BIOINFORMATICS DISCOVERY NOTE Vol. 20 no. 18 2004, pages 3293–3301
doi:10.1093/bioinformatics/bth394

A computational search for box C/D snoRNA genes in the Drosophila melanogaster genome

M. C. Accardo[1,†], E. Giordano[1,†], S. Riccardo[1], F. A. Digilio[2], G. Iazzetti[1], R. A. Calogero[3,*] and M. Furia[1]

[1]Department of Genetics, General and Molecular Biology, University of Naples 'Federico II', via Mezzocannone 8, 80134 Napoli, Italy, [2]Institute of Genetics and Biophisics Adriano Buzzato Traverso, Consiglio Nazionale delle Ricerche, 80131 Naples, Italy and [3]Department of Clinical and Biological Sciences, University of Torino, Regione Gonzole 10, 10043 Orbassano (TO), Italy

Received on December 23, 2003; revised on May 9, 2003; accepted on June 26, 2004
Advance Access publication July 9, 2004

ABSTRACT

Motivation: In eukaryotes, the family of non-coding RNA genes includes a number of genes encoding small nucleolar RNAs (mainly C/D and H/ACA snoRNAs), which act as guides in the maturation or post-transcriptional modifications of target RNA molecules. Since in *Drosophila melanogaster* (Dm) only few examples of snoRNAs have been identified so far by cDNA libraries screening, integration of the molecular data with *in silico* identification of these types of genes could throw light on their organization in the Dm genome.

Results: We have performed a computational screening of the Dm genome for C/D snoRNA genes, followed by experimental validation of the putative candidates. Few of the 26 confirmed snoRNAs had been recognized by cDNA library analysis. Organization of the Dm genome was also found to be more variegated than previously suspected, with snoRNA genes nested in both the introns and exons of protein-coding genes. This finding suggests that the presence of additional mechanisms of snoRNA biogenesis based on the alternative production of overlapping mRNA/snoRNA molecules.

Availability: Additional information is available at http://www.bioinformatica.unito.it/bioinformatics/snoRNAs

Contact: raffaele.calogero@unito.it

INTRODUCTION

Genes producing functional RNAs rather than protein products form a large and variegated class in all genomes, from bacteria to mammals. However, despite the importance of their functional roles, most of them have not yet been identified even in organisms whose genome has been completely sequenced. In eukaryotes, the family of nc-RNA genes comprises many small nucleolar RNAs (snoRNAs) that guide the maturation or post-transcriptional modification of target RNA molecules. Most snoRNAs fall into two classes called box C/D and box H/ACA snoRNAs. Each class is defined by the presence of common sequence motifs and common associated proteins (reviewed by Bachellerie *et al*., 2002). A few snoRNAs in either class are required for definite pre-rRNA cleavages and essential for viability, whereas most are responsible for the $2' - O$-ribose methylation (C/D) or pseudouridylation (H/ACA) of target RNA molecules, respectively (reviewed by Kiss, 2002; Bachellerie *et al*., 2002). The H/ACA class directs pseudouridylation at specific selected sites. The C/D class guides site-specific $2' - O$-ribose methylation by base pairing of the 10–21 nt long sequence positioned upstream from a D (or an internal D′) box to the target RNA, with the nucleotide positioned 5 bp upstream from the D/D′ box selected for methylation (Nicoloso *et al*., 1996). Most of the C/D and H/ACA snoRNAs are involved in the modification of rRNA, though tRNA, snRNAs and possibly mRNAs are recognized as targets. Their range of action may thus extend beyond ribosome biogenesis. Mammalian guide RNAs that lack complementarities for rRNA or spliceosomal snRNAs have, in fact, been identified, and it has been suggested that they may target mRNAs (Cavaillè *et al*., 2000). Intriguingly, several of these specimens devoid of significant base pairing with rRNA or snRNA do not show an ubiquitous expression, but are tissue-specific in both mouse and man (reviewed by Bachellerie *et al*., 2002).

Many snoRNAs guiding the maturation or post-transcriptional modification of target RNA molecules have been described in several eukaryotic organisms. Their identification has been primarily achieved by computer-assisted genome analysis or the production of specialized cDNA libraries. In *Drosophila melanogaster* (Dm), few examples of the two main classes had been described until very recently.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

Among 66 candidates for small non-messenger RNAs (snmRNAs) identified from a Dm cDNA library generated from total RNA, sized from 50 to 500 nt, only five new RNA species correspond to C/D snoRNAs predicted to methylate specific rRNA sites (Yuan *et al*., 2003). rRNA methylated sites have not yet been mapped in Dm. Their number, however, is thought to be much greater, since computational searches have revealed 41 snoRNA genes responsible for 51 sites among the 55 rRNA methylated modifications in yeast (Lowe and Eddy, 1999), 66 C/D snoRNAs in *Arabidopsis* (Barneche *et al*., 2001) and about 107 $2' - O$-methylated residues in humans (Maden, 1990). The number of methylated residues on Dm rRNA may be reasonably expected to lie within the range for yeast and mammallian rRNAs. It may thus be supposed that a substantial number of fly box C/D snoRNA genes targeting rRNA escaped the recent molecular analysis and remain hidden in the Dm genome. We have therefore applied a computational approach to the complete Dm genome sequence (Adams *et al*., 2000) in a large-scale identification of C/D snoRNA genes potentially involved in rRNA methylation.

