

# Noncoding RNA Genes

Sean R. Eddy

Dept. of Genetics

Washington University School of Medicine

St. Louis, MO 63110 USA

Phone: +1-314-362-7666

FAX: +1-314-362-7855

eddy@genetics.wustl.edu

## Summary

Some genes produce functional RNAs instead of encoding proteins. Noncoding RNA genes are surprisingly numerous. Recently active research areas include small nucleolar RNAs, antisense riboregulator RNAs, and RNAs involved in X dosage compensation. Genome sequences and new algorithms have begun to make systematic computational screens for noncoding RNA genes possible.

# Introduction

It may sound like a script from the X-Files but it's all true. There is a class of genes whose final products are not proteins; instead, they have a very different biochemistry. Some of these genes originated before the last common ancestor of life on our planet. Some people believe that they are vestiges of ancient lifeforms that preceded modern DNA/protein based life. Their total numbers remain mysterious, in part because they are invisible to the genefinding programs used to analyze genome sequence data.

These enigmatic genes are the genes for noncoding RNAs (ncRNAs). Noncoding RNA genes produce transcripts that do not encode proteins, but rather function directly as RNAs. Interest in RNA structure and function was spurred by the discovery of RNA catalysis [1\*\*]. RNA's ability to act as both genetic template and biochemical catalyst led to the proposal of the RNA World hypothesis for the origin of life [1\*\*]. The RNA World proposes that RNA-based life preceded both DNA and protein in evolution [2]. Some have proposed that some modern ncRNA genes may be molecular fossils of the RNA World [3\*]. There is therefore great interest in studying the evolution of ncRNAs [4,5\*].

The availability of complete genome sequence data in many organisms makes it attractive to try to identify ncRNA genes systematically by computational analysis, much as new protein genes are discovered. In this review, I discuss noncoding RNA genes from a perspective of large scale computational genome analysis: first, how many ncRNA genes

are there, and second, how do we find them?

## The usual suspects

Ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), and the small nuclear RNA (snRNA) components of the spliceosome are probably the most intensively studied ncRNAs. Their genes are numerous because most (particularly rRNA and tRNA) occur in multiple redundant copies. The yeast *Saccharomyces cerevisiae*, for example, has well over 700 of these genes: 275 transfer RNA genes, 100-200 copies each of four different ribosomal RNA genes (16S, 25S, 5.8S, and 5S) and perhaps about 40 genes in all for the five spliceosomal snRNAs (U1, U2, U4, U5, and U6) [6]. In comparison, yeast has about 6000 protein-coding genes [6]. In most model organisms, many representatives of these genes have already been identified (usually biochemically), and additional copies are usually identified readily in genome sequence by standard similarity searches (e.g. BLAST) or, in the case of tRNAs, specialized programs [7].

## A rogue's gallery

Numerous other ncRNAs have been identified. They have a surprisingly diverse range of functions. RNase P RNA, which processes transfer RNAs (and some other RNAs), is one

of the few natural catalytic RNAs [8]. Signal regulatory particle RNA is involved in translocating proteins across the endoplasmic reticulum [9]. Bizarre numbers of guide RNAs are responsible for RNA editing in trypanosomes [10]. Telomerase RNA functions as the template for adding new telomeres in most eukaryotes [11]. *Schizosaccharomyces pombe* meiRNA is involved in regulating the onset of meiosis [12]. In bacteria, tmRNA is involved in targeting aberrant partial protein products of truncated mRNAs for rapid degradation [13,14]. A 120 nt bacteriophage  $\phi$ 29 RNA forms a hexameric structure that is essential (at least *in vitro*) for packaging of DNA into the phage head [15\*].

One very active area is the role of RNAs in X dosage compensation. The mammalian *Xist* RNA coats the inactivated X chromosome and is necessary for X chromosome inactivation [16\*,17]. *Xist* may itself be regulated by an antisense ncRNA transcript, *Tsix* [18]. X dosage compensation in *Drosophila* also involves two small ncRNAs called *roX1* and *roX2* [17,19,20].

