 Nucleic Acids Research, 2000, Vol. 28, No. 1 **197–200**

Non-coding, mRNA-like RNAs database Y2K

Volker A. Erdmann, Maciej Szymanski[1], Abraham Hochberg[2], Nathan de Groot[2] and Jan Barciszewski[1,*]

Institut fur Biochemie, Freie Universitat Berlin, Thielallee 63, 14195 Berlin, Germany, [1]Institute of Bioorganic Chemistry of the Polish Academy of Sciences, Noskowskiego 12, 61704 Poznan, Poland and [2]Department of Biological Chemistry, The Hebrew University, IL-91904 Jerusalem, Israel

Received September 21, 1999; Revised and Accepted October 4, 1999

ABSTRACT

In last few years much data has accumulated on various non-translatable RNA transcripts that are synthesised in different cells. They are lacking in protein coding capacity and it seems that they work mainly or exclusively at the RNA level. All known non-coding RNA transcripts are collected in the database: http://www.man.poznan.pl/5SData/ncRNA/index.html

INTRODUCTION

Eukaryotic genes contain clearly identifiable open reading frames (ORFs) that direct the translation of functional proteins. However, not all RNA transcripts (other than tRNA, rRNA or snRNA) are translated into polypeptides. Many non-translatable mRNA-like RNA transcripts have been found in the cell. They are polyadenylated, spliced and are lacking long ORFs. It has been suggested that in the absence of protein product they function as RNA riboregulators and/or are involved in regulation of expression of linked genes. There is a wealth of data that RNA plays a variety of structural, informational, catalytic and regulatory roles in the cell.

CHARACTERISTICS OF NON-CODING RNA

These RNAs can be divided into the following classes:

- Gene regulators: *Xist, roX1, Tsix, XistAS, roX2, H19, mei, LPW, KvDMR1, DGCR5, CMPD*
- Abiotic stress signals: *gadd7/adapt15, adapt33, hsrω, OxyR, DsrA, lbi, G90*
- Biotic stress signals: *His-1, ENOD40, CR20, GUT15*
- Others: *UHG, NTT, Bsr, BC1, BC200, SRA*

Xist, roX RNAs

A subset of mammalian genes is subject to genomic imprinting. Three of them: *Xist*, *H19* and *IPW* are known. The *XIST*/*Xist* gene is expressed exclusively from the inactive X-chromosome in somatic cells and produces an RNA of 17 kb in man and 15 kb in mouse. Multiple copies of *Xist* transgenes inserted into mouse chromosome 12 produced RNA transcript which coated the autosome, became hypoacetylated and late replicating. The developmental up-regulation is a consequence of the stabilisation of *Xist* RNA due to a promoter switch, which is an effect of hyperacetylation of chromatin upstream of *Xist*. Deacylation on the other hand leads to the further progression of X-inactivation (1–7). In *Drosophila* a mechanism of dosage compensation involves the products of the two genes *roX1* and *roX2* which do not encode proteins. *roX1* RNA becomes associated with the X chromosome at sites determined by binding of the *msl* (male specific lethal) gene products (3). It has been shown that deleting a 65 kb region downstream of *Xist* results in constitutive expression and X-inactivation, implying the presence of a *cis*-regulatory element. In this region there is a gene *Tsix* antisense to *Xist*. It is a 40 kb RNA originating 15 kb downstream of *Xist* and transcribed across the *Xist* locus. Human *Tsix* RNA has no conserved ORFs, occurs exclusively in the nucleus and is localised at *Xic*, similarly to antisense *Xist* (*XistAS*) of mouse (8,9).

H19 RNA

The *H19* gene codes for a capped, spliced and polyadenylated oncofetal 2.3 kb RNA. It is expressed exclusively from a maternal chromosome and is linked and co-regulated with the insulin-like growth factor 2 gene. The genes isolated from human and mouse consist of five exons separated by four unusually short introns. In the human sequence, there is a putative short ORF that would encode a 26 kDa protein, but no translation product has been identified so far. The evolutionarily conserved 42 bp element located upstream of the imprinting mark might play a role in imprinting and/or transcriptional regulation of *H19*. The G-rich repeat 1.5 kb upstream of mouse H19 was shown to be present in rats but absent in humans, and is not essential for H19 imprinting (10–12).