As reported here, this analysis predicts that 99 snoRNA genes may be responsible for methylation of Dm 18S/28S rRNA. We were able to confirm experimentally 26 snoRNA specimens (4 on 18S rRNA and 22 on 28S rRNA) out of 44 snoRNAs identified in *Drosophila* genome, using SNOSCAN program and methylation sites conserved between *Drosophila* and yeast. Inspection of the genomic sequences flanking these genes revealed that most are present in multiple copies and arranged in clusters, a feature that proved useful in identifying additional ncRNA genes not specifically targeted in the initial screening. The very limited overlap between our results obtained by computational search and those recently reported for a cDNA library screening for nc-RNAs suggests that these two techniques are complementary.

MATERIALS AND METHODS

Database search and sequence analysis

All data parsing was performed with PERL scripts (www.perl.org). The Dm genome scaffolds available at ftp://ftp.ncbi.nih.gov/genomes/Drosophila_melanogaster were searched for potential box C/D snoRNAs targeting putative rRNA methylated sites (potentially conserved between Dm and yeast), using a snoRNA search program SNOSCAN (Lowe and Eddy, 1999, http://rna.wustl.edu/snoRNAdb/code/). SNOSCAN employs a greedy search algorithm to scan for $2'$-O-methylation guide snoRNA candidates. It sequentially identified six components characteristic of these genes: box D, box C, a region of sequence complementary to ribosomal RNA, box D' if the rRNA complementary region is not directly adjacent to box D, the predicted methylation site within the rRNA based on the complementary region and the terminal stem base pairings, if present. The program also takes into account the relative distance between identified features within the snoRNA, information which is critical to reducing the number of false positives. To identify snoRNA genes, SNOSCAN needs the rRNA (28S, 18S) sequences and a list of rRNA methylation sites. Since the Dm rRNA methylated sites have not been experimentally determined, an alignment between *Saccharomyces cerevisiae* (Sc) rRNAs was generated by the BLAST 2 sequences program (Altschul *et al*., 1990) and the Sc methylated sites annotated at http://rna.wustl.edu/snoRNAdb/Sc/Sc-snos-bysno.html were mapped on Dm rRNAs. The putative Dm methylation sites were considered reliable if they allowed identification of the corresponding Sc snoRNA genes with a score higher than 20 bits, which is the default parameter defined by Lowe and Eddy (1999). Furthermore, we refined the set of putative methylation sites on *Drosophila* rRNAs by integrating human, yeast and *Drosophila* structural alignments (http://www.rna.icmb.utexas.edu/) with the snoRNA human/yeast conservation annotations (http://bioinf.scri.sari.ac.uk/cgi-bin/plant_snorna/conservation). The Dm genome scanning yielded 145 candidates and some of the putative snoRNA genes were generated by a few bases shift over the same genomic region encoding the snoRNA; in these cases, the prediction with the highest score was kept to yield the 99 snoRNAs presented in this paper. All these 99 Dm snoRNA genes had a SNOSCAN score higher than 20 bits (Lowe and Eddy, 1999). Flanking sequences of each snoRNA candidate were also examined for other C/D snoRNAs by BLAST analysis and visual inspection. Futhermore, BLAST scanning was applied to the complete Dm genome to identify variants of all snoRNA genes. Sequence alignment of snoRNA isoforms was performed with CLUSTAL W program (Thompson *et al*., 1994). The D and D' upstream regions ($\geq$13 nt) were also used to scan the Dm genome in order to identify potential targets other than rRNAs.

snoRNA experimental validation

Experimental confirmation of snoRNA putative candidates identified in *Drosophila* genome using SNOSCAN program and the methylation sites conserved between *Drosophila* and yeast was performed by developmental northern blot analysis. Total RNA was extracted from various stages of *Drosophila* development (embryos, first instar larvae, third instar larvae, and male and female adults) by the TRIzol method (Gibco-BRL). For northern blot analysis, 6 μg of total RNA were electrophoresed and transferred onto Hybond-NX (Amersham) filters for hybridization. Specific probes were PCR-amplified on genomic DNA by using the appropriate primer pairs. Basic cloning techniques, PCR amplification, DNA extraction, manipulation and labelling, screening and sequencing techniques were carried out according to the method of Sambrook *et al*. (1989). Size of snoRNAs was determined on agarose or 6% polyacrylamide gels by using the DNA Molecular Weight Marker V (Roche) end-labelled

Table 1. List of experimentally confirmed snoRNAs found associated to 18S *D.melanogaster* rRNA by SNOSCAN analysis (Lowe and Eddy, 1999)