There are many natural 'antisense' RNAs that act as 'riboregulators' [21]. In *E. coli*, OxyS RNA [22,23], DsrA RNA [24,25], and MicF RNA [21] all act at the level of translational initiation, by base pairing upstream of the initiator AUG and either blocking translational initiation or (in the case of MicF) competing away an inhibitory cis secondary structure to free the initiation site and thereby activate translation. In eukaryotes, the best studied small riboregulator is the 22 nt Lin-4 RNA, which regulates the developmental timing of larval molts in the nematode *Caenorhabditis elegans*. Lin-4 binds to the 3' UTRs of at least two target mRNAs (from the protein-coding genes *lin-14* and *lin-28*) and

posttranscriptionally represses the expression of these genes by an as yet unknown mechanism [26,27,28].

There are also several apparently noncoding RNA transcripts whose function is unclear. A fascinating example is *E. coli* 6S RNA, one of the first RNAs ever sequenced, but which still has no known function [29]. Others include the abundant *Drosophila hsr-omega* transcript induced by heat shock [30], and the human *H19* transcript [31]. A database that collates the sequences of many of these transcripts has been established [32].

## More sno storms predicted

Eukaryotic nucleoli are currently the richest source of new ncRNAs – the small nucleolar RNAs (snoRNAs) [33\*\*,34,35,36]. snoRNAs are involved in posttranscriptional processing and modification of ribosomal RNA. Almost all snoRNAs fall into two classes based on sequence features: the C/D box snoRNAs and the H/ACA box snoRNAs. Most C/D box snoRNAs guide 2'-O-ribose methylation at specific sites in ribosomal RNA, and the H/ACA snoRNAs guide site-specific pseudouridylation. Both snoRNA families use complementary base-pairing to a target rRNA site to specify the position of a modification. Eukaryotic ribosomal RNAs are heavily modified, with up to hundreds of specific methylations and pseudouridylations, and there is a corresponding abundance of distinct snoRNA species. Vertebrate genomes are thought to contain several hundred snoRNA

genes. A computational screen in the yeast genome brought the total number of known yeast C/D methylation guide snoRNAs up to 41 [37\*].

There are some early indications that snoRNA-guided modification systems operate on targets other than the ribosomal RNAs. Vertebrate spliceosomal snRNAs are heavily modified (24 pseudouridylations and 30 ribose methylations in the five rat snRNAs). Two new guide snoRNAs have been identified by a computational search and shown to direct two methylations on U6 snRNA [38\*\*]. A target sequence artificially expressed in an mRNA context can also be modified (albeit inefficiently) [39]. Since most of the known snoRNAs were originally isolated biochemically [40], the known snoRNAs could easily be biased towards abundant guide RNAs for abundant targets (e.g. rRNA); similarly, computational snoRNA screens have required antisense complementarity to specific targets as part of their search criteria [37\*,38\*\*,41]. Genes for snoRNAs that direct modifications at non-rRNA targets probably still lurk in eukaryotic genomes.

## **Inside-out genes**

In vertebrates, all known snoRNAs (except U3, U8, U13, and MRP) are found in introns of pre-mRNAs [33\*\*,35]. The snoRNAs are released from their host transcripts by nucleolytic processing.

Even more surprisingly, at least four mammalian snoRNA host genes seem to be ncRNAs themselves. The sole function of these host genes appears to be to carry a payload of intron-encoded snoRNA genes. The *gas5*, *UHG*, *U17HG*, and *U19H* genes produce mRNAs that have no apparent open reading frame, and their exons are not recognizably conserved between mouse and man [42\*,43,44,45]. In each gene, one or more introns of their pre-mRNAs has a highly conserved region corresponding to a snoRNA gene; for example, human *gas5* hosts ten snoRNAs, and human *UHG* hosts eight snoRNAs.

The *gas5* story is interesting from the standpoint of genome analysis. The gene was identified in a subtractive hybridization screen many years ago [46]. It was assumed to be a protein coding gene [47] until the U80 snoRNA was serendipitously found in *gas5* genomic sequence by a similarity search [42\*].

## Cautionary tales

Indeed, there are other examples of ncRNA genes being initially overlooked because only protein-coding genes are expected. An interesting case appeared in the recent literature. In *Escherichia coli* and the plant pathogen *Erwinia carotovora*, there is a global posttranscriptional regulatory pathway called Csr (carbon storage regulation) [48\*]. A key gene in the *Erwinia* pathway, *aepH*, was defined genetically by transposon insertion mutagenesis [49]. Several insertions fell inside and just upstream of a small 47 amino acid

open reading frame, which was then assumed to be *aepH*. The predicted product is still available in the public protein sequence database (Genbank pid g691744).