IPW RNA

IPW has been mapped in the Prader–Willi syndrome region. The IPW (human) and ipw (mouse) RNA transcripts contain no long ORFs, are alternatively spliced and contain multiple copies of a 147 bp repeat arranged in a head-to-tail orientation. However unlike *H19* and *Xist*, *IPW* is poorly conserved between mouse and human in both overall structure and nucleotide sequence. The region of similarity between mouse and human genes is restricted to a 500 bp region (13–15).

DGCR5

DGCR5 is a gene from the DiGeorge syndrome region that codes for an untranslated RNA. It could be implicated in the control of expression of nearby genes (16).

*To whom correspondence should be addressed. Tel: +48 61 852 8503; Fax: +48 61 852 0532; Email: jbarcisz@ibch.poznan.pl

KvDMR1

Loss of imprinting at IGF2 is associated with the cancer predisposition condition Beckwith–Wiedeman syndrome. Imprinting control elements are proposed to exist within the KvLQT1 locus. The KvDMR1-associated RNA is transcribed exclusively from the paternal allele and in the reverse orientation with respect to the maternally expressed KvLQT1 gene (17).

CMPD-associated RNA

Analysis of chromosomal translocations in patients with campomelic dysplasia (CMPD) suggested that disruption of genes in the 17q24–q25 region of chromosome 17 may be responsible for this disease that often involves sex reversal in genotypical males. Positional cloning allowed the identification and isolation of 3.5 kb cDNA which is specifically expressed in adult testis. There are no long ORFs detected in this RNA and no detectable protein produced *in vitro* (18).

Oxidative stress response RNAs (*gadd7*/*adapt15*, *adapt33*)

In mammals several groups of genes are expressed in response to UV irradiation or exposure to hydrogen peroxide and other reactive oxygen species. One of these genes, *gadd7*, is expressed in response to treatment with UV radiation. The *gadd7* transcript is a 0.9 kb long polyadenylated RNA species, lacking long ORFs although there are three short ORFs for peptides of length 37, 38 and 43 amino acids. Other RNA species produced in response to oxidative stress, induced by hydrogen peroxide are *adapt15* and *adapt33*. They have been isolated from Chinese hamster cells. Transcription products of both of these genes lack long ORFs and, similarly to *gadd7* RNA, are polyadenylated. The *adapt15* RNA is 0.95 kb long and is almost identical to *gadd7* RNA, whereas for *adapt33*, two homologous RNA species of 1.46 and 0.99 kb have been isolated (1).

In addition, v-src end associated peptide 1 RNA (vseap1) shows high similarity to gadd7/adapt 15 RNA (19).

G90

The murine *G90* gene has been identified by substractive hybridisation based on the differential presence of its transcript in large and small intestine. The full-length cDNA produces a 1.5 kb transcript that is polyadenylated but has no ORF larger than 248 bp. G90 is transcribed at very high levels in the small intestine and at lower levels in the large intestine, testis and kidney of the mouse (20).

OxyS RNA

In *Escherichia coli* an expression of hydrogen peroxide-inducible proteins is controlled by the transcriptional regulator, OxyR. It also induces the expression of an abundant 109 nt untranslated regulatory oxyS RNA. This RNA acts as a regulator of as many as 40 genes that integrate adaptation to hydrogen peroxide with other cellular stress responses and help to protect cells against oxidative damage. OxyS RNA can pair with a short sequence overlapping the Shine–Dalgarno sequence in the mRNA-encoding delta subunit of RNA polymerase (1,21).

DsrA RNA

The regulation of capsular polysaccharide synthesis in *E.coli* depends on the level of an unstable positive regulator, RcsA. The amount of RcsA protein is limited both by its rapid degradation and by its low level of protein synthesis. The last effect is due to transcriptional silencing by the histone-like protein H-NS. A small, 85 nt RNA, DsrA RNA, when overproduced activates transcription by counteracting H-NS silencing (1).

lbi RNA

The lbi (lipopolysaccharide biosynthesis interfering) RNA of phage Acm1, an untranslated RNA species of 97 nt, was shown to affect the biosynthesis of the O-specific polysaccharide of lipopolysaccharide (LPS) in various enterobacterial strains. Several lines of evidence suggested that lbi RNA may act *in trans* by an antisense-like mechanism regulating the activity of cellular target RNAs by formation of RNA hybrid molecules (22,23).