18S putative methylation site in Dm (homologous *S.cerevisiae* methylation site)[a] **Sc guide snoRNA**	SNOSCAN symbol **Assigned name**	**SNOSCAN prediction** (score)	**snoRNAs found in the same cluster**	**Genomic location** Chromosome, contig (start..end)	**snoRNA location within gene sequence(s)**	**snoRNA estimated length**
A 28 (A 28) ***SnR74/Z4***	Dm_18S.009[b,c] (SnoRNA:U27:54Eb) ***DmSnR74/Z4***	31.50	Dm_28S.031[b] Dm_28S.032[b] Dm_28S.041[b]	2R, NT_033778 (12766358..12766481)	dUhg1: CG14486	71
A 425 (−)	Dm_18S.011 **Dm425**	26.92	—	3R, NT_033777 (7048578..7048666)	Intron 3 CG4863-RA; intron 2 CG4863-RE; exon 3 CG4863-RD (ribosomal protein L3)	80
A 1061 (A 973) **SnR54**	Dm_18S.001 **DmSnR54**	30.15	—	2R, NT_033778 (9686765..9686892)	Intergenic CG12863/CG10131	72
G 1620 (G 1425) **SnR56**	Dm_18S.007[b,c] (SnoRNA:U25:30E) **DmSnR56 a**	30.05	—	2L, NT_033779 (9886700..9886777)	dUhg2: CR32873	67

[a]The absence of the homologous *S.cerevisiae* methylation site indicates that putative methylation site was predicted by SNOSCAN.
[b]snoRNA also identified by Tycowski and Steitz (2001).
[c]snoRNA having an antisense sequence (≥13 nt), located upstream to D or D′ box, that is not complementary to rRNA sequences and for which were found putative target sequences in Dm gene transcripts.

with [γ-^{32}P]ATP and T4 polynucleotide kinase. The 5′ ends of snoRNAs were determined by primer extension analysis using 50 μg of total RNA and suitable primers complementary to the snoRNA internal sequences. RNA quantitative analyses were carried out with the ImageQuaNT software and the Molecular Dynamics PhosphorImager.

RESULTS AND DISCUSSION

Since rRNA methylation in *Drosophila* has not yet been determined experimentally, 18S and 28S rRNA sequences from Dm were aligned with those of Sc rRNAs. We defined putative *Drosophila* rRNA methylation sites as those experimentally defined in the yeast (Kiss-Laszlo *et al*., 1996; Lowe and Eddy, 1999) and present in regions conserved between Dm and Sc rRNAs (see Materials and Methods section). This rationale was supported by the notion that, judging from sequence complementarity to rRNA and the corresponding 2′-O-methylation sites, a recent analysis has indicated that 58 of the C/D snoRNA genes from rice have homologues in other organisms, including 15 snoRNA genes that are well conserved in plants, yeasts and humans (Chen *et al*., 2003). The SNOSCAN program (Lowe and Eddy, 1999) was used to identify snoRNA genes in the Dm genome. Out of the 16 Sc methylation sites available (http://rna.wustl.edu/snoRNAdb/Sc/Sc-snos-bysite.html) on the small ribosomal subunit (SSU) we defined 'reliable' (see Materials and Methods section) 10 sites on Dm 18S rRNA and on Dm 28S rRNA, 34 sites out of the 39 available on the Sc large ribosomal subunit (LSU).

The putative snoRNAs identified by scanning the *Drosophila* genome were integrated in Table V (see additional information website) with the annotations of the snoRNAs and rRNA methylation sites described in yeast and humans (http://bioinf.scri.sari.ac.uk/cgi-bin/plant_snorna/conservation). We also used the structural alignment of human versus *Drosophila* rRNAs (http://www.rna.icmb.utexas.edu/) and the annotated human methylation sites to identify additional putative snoRNA targets in *Drosophila* rRNA sequences. Among the putative methylation sites conserved between yeast and humans, 10 on 18S and 23 on 28S, 8 (80%) and 22 (95%), respectively, were conserved in *Drosophila*. We also found four sites on 18S and six sites on 28S that represented putative methylation sites common to humans and *Drosophila* only (Table V). When used to analyse the *Drosophila* genome by SNOSCAN, only two of these did not allow the identification of snoRNAs.

In this paper, we report a validation analysis of the 44 snoRNA candidates identified in the *Drosophila* genome using SNOSCAN program and the methylation sites conserved between *Drosophila* and yeast. This analysis led to confirm the expression of 26 snoRNA genes (4 on 18S rRNA and 22 on 28S rRNA) (Tables 1 and 2; Fig. 1 for a schematic representation of a typical C/D snoRNA structure and the role of its

Table 2. List of experimentally confirmed snoRNAs found associated to 28S *D.melanogaster* rRNA by SNOSCAN analysis (Lowe and Eddy, 1999)

