The Functional Genomics of Noncoding RNA

John S. Mattick

Large numbers of noncoding RNA transcripts (ncRNAs) are being revealed by complementary DNA cloning and genome tiling array studies in animals. The big and as yet largely unanswered question is whether these transcripts are relevant. A paper by Willingham *et al.* shows the way forward by developing a strategy for large-scale functional screening of ncRNAs, involving small interfering RNA knockdowns in cell-based screens, which identified a previously unidentified ncRNA repressor of the transcription factor NFAT. It appears likely that ncRNAs constitute a critical hidden layer of gene regulation in complex organisms, the understanding of which requires new approaches in functional genomics.

Recent large-scale studies of the human and mouse transcriptomes have used both cDNA cloning approaches (*1*–*3*) and the interrogation of genome tiling arrays (*4*–*6*). The surprising but consistent finding of these studies has been that a huge number of observed transcripts—about half of the total—do not appear to encode proteins. Many of these transcripts appear to be developmentally regulated (*1*, *4*), and similar findings have been reported in *Drosophila* (*7*). The big and as yet largely unanswered question is whether these noncoding RNAs (ncRNAs) are meaningful or simply represent "transcriptional noise" (Fig. 1). A study by Schultz, Hogenesch, and colleagues (*8*) begins to answer this question by developing a strategy for large-scale functional screening of ncRNAs.

Willingham *et al.* (*8*) selected 512 ncRNA sequences from the RIKEN Fantom2 mouse cDNA collection (*1*, *9*) that showed significant conservation with human genomic sequences and constructed small interfering RNAs (siRNAs) (two each, expressed as short hairpin RNAs) against the human orthologs of these sequences. These siRNAs were then used to interrogate a battery of 12 cell-based assays representing key cellular processes and signaling pathways, with the use of reporter assays in microtiter plates (*10*). They identified eight functional ncRNAs: six essential for cell viability, one repressor of Hedgehog signaling, and one (termed NRON) that acts as a repressor of the transcription factor NFAT, which is itself required for T-cell receptor–mediated immune response and the development of the heart, vasculature, musculature, and nervous tissue.

Detailed analysis of NRON showed that this ncRNA, which has two blocks of near-perfect conservation between humans and mice but no substantial open reading frame, is enriched in placenta, muscle, and lymphoid tissues and exhibits a distinct tissue-specific distribution of splice variants, suggesting subtle but biologically relevant differences in its function in different tissues (*8*). By tagging NRON with an RNA hairpin that is bound by the MS2 phage protein, followed by affinity chromatography of whole-cell extracts, the authors showed that NRON interacts directly or indirectly with 11 proteins, including three members of the importin-beta superfamily, which mediate the nucleocytoplasmic transport of cargoes such as NFAT. siRNA knockdown of four of these proteins (including importin–beta 1) activated NFAT activity, whereas overexpression of these proteins repressed NFAT activity, as did siRNAs directed against NRON. Moreover, binding and ribonuclease protection experiments supported a direct association of NRON with importin–beta 1, which itself is known to associate with some of the other proteins that were identified as interacting with NRON (*8*).

These observations suggest that NRON may act as a modulator of NFAT nuclear trafficking, probably by regulating its subcellular location, a conclusion supported by the observation that NFAT nuclear localization is increased when the level of NRON is reduced by siRNA (*8*). The broader conclusion is that NRON may act as a scaffold for the assembly of protein complexes that regulate nuclear trafficking of this and probably other important transcription factors, opening a new dimension of organizational control in cell biology and development.

This elegant study not only points the way ahead but also illustrates the magnitude of the task that is in front of us, which may be an equal or greater challenge than that we already face in working out the biochemical function and biological role of all of the known and predicted proteins and their isoforms. The cDNA and genome tiling array studies have indicated not only that there are tens of thousands of ncRNA transcripts (both polyadenylated and nonpolyadenylated) expressed from the mammalian genome in different cells and tissues but also that these transcripts comprise a complex interlaced and overlapping network from both strands, whereby even a single nucleotide may be part of multiple differently processed transcripts (*2*, *3*, *6*, *11*, *12*).

