# Fewer Genes, More Noncoding RNA

Jean-Michel Claverie

**Recent studies showing that most "messenger" RNAs do not encode proteins finally explain the long-standing discrepancy between the small number of protein-coding genes found in vertebrate genomes and the much larger and ever-increasing number of polyadenylated transcripts identified by tag-sampling or microarray-based methods. Exploring the role and diversity of these numerous noncoding RNAs now constitutes a main challenge in transcription research.**

A few months before the publication of the first drafts of the human genome sequence (*1*, *2*), online bids predicting the number of human protein-coding genes ranged from 30,000 to 150,000 [see (*3*)]. To the surprise of many (*4*), initial bioinformatic analyses revealed no more than 35,000 human genes, an estimate that has steadily declined to the present 25,000 genes (*5*). On the other hand, the largest estimates based on the number of distinct polyadenylated transcript 3′-ends identified through the single-pass sequencing of cDNA libraries (*6*) [i.e., expressed sequence tags (ESTs)] have not followed a diminishing trend. On the contrary, more transcripts keep being discovered, many of which do not correspond to annotated genes [e.g., (*7*)], in particular when using the serial analysis of gene expression (SAGE) approach (*8*).

Over the last 5 years, this discrepancy (*4*) between the number of recognized protein-coding genes and the apparent number of transcripts has not been reduced. As early as 1997, the then-thriving genomics industry had already sequenced several million ESTs and had come up with estimates of well over 100,000 human genes. For example, Incyte Genomics estimated 140,000 genes by grouping overlapping EST sequences [cited in (*9*)]; this total did not include more than 200,000 EST sequences seen only once.

Comparable numbers emerged a few years later in the public domain. The Human Gene Index of the Institute for Genomic Research predicted in excess of 75,000 human genes (*10*), whereas the Unigene database of the National Center for Biotechnology Information indicated 84,000 genes (*6*). These sequences are still in the databases, awaiting reconciliation with the much smaller number of human genes identified by the direct analysis of the human genome sequence.

Recent results may put an end to the paradox, albeit in a rather unexpected manner: A large fraction of the human (vertebrate) genome appears to give rise to polyadenylated transcripts that do not code for proteins. The notion of noncoding RNAs is not new—for example, the 17-kb X chromosome–inactivated specific transcript (Xist) was discovered in 1991 (*11*). However, it is only recently that the sheer scale of the phenomenon has begun to be realized. Unfortunately, initial analyses of the transcriptome were based on hybridization with probes derived from predefined or predicted gene sequences, and thus they did not reveal unexpected transcripts. A vastly different picture of transcriptional activity emerged as soon as tiling arrays were introduced, allowing the interrogation of genome sequences for corresponding transcripts at fixed intervals irrespective of predicted gene locations. For instance, a tiling array with 5-nucleotide resolution that mapped transcription activity along 10 human chromosomes revealed that an average of 10% of the genome (compared to the 1 to 2% represented by bona fide exons) corresponds to polyadenylated transcripts, of which more than half do not overlap with known gene locations (*12*).