28S putative methylation site in Dm (homologous *S.cerevisiae* methylation site)[a] **Sc guide snoRNA**	SNOSCAN symbol **Assigned name**	**SNOSCAN prediction** (score)	**snoRNAs found in the same cluster**	**Genomic location** Chromosome, contig (start..end)	**snoRNA location within gene sequence(s)**	**snoRNA estimated length**
A 773 (A 647) **U18**	Dm_28S.002[b] ***DmU18 a***	31.46	—	3L, NT_037436 (12981769..12981850)	Intron 3 CG11271-RA, RB, RF; intron 2 CG11271-RC (putative ribosomal protein S12)	82
	Dm_28S.002a[b] **DmU18 b**	—	—	3L, NT_037436 (12981134..12981195)	Intron 2 CG11271-RA, RB, RC, RF; (putative ribosomal protein S12)	84
C 774 (C 648) **U18adj**	Dm_28S.004[c] **DmU18adj**	26.05	—	3L,NT_037436 (21544792..21544926)	Exon 4 CG11306 (putative alpha-1,2-mannosyltransferase activity)	135
C 787 (C 661) **SnR58**	Dm_28S.005[b] **DmSnR58 a**	25.20	Dm_28S.043[b] Dm_28S.043a[b]	3L, NT_037436 (807898..808016)	Intron 4 CG13900-RB; intron 10 CG13900-RA (putative damaged DNA-binding activity)	119
	Dm_28S.005a[b] **DmSnR58 b**	—	—	3L, NT_037436 (807903..808008) 2R, NT_033778	Intron 4 CG13900-RB; intron 10 CG13900-RA (putative damaged DNA-binding activity)	107
A 981 (A 805) **SnR39/59**	Dm_28S.008[b,c] **DmSnR39/59 a**	33.99	Dm_28S.027 Dm_28S.028	2R, NT_033778 (4157248..4157335)	Intergenic CG13741/CG8078	88
	Dm_28S.008 a[b,c,d] **DmSnR39/59 b**	—		2R, NT_033778 (4156990..4157057)	Intergenic CG13741/CG8078	79
G 1082 (G 906) **SnR60**	Dm_28S.010[b,c] **DmSnR60 a**	38.37	Dm_28S.011[b] Dm_28S.010/011a[b] Dm_28S.010/011b[b]	2R, NT_033778 (19228887..19228978)	Intron 8 mfl/Nop60B:CG3333	92
	Dm_28S.011[b,c] **DmSnR60 b**	36.40	Dm_28S.010[b] Dm_28S.010/011a[b] Dm_28S.010/011b[b]	2R, NT_033778 (199229126..19229217)	Intron 9 mfl/Nop60B:CG3333	92
	Dm_28S.010[c]/011a[b] **DmSnR60 c**	—	Dm_28S.010[b] Dm_28S.010/011b[b]	2R, NT_033778 Dm_28S.011 (19228662..19228736)	Intron 7 mfl/Nop60B:CG3333	92
	Dm_28S.010[c]/011b[b] **DmSnR60 d**	—	Dm_28S.010[b] Dm_28S.011[b] Dm_28S.010/011b[b]	2R, NT_033778 (19228415..19228490)	Intron 6 mfl/Nop60B:CG3333	92
A 1321 (A 797) ***SnR61***	Dm_28S.016[e] (Dm-797) **DmSnR61 c**	29.07	—	2R, NT_033778 (6957832..6957916)	Intron 1 Ef1alpha48D	80
G 1322 (−)	Dm_28S.040[c] ***DmG1322***	22.81	—	3R, NT_033777 (12139343..12139450)	Opposite polarity: exon 3 gene sulf1 CG6725 (*N*-acetylglucosamine-6-sulfatase-activity; required for normal pattern formation of embryonic cuticle)	108
U 1332 (−)	Dm_28S.039 ***DmU1332***	21.69	—	X, NC_004354 (17818091..17818221)	Exon 1 gene dik CG7098 (component of histone acetyltransferase complex)	131