Meanwhile, the homologous pathway was being dissected in *E. coli*. One component, the *E. coli* CsrA protein, was identified in a genetic screen for glycogen overproduction mutants, and then shown biochemically to be a small RNA-binding protein that probably acts as a translational repressor [48\*,50]. Purification of the CsrA protein from *E. coli* extracts unexpectedly showed an associated 360 nt RNA, which was named CsrB [51]. The CsrB RNA represses CsrA, probably by sequestering CsrA protein and preventing it from binding its downstream targets. The *csrB* ncRNA gene is conserved in *Erwinia*, and overlaps with the *aepH* 'protein coding' region (which is otherwise not conserved); this and other data showed that the *Erwinia* locus did not encode a protein, but instead encodes a noncoding RNA, now renamed CsrB [51].

## Finding hairpins in a haystack

From all the examples above, and with new examples arriving, it seems clear that there are many ncRNAs to be found in genomes. How should they be identified in a genome sequence? The most straightforward computational gene identification approach is database similarity searching. Statistically-based similarity searching (e.g. BLAST and FASTA) is taken for granted in protein gene analysis, but is more difficult for ncRNAs.

Many noncoding RNAs evolutionarily conserve a consensus secondary structure more than they conserve primary sequence, so primary sequence alignment methods are often not very effective.

If the consensus secondary structure of an RNA is known, and has at least one highly conserved region, one can define an exact-match pattern and use it to search sequence databases using any of several available software packages [52,53,54]. If the consensus structure tolerates significant structural variations (as most do), statistical similarity search techniques are now available. Mathematical models called ‘stochastic context-free grammars’ (SCFGs) been introduced as a sound statistical framework for RNA structure/sequence alignment [55\*]. Probably the first practical SCFG application was tRNAscan-SE, which detects transfer RNA genes in DNA sequence [7]. A model used to computationally detect a number of new snoRNA genes in the yeast genome also included SCFG components [37\*]. SCFG methods are computationally demanding on current computers; both of the cited examples used shortcuts to reduce this complexity.

## **Here’s looking for you, kid**

Probably the most significant open problem in computational ncRNA analysis is genefinding, where one tries to find novel ncRNAs in genome sequence without using similarity information. It is unclear what statistical signals can be used to detect genes

that have no open reading frames and no codon bias. Only a few approaches have been tried so far, and they have had limited success.

Maizel's group has explored methods that look for regions of a genome with predicted RNA structures that are significantly more thermodynamically stable than random sequence of the same base composition [56,57]. This approach detects a few highly structured ncRNAs (as well as a few cis-regulatory structures) but does not appear to work in general (E Rivas, SR Eddy, unpublished data).

Olivas et al. searched for polIII promoter consensus sequences in the yeast genome, and experimentally confirmed that one of their candidates was indeed a novel ncRNA [58\*\*]. However, many RNAs are expressed from polII promoters (which are typically more difficult to predict); also, the snoRNAs provide examples of ncRNAs that do not have their own promoters at all.

In coding-dense genomes, suspicious-looking large regions with little or no coding potential have been dubbed 'grey holes' [59]. Olivas et al. examined 59 grey holes of  $\geq 2$  kb in the *Saccharomyces cerevisiae* genome [58\*\*]. Northern analysis detected distinct transcripts from 15 of the grey holes. One transcript appears to be an H/ACA snoRNA. The remaining transcripts may either be additional ncRNAs, or may encode short ORFs. One 2.1 kb grey hole in yeast was already known to contain the telomerase ncRNA gene *tlc1*, originally identified in a genetic screen for high-copy suppressors of telomeric silencing [60].

## Conclusions

A large number of ncRNA genes have been discovered, but on the other hand, most systematic genomic screens for new genes are biased against discovering ncRNAs. (As just one example, the whole reason to prepare an oligo dT selected cDNA library is to deplete an RNA population of noncoding RNAs.) Genetic screens are also somewhat biased against ncRNAs, because ncRNAs are usually small, sometimes present in multiple redundant copies, and are immune to frameshift or nonsense mutations. There are undoubtedly many more ncRNAs to be found.