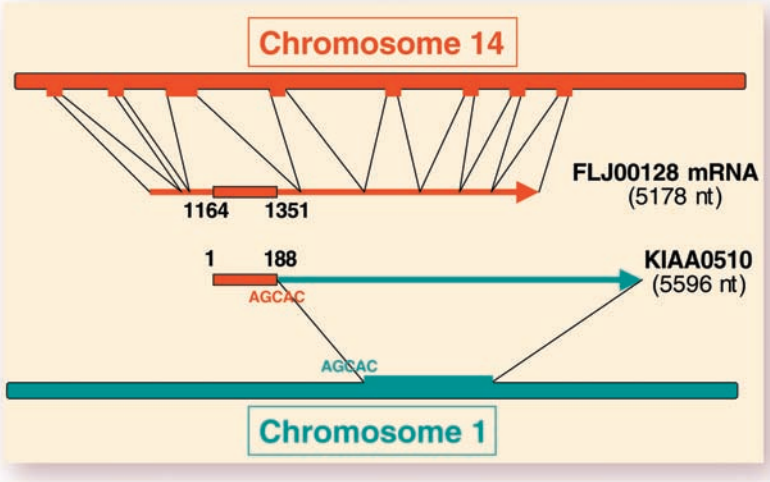


**Fig. 1.** Relationship between the KIAA0510 cDNA sequence and a FLJ00128 protein-encoding transcript. The FLJ00128 cDNA (GenBank identification number 18676462) looks like a standard transcript with more than 20 exons (not drawn), all mapping to human chromosome 14. This transcript encodes a large protein of more than 1500 residues without known or predicted function. The KIAA0510 cDNA sequence (GenBank identification number 3413954) corresponds to a single exon, mapping on chromosome 1 and devoid of significant open reading frames. The 3′ noncoding part of this cDNA is fused to a 188-nucleotide sequence (boxed) 100% identical to a sequence unique to chromosome 14 and encoding 62 residues of protein FLJ00128. This region does not match the boundaries of an exon (as would be expected for trans-splicing) in the gene encoding FLJ00128. Both transcript sequences were assembled from multiple independently isolated ESTs and are devoid of low-complexity regions or repeats. Thus, they cannot easily be dismissed as cloning or sequencing artifacts.

Recent data from the FANTOM 3 project (*13*, *14*) confirm and amplify these findings. Through a technical tour de force, the members of this consortium have established that a staggering 62% of the mouse genome is transcribed. They have identified more than 181,000 independent transcripts, of which half consist of noncoding RNA. Moreover, they found that more than 70% of the mapped transcription units overlap to some extent with a transcript from the opposite strand (*13*, *14*).

These results provide a solution to the discrepancy between the number of (protein-coding) genes and the number of transcripts—noncoding polyadenylated mRNA contributes to a large fraction of the 3′-EST sequences (and SAGE tags) subsequently clustered or remaining as singletons. Indeed, the noncoding Xist mRNA is abundantly represented in all EST projects. It is thus likely that sequences of noncoding transcripts have been accumulating

Structural and Genomics Information Laboratory, CNRS UPR 2589, Institut de Biologie Structurale et Microbiologie, 31 chemin Joseph Aiguier, Marseille 13402, France, and University of Méditerranée School of Medicine, Marseille 13385, France. E-mail: jean-michel.claverie@igs.cnrs-mrs.fr

in EST databases and have for the most part (including singleton and antisense ESTs) been erroneously interpreted as coming from the 3′-untranslated regions of protein-coding transcripts. Noncoding transcripts originating from intergenic regions, introns, or antisense strands have probably been right before our eyes for 8 years without having been discovered!

The notion that transcription is limited to protein-coding genes is also being challenged in microbial systems. For *Escherichia coli*, the first analysis with a genome tiling microarray revealed a substantial number of antisense and intergenic transcripts (*15*). Noncoding short-lived "cryptic" mRNAs have also recently been seen in yeast, the transcription of which may maintain chromatin in an open state (*16*). The consequences of certain RNA polymerase II mutations for the status of pericentromeric heterochromatin also suggest a direct coupling between the transcription of noncoding RNAs and chromatin structure (*17*).

The intergenic, intronic, and antisense transcribed sequences that were once deemed artifactual are now a testimony to our collective refusal to depart from an oversimplified gene model. But what if transcription is even more complex? Could it, for instance, lead to mRNAs generated from two different chromosomes (Fig. 1)? A year ago, we would have immediately suspected such sequences as further artifacts arising from large-scale cDNA sequencing programs. But now? Perhaps it's time to go back to the cDNA sequence databases and reevaluate the numerous unexpected objects they contain (*18*). Transcription will never be simple again, but how complex will it get?

### References and Notes

1. E. S. Lander *et al.*, *Nature* **409**, 860 (2001).
2. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
3. Editorial, *Nat. Genet.* **25**, 127 (2000).
4. J.-M. Claverie, *Science* **291**, 1255 (2001).
5. International Human Genome Sequencing Consortium, *Nature* **431**, 931 (2004).
6. D. L. Wheeler *et al.*, *Nucleic Acids Res.* **29**, 11 (2001).
7. E. E. Schadt *et al.*, *Genome Biol.* **5**, R73 (2004).
8. K. R. Boheler, M. D. Stern, *Trends Biotechnol.* **21**, 55 (2003).
9. D. B. Davison, J. F. Burke, *IBM J. Res. Dev.* **45**, 439 (2001).
10. F. Liang *et al.*, *Nucleic Acids Res.* **28**, 3657 (2000).
11. C. J. Brown *et al.*, *Nature* **349**, 38 (1991).
12. J. Cheng *et al.*, *Science* **308**, 1149 (2005); published online 24 March 2005 (10.1126/science.1108625).
13. FANTOM Consortium and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group), *Science* **309**, 1559 (2005).
14. RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) and FANTOM Consortium, *Science* **309**, 1564 (2005).
15. D. W. Selinger *et al.*, *Nat. Biotechnol.* **18**, 1262 (2000).
16. F. Wyers *et al.*, *Cell* **121**, 725 (2005).
17. H. Kato *et al.*, *Science* **309**, 467 (2005); published online 9 June 2005 (10.1126/science.1114955).
18. J. Shendure, G. M. Church, *Genome Biol.* **3**, research0044.1 (2002).
19. The Structural and Genomics Information Laboratory is supported by CNRS and the Marseille-Nice Genopole. I thank N. Baeza for drawing my attention to the KIAA0510 transcript.