Table 2. Continued

28S putative methylation site in Dm (homologous *S.cerevisiae* methylation site)[a] **Sc guide snoRNA**	SNOSCAN symbol **Assigned name**	**SNOSCAN prediction** (score)	**snoRNAs found in the same cluster**	**Genomic location** Chromosome, contig (start..end)	**snoRNA location within gene sequence(s)**	**snoRNA estimated length**
C 1652 (C1435) **U24**	Dm_28S.017 **DmU24**	23.15	—	3L, NT_037436 (3201959..3202060)	Intron 3 CG12740-RA, RB, RD; exon 3 CG12740-RC (putative ribosomal L28e protein)	102
U 2133 (U 1886) ***SnR62***	Dm_28S.018[b,c] **DmSnR62 a**	34.24	—	3R, NT_033777 (1451090..1451158)	Exon 3 CG 1475, RH27364; intron 2 CG 1475, for SD27659 (putative ribosomal L13 protein)	69
	Dm_28S.018a[b] **DmSnR62 b**	—	—	3R, NT_033777 (1450906..1450971)	Intron 2 CG 1475 (putative ribosomal L13 protein)	65
A 2527 (A 2218) ***SnR47***	Dm_28S.019 **DmSnR47 a**	24.53	—	X, NC_004354 (19775793..19775877)	Intergenic CG9570/CG9571	85
C 2644 (C 2335) **SnR64**	Dm_28S.023[b] **DmSnR64 a**	32.17	—	2L, NT_033779 (20617555..20617735)	GH14469, putative RNA non-coding gene	181
	Dm_28S.023a[b,d] **DmSnR64 b**	—	—	2L, NT_033779 (20617800..20617970)	GH14469, putative RNA non-coding gene	171
	Dm_28S.023b[b,d] **DmSnR64 c**	—	—	2L, NT_033779 (20618031..20618173)	GH14469, putative RNA non-coding gene	143
G 3080 (G 2616) **SnR67**	Dm_28S.041[c,f] (snoRNA:U31:54Ea) **DmSnR67**	29.38	Dm_28S.031[f] Dm_18S.009[f] Dm_18S.032[f]	2R, NT_033778 (12762927..12763009)	dUHG1: CG14486	83
G 3112 (−)	Dm_28S.043[b] ***Dm3112***	23.09	Dm_28S.005[b] Dm_28S.005a[b]	3L, NT_037436 (808163..808253)	Intron 2 CG13900-RB; intron 8 CG13900-RA (putative damaged DNA-binding activity)	91
	Dm_28S.043a[b,c] ***Dm3112 a***	—	—	3L, NT_037436 (808410..808451)	Intron 1 CG13900-RB; intron 7 CG13900-RA (putative damaged DNA-binding activity)	91
G 3254 (G 2790) ***SnR48***	Dm_28S.026[c] **DmSnR48 a**	23.19	—	2R, NT_033778 (19961174..19961258)	Intron 3 CG18506 (function unknown)	85
	Dm_28S.027 **DmSnR48 b**	27.93	Dm_28S.008[b] Dm_28S.008a[b] Dm_28S.028	2R, NT_033778 (4157505..4157597)	Intergenic CG13741/CG8078	93
	Dm_28S.028[c] **DmSnR48 c**	26.96	Dm_28S.008[b] Dm_28S.008a[b] Dm_28S.027	2R, NT_033778 (4156738..4156833)		96
C 3403 (−)	Dm_28S.046 a[b] **Dm3403 a**	20.62	—	4, NC_004353 (779886..780018)	Exon 33 gene bt CG32019	133
	Dm_28S.046 b[b] **Dm3403 b**	—	—	4, NC_004353 (772774..772883)	Exon 32 gene bt CG32019	109
	Dm_28S.046 c[b] **Dm3403 c**	—	—	4, NC_004353 (771067..771181)	Exon 30 gene bt CG32019	114
	Dm_28S.046 d[b] **Dm3403 d**	—	—	4, NC_004353 (767748..767852)	Exon 29 gene bt CG32019	104

Table 2. Continued

28S putative methylation site in Dm (homologous *S.cerevisiae* methylation site)[a] **Sc guide snoRNA**	SNOSCAN symbol **Assigned name**	**SNOSCAN prediction** (score)	**snoRNAs found in the same cluster**	**Genomic location** Chromosome, contig (start..end)	**snoRNA location within gene sequence(s)**	**snoRNA estimated length**
A 3406 (A 2943) ***SnR71***	Dm_28S.031[c,f] (snoRNA:U29:54Eb) **DmSnR71 a**	29.25	Dm_18S.009[f] Dm_28S.032[f] Dm_28S.041[f]	2R, NT_033778 (12763939..12764038)	dUHG1: CG14486	100
	Dm_28S.032[c,f] (snoRNA:U29:54Ed) **DmSnR71 b**	28.84	Dm_18S.009[f] Dm_28S.031[f] Dm_28S.041[f]	2R, NT_033778 (12764695..12764795)	dUHG1: CG14486	101
C 3408 (C 2945) ***SnR69***	Dm_28S.036 a[b] **DmSnR69 a**	27.55	—	3L, NT_037436 (17767021..17767175)	Exon 1 CG14586 (glutamated-gated-ion channel activity)	155
	Dm_28S.036 b[b] **DmSnR69 b**	—	—	3L, NT_037436 (17764880..17765031)	Exon 3/intron 3 CG14586 (glutamated-gated-ion channel activity)	151

[a]The absence of the homologous *S.cerevisiae* methylation site indicates that in *D.melanogaster* the putative methylation site was predicted by SNOSCAN.
[b]All snoRNAs indicated with the same number (i.e. Dm_28S.018, Dm_28S.018a) are located in the same cluster; when SNOSCAN score is not present snoRNA isoforms have been identified by BLAST scanning of the sequences surrounding the gene originally identified by SNOSCAN and have highly related sequences.
[c]snoRNA having an antisense sequence (≥13 nt) located upstream to D or D′ box that is not complementary to rRNAs sequences, and for which were found target sequences in the Dm gene transcripts.
[d]The antisense sequence located upstream to the D and D′ boxes is not complementary to rRNAs.
[e]snoRNA also identified by Yuan *et al.* (2003).
[f]snoRNA also identified by Tycowski and Steitz (2001).