The functions of the known ncRNAs are diverse. A common theme is that many are involved in specific recognition of nucleic acid targets via complementary base pairing. This is a function that RNA is well suited for, more so than protein. For example, it must have been easier to evolve one protein 2'-O-ribose methylase that interacts with a hundred small guide snoRNAs with different ~15 nucleotide complementary targeting sequences, than to evolve a couple of hundred protein methylases with ~100 amino acid RNA-binding domains having different precise binding specificities. This observation supports an argument against an RNA World origin for many ncRNAs. There is ample reason for evolution to invent new ncRNAs even in a protein/DNA world. From this perspective, the ncRNAs with no apparent nucleic acid target are interesting: what is signal recognition particle RNA doing, for example?

Genome sequence data makes ncRNA discovery easier, but only somewhat. The advent of SCFG search methods means that similarity searching for structurally homologous ncRNAs is headed for firmer ground. However, the lack of useful statistical signals common to all ncRNA genes means that novel ncRNA gene finding will remain difficult. ncRNA gene finders may have to be specialized for particular subsets of genes (polIII-transcribed ncRNAs, highly structured ncRNAs, and so on).

One powerful computational approach that will become possible is comparative genome analysis. Comparison of two genomes that have diverged just the right distance, so that alignments show functional regions standing out as islands of conservation, reveals regions that are under selective pressure, including ncRNA genes. Most complete genome sequences are too divergent to permit powerful comparative analysis. However, the *Caenorhabditis briggsae* and mouse genomes are slated for sequencing, to enable comparisons to *C. elegans* and human, and many closely related microbial genomes are becoming available. Well-chosen pairs of genomes may be the most fertile hunting grounds for novel ncRNAs.

## Acknowledgements

Research in my group is supported by NIH R01-HG01363, Monsanto, and Paracel. My apologies to those whose work I've been unable to cite due to space limitations.

## References and recommended reading

\*\*1. Gesteland R, Cech T, Atkins J, editors: *The RNA World, Second Edition*, New York: Cold Spring Harbor Laboratory Press, 1999.

The indispensable and recently updated guide to all things RNA.

2. Poole A, Jeffares D, Penny D: **The path from the RNA world**. *J Mol Evol* 1998, **46**:1--17.

\*3. Jeffares D, Poole A, Penny D: **Relics from the RNA world**. *J Mol Evol* 1998, **46**:18--36.

A speculative article that is probably the strongest written argument that modern ncRNAs are fossils of the RNA world. I find the argument fascinating, but I don't accept the notion that billions of years ago, ncRNAs lost a war against proteins for functional supremacy, and that modern ncRNAs are like shell-shocked snipers starving in remote jungle caves. I think many ncRNAs are recent evolutionary innovations.

4. Brosius J: **Transmutation of tRNA over time**. *Nature Genet* 1999, **22**:8--9.

A short speculative paper about a transition from an RNA World to the modern world. Here again, one may disagree with the details, as with any hypothetical evolutionary scenario, but the ideas are interesting.

\*5. Ribas de Pouplana L, Turner RJ, Steer BA, Schimmel P: **Genetic code origins:**

**tRNAs older than their synthetases?** *Proc Natl Acad Sci USA* 1998, **95**:11295--11300.

The RNA World model doesn't generate many testable hypotheses, but here the authors propose one. If modern RNAs descend from the RNA world, phylogenetic analysis might indicate that RNA sequence families diversified before proteins did. This is probably the first such test of the RNA World model. The results suggest that lysine tRNAs evolved before the modern forms of lysyl-tRNA synthetase.

6. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG: **Life with 6000 genes.** *Science* 1996, **274**:546--567.

7. Lowe TM, Eddy SR: **tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence.** *Nucl Acids Res* 1997, **25**:955--964.

This has become a standard tRNA gene detection program in many genome sequence analysis groups.

8. Altman S, Kirsebom L: **Ribonuclease P.** In *The RNA World, Second Edition*. Edited by Gesteland R, Cech T, Atkins J. New York: Cold Spring Harbor Laboratory Press; 1999:351--380.