Heat shock response RNA (hsrω)

Protection against various environmental stresses is mainly conferred by the induction of heat shock genes. In *Drosophila* a major site of transcription in heat shock is the *hsrω* gene. If this gene is polyadenylated and spliced, the product RNA has only very small ORFs. The expression of *hsrω* is constitutive and is elevated by heat shock. In both normal and stressed cells the transcription from the *hsrω* locus gives rise to three transcripts *ω1*, *ω2* and *ω3*. All of them have the same transcription start site. The hsr*ω1* transcript is ~10 kb long and results from transcription of the whole locus. At the 3′-end of hsr*ω1* transcript there is a 7–8 kb long region consisting of a short tandem repeat unit. The hsr*ω2* (~1.9 kb) transcript results from the use of an alternative polyadenylation site located just upstream of the tandem repeats region. The *ω*3 transcript (1.2–1.3 kb) is produced from the *ω2* transcript by removal of a single intron and is a cytoplasmic species. This RNA does not contain any long ORFs. One of the short ORFs that is conserved in three *Drosophila* species would encode a peptide of length 23–27 amino acids (1).

His-1 RNA

Upon viral insertion in murine myeloid leukemias, the *His-1* gene is activated. The *His-1* gene is a single-copy sequence that has been found in a variety of vertebrate species. *His-1* RNA is expressed as a 3 kb long, spliced and polyadenylated RNA. It lacks long ORFs. *His-1* RNA transcription is correlated with viral insertion and carcinogenesis, since no transcripts have been detected in normal tissues. This finding suggests that the expression of the *His-1* gene is highly restricted and that its inappropriate activation may contribute to carcinogenesis (1).

Brain specific RNAs (Bsr RNA, BC1 RNA, BC200 RNA)

Brain specific repetitive RNA (Bsr RNA) is a non-coding RNA that consists of tandem repeats of similar sequences, ~0.9 kb. Bsr RNA is preferentially expressed in the rat central nervous system (24).

BC1 RNA is generated by retroposition of $tRNA^{Ala}$. It is 120 nt long and is expressed preferentially in the brain by polymerase III. The sequence similarity between $tRNA^{Ala}$ and the 5′ half of BC1 RNA is ~80%, however this domain does

not fold in a tRNA like manner, but forms a stable stem–loop structure. The gene arose after mammalian radiation but prior to the diversification of the order Rodentia. The level of this RNA increases during the period of synaptogenesis. It plays a role in transport and translational regulation (25).

BC200 RNA is a small cytoplasmic non-coding RNA that has been identified in a number of primate species. Its expression is almost exclusively restricted to neural tissues. In the cell it is present as a ribonucleinoprotein particle. The gene encoding BC200 arose from a monomeric Alu element and the encoded RNA had been recruited into a function in the nervous system (26).

SRA RNA

SRA RNA, steroid receptor RNA activator, is selective for steroid hormone receptors and mediates transactivation via their N-terminal activation function. The transcript is 0.7–0.85 kb. It is a bona fide transcriptional coactivator selective for AF1 of steroid receptors. It is expressed as multiple isoforms in a cell-specific manner and is present in a steady-state coregulator complex with the AF2 coactivator SRC-1 (27).

MeiRNA

Fission yeast protein Mei2 is an RNA binding protein required for both premeiotic DNA synthesis and meiosis I. It binds to polyadenylated MeiRNA of ~0.5 kb, loss of which blocks meiosis I. It is required for nuclear localisation of Mei2 (28).

NTT RNA

NTT is human gene (non-coding transcript in T cells) which produces 17 kb non-coding, polyadenylated, but not spliced, nuclear RNA, expressed in activated $CD4^+$ T cells. NTT has no ORF larger than 270 bp. There is a single copy of the *NTT* gene per haploid genome and both alleles are transcriptionally active (1).