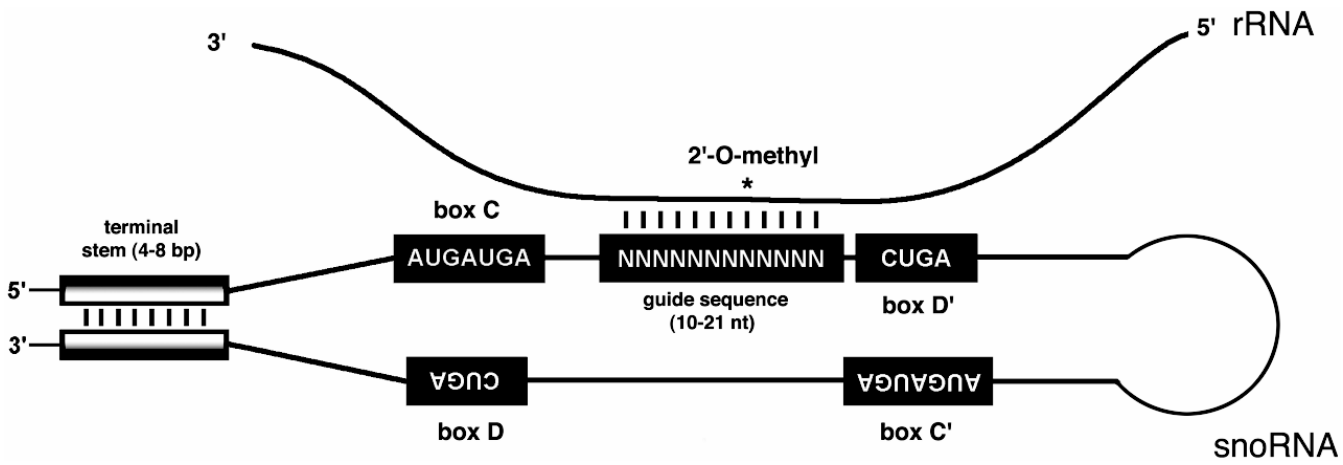

Fig. 1. Schematic representation of a typical C/D box snoRNA structure. The rRNA 2′-O-methylation site is indicated (*) within the helix formed by the snoRNA antisense guide sequence and the target rRNA.

guide antisense sequence). Furthermore, 3/4 (75%) and 11/22 (50%) of the experimentally confirmed *Drosophila* snoRNAs had their methylation target conserved in man and yeast 18S and 28S rRNAs, respectively.

In our validation assay, a panel of total RNA samples extracted from various stages of Dm development was analysed by northern blot analysis using each specific probe (Additional information, Fig. 2). Most of the genes examined were constitutively expressed during Dm development, as expected for snoRNA molecules targeting rRNA. The length of each molecule was more accurately established by electrophoresis on 6% polyacrylamide gels and, in some cases,

by determination of the 5′ end by primer extension analysis (Additional information, Fig. 2). The full set of validated and not validated snoRNA genes is available, as additional information, in Table III, while snoRNAs sequences are available, as additional information, in Table IV. Computationally predicted snoRNAs not detected by northern blot analysis may represent false positives, or snoRNAs that escaped our molecular analysis due to their low abundance or very sharp expression profile. We cannot at present say whether assays more sensitive than northern blots would validate more of our predicted candidates.

Six validated snoRNAs correspond to specimens already identified (Tycowski and Steitz, 2001; Yuan *et al*., 2003). Only one, however, (called DmSnR61c in our analysis, previously indicated as Dm-797 and already deposited in GenBank as Z1) was represented in a *Drosophila*-sized cDNA library recently screened by Yuan *et al*. (2003), whereas the other five were members of the dUHG1 or the dUHG2 cluster identified by sequence homology to the mammalian UHG polycistronic ncRNA (Tycowski and Steitz, 2001; Tycowski *et al*., 1996). Remarkably, the finding that four dUHG1 members (U27:54Eb; U29:54Eb; U29:54Ed; U31:54Ea), together with one dUHG2 member (U25:30E), are potentially able to methylate sites conserved between yeast and *Drosophila* 18S/28S rRNAs provides a clear hint of their evolutionary origin, as well as that of their highly related mammalian counterparts (Tycowski *et al*., 1996).