9. Bovia F, Strub K: **The signal recognition particle and related small cytoplasmic ribonucleoprotein particles.** *J Cell Sci* 1996, **109**:2601--2608.

10. Simpson L: **RNA editing – an evolutionary perspective**. In *The RNA World, Second Edition*. Edited by Gesteland R, Cech T, Atkins J. New York: Cold Spring Harbor Laboratory Press; 1999:585--608.
11. Blackburn E: **Telomerase**. In *The RNA World, Second Edition*. Edited by Gesteland R, Cech T, Atkins J. New York: Cold Spring Harbor Laboratory Press; 1999:609--635.
12. Watanabe Y, Yamamoto M: ***S. pombe* mei2+ encodes and RNA-binding protein essential for premeiotic DNA synthesis and meiosis I, which cooperates with a novel RNA species meiRNA**. *Cell* 1994, **78**:487--498.
13. Zwieb C, Wower I, Wower J: **Comparative sequence analysis of tmRNA**. *Nucl Acids Res* 1999, **27**:2063--2071.
14. Keiler K, Waller P, Sauer R: **Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA**. *Science* 1996, **271**:990--993.
- \*15. Zhang F, Lemieux S, Wu X, St.-Arnaud D, McMurray C, Major F, Anderson D: **Function of hexameric RNA in packaging of bacteriophage  $\phi$ 29 DNA in vitro**. *Mol Cell* 1998, **2**:141--147.

An example of a functional high-order RNA quaternary structure. Stunningly, the pRNA hexamer appears to form a ring structure, which is thought to be a rotating machine essential for packaging phage DNA into the prohead.

\*16. Panning B, Jaenisch R: **RNA and the epigenetic regulation of X chromosome inactivation.** *Cell* 1998, **93**:305--308.

\*17. Willard HF, Salz HK: **Remodelling chromatin with RNA.** *Nature* 1997, **386**:228--229.

18. Lee J, Davidow L, Warshawsky D: **Tsix, a gene antisense to Xist at the X-inactivation centre.** *Nature Genet* 1999, **21**:400--404.

19. Amrein H, Axel R: **Genes expressed in neurons of adult male *Drosophila*.** *Cell* 1997, **88**:459--469.

Discovery of the *Drosophila roX1* and *roX2* ncRNA transcripts in a subtractive hybridization screen.

20. Meller VH, Wu KH, Roman G, Kuroda MI, Davis RL: **roX1 RNA paints the X chromosome of male *Drosophila* and is regulated by the dosage compensation system.** *Cell* 1998, **88**:445--457.

Independent discovery of the *Drosophila roX1* ncRNA transcript by its expression pattern in an enhancer trap line.

21. Delihans N: **Regulation of gene expression by trans-encoded antisense RNAs.** *Mol Microbiol* 1995, **15**:411--414.

A nice review of prokaryotic antisense riboregulatory RNAs.

22. Altuvia S, Zhang A, Argaman L, Tiwari A, Storz G: **The *Escherichia coli* OxyS regulatory RNA represses *fhlA* translation by blocking ribosome binding.**

*EMBO J* 1998, **17**:6069--6075.

23. Zhang A, Altuvia S, Tiwari A, Argaman L, Hengge-Aronis R, Storz G: **The OxyS regulatory RNA represses *rpoS* translation and binds the Hfq (Hf-I) protein.**

*EMBO J* 1998, **17**:6061--6068.

24. Lease R, Cusick M, Belfort M: **Riboregulation in *Escherichia coli*: DsrA RNA acts by RNA:RNA interactions at multiple loci.** *Proc Natl Acad Sci USA* 1998,

**95**:12456--12461.

25. Majdalani N, Cuning C, Sledjeski D, Elliott T, Gottesman S: **DsrA RNA regulates translation of RpoS message by an anti-antisense mechanism, independent of its action as an antisilencer of transcription.** *Proc Natl Acad Sci USA* 1998,

**95**:12462--12467.

26. Moss E, Lee R, Ambros V: **The cold shock domain protein LIN-28 controls developmental timing in *C. elegans* and is regulated by the *lin-4* RNA.** *Cell*

1997, **88**:637--646.

27. Lee R, Feinbaum R, Ambros V: **The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*.** *Cell* 1993,

75:843--854.

28. Wightman B, Ha I, Ruvkun G: **Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*.** *Cell* 1993, **75**:855--862.