UHG

Some small nucleolar RNAs (snoRNAs) are encoded within different introns of the unusually compact mammalian U22 host gene (*UHG*). U22 RNA is essential for the maturation of 18S rRNA. The U22 snoRNA gene lies within an intron of the single copy gene *UHG* which encodes a polyadenylated non-coding RNA. UHG RNAs in human and in mouse are 1114 and 590 nt long, respectively. Recently three other non-protein-coding snoRNA host genes: *Gas5*, *U17HG* and *U19HG* have been identified. Gas5 spliced RNA becomes polyadenylated and associates with translating ribosomes. Both the Gas5 and H17HG transcripts initiate with a cytosine followed by a pyrimidine-rich tract and belong to the 5′ terminal oligopyrimidine gene family (5′-TOP). U19HG after splicing remains in the nucleus and has a 5′ sequence reminiscent of 5′-TOP (29).

Early nodulin 40 (*ENOD40*)

Genes that are transcribed early in the plant–bacteria interaction (*ENOD* genes) are hypothesised to play a role in organogenesis and bacterial invasion. One of them, *ENOD40*, an early nodulin gene, is expressed following inoculation with *Rhizobium meliloti* or by adding nodulation factors or cytokinin to non-inoculated roots. *ENOD40* encodes an RNA with a regulatory function and shows a very short ORF of only 10–13 amino acids (30,31).

Cytokinin response RNA (CR20)

The *CR20* gene has been found to be one of several genes repressed by cytokinins. There are several CR20 transcripts generated by alternative splicing of the precursor RNA which do not contain a long ORF. A comparison of the two known CR20 sequences from cucumber and *Arabidopsis thaliana* has revealed a highly conserved 180 nt region that seems to form a stable secondary structure (1).

GUT 15 RNA

The tobacco genes with unstable transcripts (*GUT*) have been isolated and characterised. The GUT15 RNA is polyA+, contains at least one intron but does not contain a long ORF. It has a long segment similar to that of CR20 (32). Similar RNAs may also be present in aspen, poplar, soybean, rice and maize (A.van Hoof, personal communication).

Xlsirt RNA

Xlsirt RNAs have been identified as RNA species that are localised to the vegetal cortex of *Xenopus* oocytes during early stages of oogenesis. It has been proposed that they may play a structural role and may be used to localise other RNAs. Xlsirts form a family of heterogenous transcripts originating from both strands of the genes. The transcripts from one strand (sense strand) are localised in the vegetal cortex, while others are found throughout the cytoplasm. The transcripts contain 3–13 repeat units that are flanked by unique sequences (33,34).

ACKNOWLEDGEMENTS

We wish to thank Dr A.van Hoof for helpful comments. This work has been supported by the DFG (Gottfried Wilhelm Leibnitz Prize to V.A.E.), the SFB 344-C8, the DFG Trilateral Research Project Genomic Imprinting in Human Bladder Cancer and the Polish State Committee for Scientific Research.

REFERENCES

1. Erdmann,V.A., Szymanski,M., Hochberg,A., de Grot,N. and Barciszewski,J. (1999) *Nucleic Acids Res.*, **27**, 192–195.
2. Brockdorf,N. (1988) *Curr. Opin. Genet. Dev.*, **8**, 328–333.
3. Franke,A. and Baker,B.S. (1999) *Mol. Cell*, **4**, 117–122.
4. Chao,Y.-C., Lee,S.-T., Chang,M.-C., Chen,H.-H., Chen,S.-S., Wu,T.-Y., Liu,F.-H., Hsu,E.-L. and Hou,R.F. (1998) *J. Virol.*, **72**, 2233–2245.
5. Brockdorf,N. and Duthie,S.M. (1998) *Cell Mol. Life Sci.*, **54**, 104–112.
6. Duthie,S.M., Nesterova,T.B., Formstone,E.J., Keohane,A.M., Turner,B.M., Zakian,S.M. and Brockdorf,N. (1999) *Hum. Mol. Genet.*, **8**, 195–204.
7. O'Neill,L.P., Keokhane,A.M., Lavender,J.S., McCabe,V., Heard,E., Avner,P., Brockdorf,N. and Turner,B.M. (1999) *EMBO J.*, **18**, 2897–2907.
8. Lee,J.T., Davidov,L.S. and Varshawsky,D. (1999) *Nature Genet.*, **21**, 400–404.
9. Mise,N., Goto,Y., Nakajima,N. and Takagi,N. (1999) *Biochem. Biophys. Res. Commun.*, **258**, 537–541.
10. Hurst,L.D. and Smith,N.G.C. (1999) *Trends Genet.*, **15**, 134–135.
11. Frevel,M.A.E., Hornberg,J.J. and Reeve,A.E. (1999) *Trends Genet.*, **15**, 216–218.
12. Stadnick,M.P., Pieracci,F.M., Cranston,M.J., Taksel,E., Thorvaldsen,J.L. and Bartolomei,M.S. (1999) *Dev. Genes Evol.*, **209**, 239–248.
13. Tilgman,S.M. (1999) *Cell*, **96**, 185–193.
14. Brannan,C.I. and Bartolomei,M. (1999) *Curr. Opin. Genet. Dev.*, **9**, 164–170.
15. Wevrick,R. and Francke,U. (1997) *Hum. Mol. Genet.*, **6**, 325–332.
16. Southerland,H.F., Wadey,R., McKie,J.M., Taylor,C., Atif,U., Johnstone,K.A., Halford,S., Kim,U.J., Goodship,J., Baldini,A. and Scambler,P.J. (1996) *Am. J. Hum. Genet.*, **59**, 23–31.