We found that 18/44 genes were arranged in clusters. In addition to the previously identified dUHG1 and dUHG2 polycistronic ncRNAs (Tycowski and Steitz, 2001), we identified eight new clusters, either simply comprised of tandem repeats of highly related copies (the DmU18, DmSnR60, DmSnR62, DmSnR64, Dm3403 and DmSnR69 clusters; Table 2) or a mixture of homologous and heterologous snoRNAs (the DmSnR58/Dm3112 and DmSnR48/DmSnR39/59 clusters; Table 2). This finding provides further evidence that polycistronic organization is common in both invertebrates and vertebrates (Tycowski and Steitz, 2001), and implies that snoRNA gene duplication has frequently occurred during the evolution of the Dm genome. All the snoRNA genes in each cluster were usually arranged in a head-to-tail fashion and closely linked. Tandem gene duplication events generate functional redundancy and can establish sequence variability allowing the generation of new snoRNAs for selection. Consistent with this assumption, the identified clusters present both copies in which the antisense motifs were perfectly conserved among tandemly repeated snoRNA coding units (occasionally with polymorphism substantially restricted to sequences immediately downstream from the C or D′ boxes), and divergent copies displaying significant nucleotide changes within the antisense motifs and hence unable to target Dm rRNA (e.g. see the DmSnR39/59 b and DmSnR 64 b–c specimens, marked by • in Table 2). Polycistrons in introns show that the one-snoRNA-per-intron organization previously observed in vertebrates (reviewed by Bachellerie *et al*., 2002) is also present in invertebrates. Cluster arrangement, indeed, proved useful in identifying ncRNA genes not specifically targeted by our screening. Inspection of flanking sequences, in fact, occasionally revealed the presence of genes encoding either snmRNAs or snoRNAs of the H/ACA class. This occurred in the case of the DmSnR60 a–d and DmSnR64 a–c clusters, as well as the DmU24 snoRNA gene (manuscript in preparation). In six cases, the snoRNA genes were located in apparently intergenic regions of the Dm genome and may represent new examples of genes devoid of protein-coding potential, thus broadening the repertoire of Dm genes that produce solely ncRNAs. The DmSnR54 and the DmSnR48/DmSnR39/59 clusters, in fact, are probably comprised of new polycistronic Dm ncRNA genes. In vertebrates, snoRNAs are frequently found in introns of genes responsible for ribosome biogenesis, or more generally involved in translation, and it has been suggested that this organization has evolved for coordinate expression of functionally related genes. We found seven cases in which the snoRNA gene was located within an intron of a protein-coding gene involved, or putatively involved, in ribosome synthesis or more generally in translation. These cases also include examples of intronic clusters, in which we found that several introns of the same gene hosted a snoRNA according to the 'one-snoRNA-per-intron' organization (Table 1, CG4863; Table 2, CG11271, Nop60B, Ef1alpha48D, CG12740, CG 1475).

However, in three cases the snoRNA gene was located within an intron of a protein-coding gene whose function is unknown or thought to be completely unrelated to translation (Table 2: CG11271, CG13900, CG18506). This was quite unexpected, especially since our approach was specifically focused on the identification of snoRNAs potentially able to guide methylation of rRNA, and raises the possibility that these snoRNAs may target other types of RNA molecules. The vast majority of the snoRNAs identified in our analysis, indeed, are characterized by the presence of the D and D′ boxes and in all but one of those validated the rRNA recognition sequence is bound to the region upstream from the internal additional D′ box (Additional information, Table III). By scanning the Dm genome for the presence of target sequences complementary to the antisense sequence ($\geq$13 nt) located upstream from the D/D′ boxes not involved in rRNA recognition, we were able to identify for 25 snoRNAs (marked by $\otimes$ in Tables 1 and 2) the presence of potential recognition sites within Dm gene transcripts. These snoRNAs may thus be involved in both rRNA and mRNA modification. In five cases (DmU18adj, DmG1322, DmU1322, Dm3403 and DmSnR69), the snoRNA was encoded within an exon of a protein-coding gene, and in one case (DmG1322) even with polarity opposite to that of the host gene. This type of arrangement has not been previously described for any snoRNA coding unit, and may reflect a still uncharacterized mechanism of snoRNA biogenesis based on the alternative production of

the overlapping mRNA/snoRNA molecules. Indeed, in three additional examples we noticed that the snoRNA gene was located within a genomic sequence subjected to alternative splicing. Location of the snoRNA gene may thus be either intronic or exonic, according to the splicing pattern (Table 1, Dm425; Table 2, DmU24 and DmSnR62). Finally, we also found an example of snoRNA coding unit located across an exon/intron boundary (the DmSnR69 b gene, mapping across exon 3/intron 3 boundary in CG14586). Taken as a whole, these data suggest that snoRNA organization in the Dm genome is more variegated than had been previously supposed. Location within open reading frames (ORFs) (or putative ORFs) has been reported for eight Dm snmRNAs, while opposite orientation relative to the host protein-coding gene has been noted for two intron-encoded snmRNAs (Yuan *et al*., 2003). Further experiments are required to clarify the mechanism accounting for the expression of Dm ORF-encoded snoRNAs and/or snmRNAs.

Comparison of our data with those provided by cloning (Yuan *et al*., 2003) shows that, with the exception of the DmSnR61c/Dm797 specimen detected in both analyses, and the Dm-442 and Am2564 on 28S, which our analysis failed to detect, all the C/D snoRNAs targeting rRNA identified by cloning were predicted to guide methylation at putative sites on Dm rRNA that are not modified in yeast (Am1374/18S, Um1906/18S, Cm1813/18S and Cm2933/28S). Hence, these specimens could not be targeted by our approach. Conversely, we noted that SNOSCAN also allowed identification of snoRNAs predicted to methylate Dm rRNA at sites not homologous to yeast (Am425/18S, Cm1643/18S, Gm1322/28S, U1332/28S, Am2466/28S, Cm3049/28S, Gm3112/28S and Gm3403/28S), three of which were experimentally confirmed (Am425/18S, Um1332/28S and Gm3112/28S). However, for each of the newly identified snoRNAs conclusive corroboration of their function as methylation guides will require additional experiments. In view of the large bulk of *Drosophila* genetic resources available on the Fly base (http://flybase.bio.indiana.edu), mutations affecting the transcription of the host gene or deficiencies covering the genomic position of the snoRNA, may have already been described for some of these snoRNAs. If homozygous and viable, such strains can easily be analysed for rRNA methylation at the predicted sites to confirm the functional role of the specific snoRNA. An alternative and more generally applicable approach would be to prepare strains (or cultured cells) in which each snoRNA gene has been specifically and stably silenced. dsRNA interference (RNAi) may prove an effective and straightforward tool for this type of experiment, since it has achieved stable silencing of snoRNA genes in *Trypanosoma* (Liang *et al*., 2003).