29. Brownlee G: **Sequence of 6S RNA of *E. coli*.** *Nature New Biol* 1971, **229**:147--149.

30. N.C.Hogan, Slot F, Traverse K, Garbe J, Bendena W, Pardue ML: **Stability of tandem repeats in the *Drosophila melanogaster* *Hsr-omega* nuclear RNA.** *Genetics* 1995, **139**:1611--1621.

31. Brannan CI, Dees EC, Ingram RS, Tilghman SH: **The product of the H19 gene may function as an RNA.** *Mol Cell Biol* 1990, **10**:28--36.

32. Erdmann V, Szymanski M, Hochberg A, de Groot N, Barciszewski J: **Collection of mRNA-like non-coding RNAs.** *Nucl Acids Res* 1999, **27**:192--195.

The beginnings of a database of polyadenylated, mRNA-like ncRNA transcripts.

\*\*33. Weinstein L, Steitz JA: **Guided tours: From precursor snoRNA to functional snoRNP.** *Curr Opin Cell Biol* 1999, **11**:378--384.

An incisive review of the recent and voluminous snoRNA literature.

34. Smith CM, Steitz JA: **Sno storm in the nucleolus: New roles for myriad small RNPs.** *Cell* 1997, **89**:669--672.

35. Tollervey D, Kiss T: **Function and synthesis of small nucleolar RNAs.** *Curr Opin Cell Biol* 1997, **9**:337--342.

36. Bachellerie JP, Cavaille J: **Guiding ribose methylation of RNA.** *Trends Biochem Sci* 1997, **22**:257--261.

\*37. Lowe TM, Eddy SR: **A computational screen for methylation guide snoRNAs in yeast.** *Science* 1998, **283**:1168--1171.

A computational model was used to screen the yeast genome and discover almost all of the C/D snoRNAs that guide ribosomal RNA that hadn't yet been found.

\*\*38. Tycowski K, Yao ZH, Graham P, Steitz J: **Modification of U6 spliceosomal RNA is guided by other small RNAs.** *Mol Cell* 1998, **2**:629--638.

Definitive evidence that snoRNAs guide modification of RNAs other than rRNA. This paper may open the floodgates for an even larger deluge of snoRNA genes.

39. Cavaille J, Nicoloso M, Bachellerie JP: **Targeted ribose methylation of RNA in vivo directed by tailored antisense RNA genes.** *Nature* 1996, **383**:732--735.

40. Maxwell E, Fournier M: **The small nucleolar RNAs.** *Ann Rev Biochem* 1995,

64:897--934.

A masterful review of the history of the snoRNA field, just before it was realized that almost all known snoRNAs guide rRNA ribose methylations and pseudouridylations.

41. Nicoloso M, Qu LH, Michot B, Bachellerie JP: **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* 1996, **260**:178--195.

\*42. Smith C, Steitz J: **Classification of *gas5* as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5'-terminal oligopyrimidine gene family reveals common features of snoRNA host genes.** *Mol Cell Biol* 1998, **18**:6897--6909.

Besides discovering that *gas5* is another example of an ncRNA functioning as a snoRNA host gene, the authors observe that all the known snoRNA host genes belong to the so-called 5'-TOP gene family. It has long been observed that intronic snoRNAs tend to occur predominantly in introns of ribosomal protein genes and other proteins involved in translation, but the correlation was not absolute. The correlation with the 5'-TOP gene family is stronger. It points to the snoRNA host genes being 'chosen' because they share a special global transcriptional program.

43. Tycowski KT, Shu MD, Steitz JA: **A mammalian gene with introns instead of exons generating stable RNA products.** *Nature* 1996, **379**:464--466.

44. Bortolin ML, Kiss T: **Human U19 intron-encoded snoRNA is processed from a long primary transcript that possesses little potential for protein coding.** *RNA* 1998, **4**:445--454.

45. Pelczar P, Filipowicz W: **The host gene for intronic U17 small nucleolar RNAs in mammals has no protein-coding potential and is a member of the 5'-terminal oligopyrimidine gene family.** *Mol Cell Biol* 1998, **18**:4509--4518.

Independently, these authors also note the correlation of intronic snoRNAs falling only in host genes belonging to the 5'-TOP family.