10.1126/science.1116800

VIEWPOINT

# Capping by Branching: A New Ribozyme Makes Tiny Lariats

**Anna Marie Pyle**

**The number of naturally occurring RNA enzymes has just been expanded by the discovery of a new branching ribozyme. But this ribozyme has unexpected relatives: group I introns.**

Before RNA molecules are ready for action, they usually undergo splicing, whereby the noncoding sequences (introns) are removed from the coding sequences (exons) and the latter are stitched back together. In some eukaryotes and prokaryotes, two classes of specialized introns (known as group I and group II introns) fold into catalytic structures that promote their own removal from flanking exons (a process called self-splicing) (*1*). Group II introns first caught the attention of an observant investigator because of their exceptional stability. During electron microscopy studies of yeast mitochondrial RNA, Arnberg *et al.* noticed abundant RNA circles (*2*) that were later shown to result from the self-splicing activity of group II introns (*3*). These introns catalyze a branching reaction in which an unpaired adenosine within the intron uses one of its sugar groups (the 2′-hydroxyl) to join with the intron terminus, thereby freeing the adjacent exon (coding region) and creating a circular "lariat" form of the intron (Fig. 1A). Once liberated, the intron lariats can function as parasitic RNAs, or mobile genetic elements, that migrate within a genome or into new genomes (*4*).

Exceptional stability is obviously a useful trait for infectious RNAs that have their own agendas in a cell (like group II introns). However, stability is important for many other RNAs, including mRNAs that encode proteins and RNAs involved in cellular function. Most mRNAs are capped at their upstream (5′) terminus with a modified guanosine residue that protects the mRNA against predation by the abundant 5′-exonucleases that prowl the cell, ready to pounce on unprotected, linear RNA strands (*5*). In this issue, Nielsen *et al.* (*6*) report that an mRNA (called I-*Dir* I) encoded by a homing endonuclease gene (HEG) from the slime mold *Didymium iridis* is capped by a different mechanism that takes a page from the group II intron playbook. The upstream terminus of the mature I-*Dir* I mRNA is a tiny circle that results from a branching reaction in which a 2′-hydroxyl group near the beginning of the mRNA reacts with a nearby phosphodiester linkage, thereby creating a circular cap and liberating the upstream RNA (Fig. 1B). Remarkably, the branching reaction is catalyzed not by a group II intron but by an unrelated group I intron–like ribozyme that is upstream from the branch site.

This group I–like ribozyme had long been known to catalyze cleavage at its junction with the I-*Dir* I mRNA (*7*), but careful primer extension analysis of the 5′ end of the transcript revealed that the mRNA cleavage product contained an unusual RNA structure. Classic chemical and enzymatic analysis of this terminal structure, together with studies on the reversibility of the cleavage reaction, strongly suggested the existence of a tiny lariat at the mRNA terminus. Importantly, specific deletion of the 2′-hydroxyl group at the putative branch site (i.e., the nucleophile in the branching reaction) eliminates this reaction.

## Implications of Branching RNAs

By finding that a group I–like ribozyme can catalyze a branching reaction, Nielsen and colleagues provide strong evidence that branching may be a common activity that is shared by many different nucleic acid molecules, regardless of their evolutionary heritage. This suggests that branching is a facile process and that branching ribozymes may have evolved independently on

Department of Molecular Biophysics and Biochemistry, Howard Hughes Medical Institute, Yale University, 266 Whitney Avenue, Bass Building Room 334, New Haven, CT 06520, USA.

**Fewer Genes, More Noncoding RNA**

Jean-Michel Claverie (September 1, 2005)

*Science* **309** (5740), 1529-1530. [doi: 10.1126/science.1116800]