Though certainly not exhaustive, our search has identified 20 new snoRNA genes in the Dm genome that had escaped previous molecular analyses. Most of them are represented in multiple copies. It is thus clear that many snoRNAs may be missed by cDNA cloning procedures, either because of their relatively low abundance or because of factors potentially affecting their representation in cDNA libraries. These factors may include sharply restricted expression profiles, sequences of the linker oligonucleotides utilized for their capture or strategies for excluding abundant RNA species. Conversely, snoRNAs identified by cloning may be missed in a computational analysis, mainly because their structure is significantly different from the characteristic key signatures. cDNA cloning and computational screening are thus complementary approaches. A comprehensive picture of the full complement of ncRNAs genes hidden in the genome evidently demands the large-scale employment of different and complementary procedures.

CONCLUSIONS

The results of our genome-wide computational search for new box C/D snoRNA genes in Dm show that this approach is straightforward and effective, since it led to the identification of 20 snoRNA genes that had escaped previous molecular analyses. Several snoRNAs were encoded within an exon of a protein-coding gene, in genomic regions subjected to alternative splicing or located across an exon/intron boundary. These new types of arrangement may reflect a still uncharacterized mechanism of snoRNA biogenesis based on alternative production of the overlapping mRNA/snoRNA. The very limited overlap between our results obtained by computational search and those provided by the molecular approach based on the preparation of sized cDNA libraries, indicates that these two procedures are complementary. This assumption, indeed, may be universally valid, since the same conclusion has been drawn with regard to a recent comprehensive search for snoRNA genes in the rice genome (Chen *et al*., 2003). Development of new and more effective programs for RNA computational analysis is expected to strongly contribute to the rapid progress of the RNomics field in the near future.

ACKNOWLEDGEMENTS

M.F. was supported by Telethon Grant no. GGPO30120 and R.C. was supported by FIRB Grant RBNE0157EH_006.

REFERENCES

Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F., *et al*. (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **24**, 2185–2195.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J Mol. Biol*., **5**, 403–410.

Bachellerie,J.P., Cavaillè,J. and Hüttenhofer,A. (2002) The expanding snoRNA world. *Biochimie*, **84**, 775–790.

Barneche,F., Gaspin,C., Guyot,R. and Echeverria,M. (2001) Identification of 66 box C/D snoRNAs in *Arabidopsis thaliana*: extensive gene duplications generated multiple isoforms predicting new ribosomal RNA 2′-O-methylation sites. *J. Mol. Biol.*, **311**, 57–73.

Cavaillè,J., Buiting,K., Kiefmann,M., Lalande,M., Brannan,C.I., Horsthemke,B., Bachellerie,J.P., Brosius,J. and Hüttenhofer,A. (2000) Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *Proc. Natl Acad. Sci. USA*, **19**, 14311–14316.

Chen,C.L., Liang,D., Zhou,H., Zhuo,M., Chen,Y.Q. and Qu,L.H. (2003) The high diversity of snoRNAs in plants: identification and comparative study of 120 snoRNA genes from *Oryza sativa*. *Nucleic Acids Res.*, **15**, 2601–2613.

Kiss,T. (2002) Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell*, **19**, 145–148.

Kiss-Laszlo,Z., Henry,Y., Bachellerie,J.P., Caizergues-Ferrer,M. and Kiss,T. (1996) Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. *Cell*, **28**, 1077–1088.

Liang,X.H., Liu,Q. and Michaeli,S. (2003) Small nucleolar RNA interference induced by antisense or double-stranded RNA in trypanosomatids. *Proc. Natl Acad. Sci. USA*, **100**, 7521–7526.

Lowe,T.M. and Eddy,S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, **19**, 1168–1171.

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

Nicoloso,M., Qu,L.H., Michot,B. and Bachellerie,J.P. (1996) Intron-encoded, antisense small nucleolar RNAs: the characterization of nine novel species points to their direct role as guides for the $2' - O$-ribose methylation of rRNAs. *J. Mol. Biol.*, **12**, 178–195.

Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Tycowski,K.T., Shu,M.D. and Steitz,J.A. (1996) A mammalian gene with introns instead of exons generating stable RNA products. *Nature*, **379**, 464–466.

Tycowski,K.T. and Steitz,J.A. (2001) Non-coding snoRNA host genes in *Drosophila*: expression strategies for modification guide snoRNAs. *Eur. J. Cell Biol.*, **80**, 119–125.

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.*, **15**, 2495–2507.