17. Smilinich,N.J., Day,C.D., Fitzpatrick,G.V., Caldwell,G.M., Lossie,A.C., Cooper,P.R., Smallwood,A.C., Joyce,J.A., Schofield,P.N., Reik,W., Nicholls,R.D., Weksberg,R., Driscoll,D.J., Maher,E.R., Shows,T.B. and Higgins,M.J. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 8064–8069.
18. Ninomiya,S., Isomura,M., Narahara,K., Seino,Y. and Nakamara,Y. (1996) *Hum. Mol. Genet.*, **5**, 69–72.
19. Mizenina,O., Yanushevich,Y., Musatkina,E., Rodina,A., Camonis,J., Tavitian,A. and Tatosyjan,A. (1998) *FEBS Lett.*, **422**, 79–84.
20. Krause,R., Hemberger,M., Himmelbauer,H., Kalscheuer,V. and Fundele,R.H. (1999) *Gene*, **232**, 35–42.
21. Gonzalez-Flecha,B. and Demple,B. (1999) *J. Bacteriol*., **181**, 3833–3836.
22. Mammat,U., Rietschel,E.T. and Schmidt,G. (1995) *Mol. Microbiol.*, **15**, 1115–1125.
23. Warnecke,J.M., Moolenaar,C.E.C., Ritschel,E.Th., Hartmann,R.K. and Mamat,U. (1999) 4th Annual Meeting of the RNA Society, Edinburgh, Scotland: 23–27 June, Abstract 593.
24. Komine,Y., Tanaka,N.K., Yano,R., Takai,S., Yuasa,S., Shiiroishi,T., Tsuchiya,K. and Yamamori,T. (1999) *Brain Res. Mol. Brain Res.*, **66**, 1–13.
25. Brosius,J. (1999) *Gene*, in press.
26. Skryabin,B.V., Kremerskothen,J., Vassilacopoulou,D., Disotell,T.R., Kapitonov,V.V., Jurka,J. and Brosius,J. (1998) *J. Mol. Evol.*, **47**, 677–685.
27. Lanz,R.B., McKenna,N.J., Onate,S.A., Albrecht,U., Wong,J., Tsai,S.Y., Tsai,M.-J. and O'Malley,B.W. (1999) *Cell*, **97**, 17–27.
28. Ohno,M. and Mattaj,I.W. (1999) *Curr. Biol.*, **9**, R66–R69.
29. Weinstein,L.B. and Steitz,J.A. (1999) *Curr. Opin. Cell Biol.*, **11**, 378–384.
30. Mirabella,R., Martirani,L., Lamberti,A., Iaccarino,M. and Chiurazzi,M. (1999) *Plant Mol. Biol*., **39**, 177–181.
31. Kouchi,H., Takane,K., So,R.B., Ladha,J.K. and Reddy,P.M. (1999) *Plant J*., **18**, 121–129.
32. van Hoff,A., Kastenmayer,J.P., Crispin,B., Taylor,C.B. and Gree,P.J. (1995) *Plant Gene Register Plant Physiol*., **113**, 1004.
33. Kloc,M., Spohr,G. and Etkin,L.D. (1993) *Science*, **262**, 1712–1714.
34. Kloc,M. and Etkin,L.D. (1994) *Science*, **265**, 1101–1103.