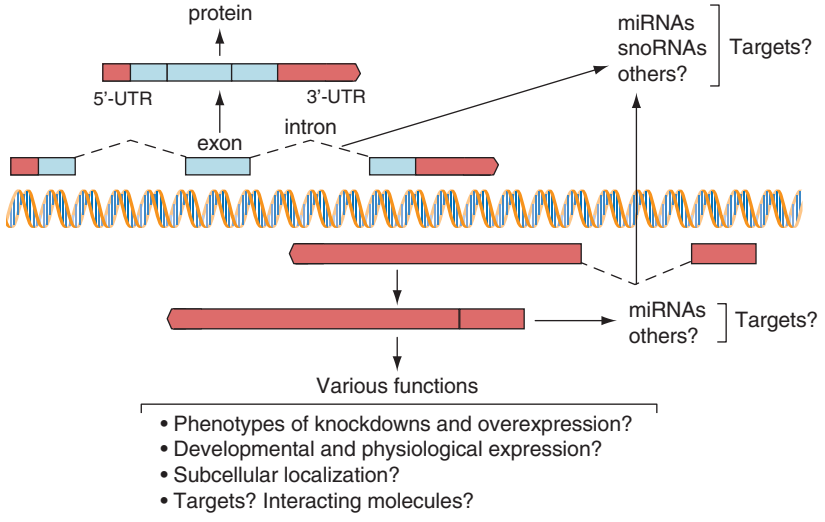

Fig. 1. The complexity of transcription of protein-coding (blue) and noncoding (red) RNA sequences. Transcripts may be derived from either or both strands, and they may be overlapping and interlaced (*2*, *3*, *6*, *11*, *12*). Many transcripts (including some noncoding transcripts) are alternatively spliced. Both exons and introns may transmit information. Many miRNAs and all small nucleolar RNAs in animals are sourced from introns [see (*13*) for a review]. The range of types and functions of noncoding RNAs is unknown.

Australian Research Council Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4072, Australia. E-mail: j.mattick@imb.uq.edu.au

Ascribing function to these ncRNAs will not be simple, nor occur quickly, given that this will require in vivo and in vitro assays, the interpretation of which will be compromised by ambiguity in the former (for example in discriminating between mutations that affect cis-acting regulatory sequences and those that affect functional trans-acting RNAs) and in both cases by the ability to detect a phenotype when the expression of targeted ncRNA sequences is altered by siRNA-mediated knockdown or ectopic expression. Only 8 of 512 ncRNAs showed function in the assays undertaken by Willingham *et al*. (*8*), although this is not a bad rate of return given the limited scope of these assays. Nonetheless, these initial findings will have a big impact, because they reveal the involvement of hitherto unsuspected ncRNAs in already intensively studied pathways such as Hedgehog signaling and nuclear trafficking. Notably, genome tiling array studies have also revealed unknown transcript and splice variants of *sonic hedgehog* (*11*), indicating just how much remains to be done.

The selection of phenotypic assays may be guided by other studies, such as the analysis of the patterns of expression and the subcellular location of the ncRNAs under analysis, as is already routinely done for proteins with unknown functions. Indeed, most would regard tissue-specific expression as a reasonable prima facie indicator of function. On the other hand, faced with the uncomfortable implications of large numbers of such RNAs and the evidence that many are expressed only at low levels, others may suggest that these RNAs are merely transcriptional noise from illegitimate promoters, which may be variable in different cells, because of, for example, different chromatin architectures, although it also seems likely that chromatin architecture is itself controlled by RNA signaling (*13*, *14*).

Notably, evolutionary conservation may not be a reliable signature of functional ncRNAs. The ncRNAs selected by Willingham *et al*. were those that were most highly conserved between humans and mouse, a reasonable filter given that conservation is normally a good indicator of function. However, the reverse—i.e., that lack of conservation indicates lack of function—is not necessarily true. Sequence conservation is normally mandated by the preservation of structure-function relationships (as in proteins) and/or multilateral interactions (as in ribosomal RNA). If many of these newly discovered ncRNAs are regulatory, as one might reasonably suppose them to be, they may have quite different evolutionary constraints. Many microRNAs (miRNAs)—small 20- to 25-nucleotide RNAs that control many aspects of plant and animal development by sequence-specific interactions with other RNAs—are highly conserved (and have been mainly identified on this basis), but these appear to be central regulators that have many targets (making covariation difficult) and there are likely to be many more that are not so constrained (*13*).

