D. (2004) Searching for ncRNAs in eukaryotic genomes: maximizing biological input with RNAmotif. *J. Integr. Bioinformatics*, **1**, 61–77.
60. Riedel,N., Wise,J.A., Swerdlow,H., Mak,A. and Guthrie,C. (1986) Small nuclear RNAs from Saccharomyces cerevisiae: unexpected diversity in abundance, size, and molecular complexity. *Proc. Natl Acad. Sci. USA*, **83**, 8097–8101.
61. Yean,S.L., Wuenschell,G., Termini,J. and Lin,R.J. (2000) Metal-ion coordination by U6 small nuclear RNA contributes to catalysis in the spliceosome. *Nature*, **408**, 881–884.
62. Tani,T. and Ohshima,Y. (1991) mRNA-type introns in U6 small nuclear RNA genes: implications for the catalysis in pre-mRNA splicing. *Genes Dev.*, **5**, 1022–1031.