46. Schneider C, King R, Philipson L: **Genes specifically expressed at growth arrest of mammalian cells.** *Cell* 1988, **54**:787--793.

47. Coccia E, Cicala C, Charlesworth A, Ciccarelli C, Rossi G, Philipson L, Sorrentino V: **Regulation and expression of a growth arrest-specific gene (*gas5*) during growth, differentiation, and development.** *Mol Cell Biol* 1992, **12**:3514--3521.

\*48. Romeo T: **Global regulation by the small RNA-binding protein CsrA and the non-coding RNA molecule CsrB.** *Mol Microbiol* 1998, **29**:1321--1330.

Excellent review of the CsrB story in *E. coli* and *Erwinia carotovora*.

49. Murata H, Chatterjee A, Liu Y, Chatterjee A: **Regulation of the production of extracellular pectinase, cellulase, and protease in the soft rot bacterium**

*Erwinia carotovora* subsp. *carotovora*: Evidence that aepH of *E. carotovora* subsp. *carotovora* 71 activates gene expression in *E. carotovora* subsp. *carotovora*, *E. carotovora* subsp. *atroseptica*, and *Escherichia coli*. *Appl Environ Microbiol* 1994, **60**:3150--3159.

50. Romeo T, Gong M, Liu M, Brun-Zinkernagel AM: **Identification and molecular characterization of csrA, a pleiotropic gene from *Escherichia coli* that affects glycogen biosynthesis, gluconeogenesis, cell size, and surface properties.** *J Bacteriol* 1993, **175**:4744--4755.

51. Liu M, Gui G, Wei B, III JP, Oakford L, Yüksel U, Giedroc D, Romeo T: **The RNA molecule CsrB binds to the global regulatory protein CsrA and antagonizes its activity in *Escherichia coli*.** *J Biol Chem* 1997, **272**:17502--17510.

52. Dandekar T, Hentze MW: **Finding the hairpin in the haystack: Searching for RNA motifs.** *Trends Genet* 1995, **11**:45--50.

A review of methods for identifying homologous RNA structures in sequence databases, focusing primarily on pattern-based searches.

53. Dsouza M, Larsen N, Overbeek R: **Searching for patterns in genomic data.** *Trends Genet* 1997, **13**:497--498.

Describes a program called PATSCAN for searching a database with a pattern that can include base-pairing constraints.

54. Laferrière A, Gautheret D, Cedergren R: **An RNA pattern matching program with enhanced performance and portability.** *Comput Applic Biosci* 1994, **10**:211--212.

Describes an improved version of the widely used RNAMOT program, which (like PATSCAN) can search a sequence database with a descriptor of an RNA structure consensus.

\*55. Durbin R, Eddy SR, Krogh A, Mitchison GJ: *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge: Cambridge University Press, 1998.

A technical introduction to probabilistic modeling methods for computational sequence analysis. Two chapters are devoted to RNA structure analysis using stochastic context-free grammar methods, including database search applications.

56. Chen JH, Le SY, Shapiro B, Currey KM, Maizel J: **A computational procedure for assessing the significance of RNA secondary structure.** *Comput Applic Biosci* 1990, **6**:7--18.

57. Le SY, Chen JH, Maizel JV: **Efficient searches for unusual folding regions in RNA sequences.** In *Structure and Methods: Human Genome Initiative and DNA Recombination*. Edited by Sarma RH, Sarma MH. New York: Adenine Press; volume 1, 1990:127--136.

\*\*58. Olivas WM, Muhlrاد D, Parker R: **Analysis of the yeast genome:**

**Identification of new non-coding and small ORF-containing RNAs.** *Nucl Acids*

*Res* 1997, **25**:4619--4625.

A landmark paper in the sparse ncRNA genefinding literature. The authors take two different approaches to discovering new ncRNA genes in the complete yeast genome.

59. Daniels DL, Plunkett G, Burland V, Blattner FR: **Analysis of the *Escherichia coli***

**genome: DNA sequence of the region from 84.5 to 86.5 minutes.** *Science* 1992,

**257**:771--778.

60. Singer MS, Gottschling DE: **TLC1: Template RNA component of**

***Saccharomyces cerevisiae* telomerase.** *Science* 1994, **266**:404--409.