This possibility is supported by a recent study that did not require substantial evolutionary conservation and (thereby) identified many new human miRNAs, a significant number of which appear to be primate specific (*15*). The number of known human miRNAs stands at well over 1500 and is rising rapidly (*13*, *15*, *16*). Sensitive genetic screens in *Caenorhabditis elegans* have also identified rare miRNAs with limited evolutionary conservation such as *lys-6*, which is required for left-right neuronal patterning, suggesting that many more remain to be found (*17*). Moreover, a number of well-studied ncRNAs are poorly conserved, such as XIST, which controls X-chromosome inactivation in mammals, and Air, a ncRNA of over 100 kb that is involved in imprinting of the *Igf2r* locus in mouse (*18*, *19*). All of these considerations suggest that many ncRNAs are evolving quickly (by drift under mild negative selection or under positive selection for the rewiring of regulatory circuitry in phenotypic radiation) and that those that have been identified (or prioritized for study) on the basis of evolutionary conservation are probably just the tip of a very large iceberg. Nonetheless, there is considerable scope for using more sophisticated bioinformatic approaches, including intragenomic sequence matching.

It is also clear that the majority of the genomes of animals is indeed transcribed (*12*), which suggests that these genomes are either replete with largely useless transcription or that these noncoding RNA sequences are fulfilling a wide range of unexpected functions in eukaryotic biology. These sequences include introns (Fig. 1), which account for at least 30% of the human genome but have been largely overlooked because they have been assumed to be simply degraded after splicing. However, it has been shown that many miRNAs and all known small nucleolar RNAs in animals are sourced from introns (of both protein-coding and noncoding transcripts) (*13*), and it is simply not known what proportion of the transcribed introns are subsequently processed into smaller functional RNAs. It is possible, and logically plausible, that these sequences are also a major source of regulatory RNAs in complex organisms (*20*).

The studies of Willingham *et al*. and others that have begun to explore the underworld of RNA in eukaryotes raise more questions than they answer. That complex organisms have complex genetic programming should come as no surprise. That much of this programming may be transacted by noncoding RNAs may be. However, given the sheer extent of noncoding RNA transcription, it seems more and more likely that a large portion of the human genome may be functional by means of RNA. This also means that we may have seriously misunderstood the nature of genetic programming in the higher organisms (*21*) by assuming that most genetic information is expressed as and transacted by proteins, as it largely is in prokaryotes (*22*). If so, there is a long road ahead in functional genomics.

References and Notes

1. Y. Okazaki *et al.*, *Nature* **420**, 563 (2002).
2. FANTOM Consortium and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group), *Science* **309**, 1559 (2005).
3. RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) and FANTOM Consortium, *Science* **309**, 1564 (2005).
4. S. Cawley *et al.*, *Cell* **116**, 499 (2004).
5. P. Bertone *et al.*, *Science* **306**, 2242 (2004).
6. J. Cheng *et al.*, *Science* **308**, 1149 (2005).
7. V. Stolc *et al.*, *Science* **306**, 655 (2004).
8. A. T. Willingham *et al.*, *Science* **309**, 1570 (2005).
9. K. Numata *et al.*, *Genome Res.* **13**, 1301 (2003).
10. J. B. Hogenesch, P. G. Schultz, personal communication.
11. P. Kapranov *et al.*, *Genome Res.* **15**, 987 (2005).
12. M. C. Frith, M. Pheasant, J. S. Mattick, *Eur. J. Hum. Genet.* **13**, 894 (2005).
13. J. S. Mattick, I. V. Makunin, *Hum. Mol. Genet.* **14**, R121 (2005).
14. A. H. Ting, K. E. Schuebel, J. G. Herman, S. B. Baylin, *Nat. Genet.* **37**, 906 (2005).
15. I. Bentwich *et al.*, *Nat. Genet.* **37**, 766 (2005).
16. P. D. Zamore, B. Haley, *Science* **309**, 1519 (2005).
17. R. J. Johnston, O. Hobert, *Nature* **426**, 845 (2003).
18. C. Chureau *et al.*, *Genome Res.* **12**, 894 (2002).
19. C. B. Oudejans *et al.*, *Genomics* **73**, 331 (2001).
20. J. S. Mattick, *Curr. Opin. Genet. Dev.* **4**, 823 (1994).
21. J. S. Mattick, *Nat. Rev. Genet.* **5**, 316 (2004).
22. J.-M. Claverie, *Science* **309**, 1529 (2005).
23. I am grateful for the support of the Australian Research Council, the Queensland State Government, and the University of Queensland.

10.1126/science.1117806