

# THE PRODUCTION LINE

If more than 90% of the genome is 'junk' then why do cells make so much RNA from it? **Anna Petherick** goes in search of some answers.

**H**OTAIR is a molecule with a future. Created from a DNA sequence on human chromosome 12, it affects genes on chromosome 2, apparently working as part of the system that enables skin cells to tell where on the body's surface they are, and thus what they should be doing.

Beyond these specifics, HOTAIR may also serve as a model for understanding a whole slew of similar molecules, the existence of which was not even dreamed of ten years ago and the function of which — if any — is still hotly debated. HOTAIR stands out because it is a long piece of RNA that doesn't encode a protein but still does something biologically important<sup>1</sup>. "HOTAIR was a gem in a sea [of long RNAs]," says John Rinn, a genome biologist who discovered the RNA while working at Stanford University in California. "It told us little about what the bulk of these things are doing. For that, we can't even see a common trend."

It is hard to comprehend the upheaval that RNA has been causing in molecular biology over the past few years. Once viewed as a passive intermediary, it was thought to faithfully carry genetic messages from the DNA sequence to the protein-making machinery, where things were made that actually got things done. Biologists were comfortable in the knowledge that only 1–2% of the human genome made protein-coding RNA in this way, and most of the rest was filler. So when, in 2005, geneticist Thomas Gingeras announced that some cells churn out RNA molecules from about 80% of their DNA, he astonished scientists attending the Biology of Genomes meeting at Cold Spring Harbor Laboratory in New York. Why should cells bother with so much manufacturing if, as it seemed, such a tiny fraction was involved in the important business of protein-making?

Over the past three years or so the case for this 'pervasive transcription' has strengthened. The phenomenon has now been ascribed to mice, fruitflies, nematode worms and yeast. These studies, and Gingeras's original reports, came from microarrays — a technology that relies on the tendency of nucleic acids to find their complementary cousins in a solution. Gingeras works for the microarray

manufacturer Affymetrix in Santa Clara, California. But not everyone has been persuaded of the extent of pervasive transcription, in part because microarrays are subject to background 'noise'. Even using no RNA, control chips will give off some signals, and results can be a matter of interpretation.

For anyone who still doubts that the genomes of nucleated organisms are first and foremost RNA machines rather than protein-coding ones, sequence data are starting to provide "ultimate information", Gingeras says. There is something about the nitty gritty of nucleotide sequences that is enticingly reassuring to molecular biologists. New sequencing machines that can stream out data many times faster than their predecessors have made the mass sequencing of cellular transcripts possible.

In 2008, this process was completed for two species of yeast<sup>2,3</sup> using machines made by Illumina, based in San Diego, California. The results broadly agree with the microarray findings, showing transcription from 74%

of the genome of brewer's yeast (*Saccharomyces cerevisiae*) and 90% from that of fission yeast (*Schizosaccharomyces pombe*). Gingeras and other researchers are now working to sequence all the RNA produced by 44 kinds of human cell as part of the Encyclopedia of DNA Elements (ENCODE) project, which aims to identify all the functional parts of the human genome. At that point, any remaining sceptics will be able to overlay the many thousands of different human RNAs onto DNA regions from whence they came. At the end of this process, the covered regions will be those that give rise to RNA — and the uncovered ones, probably just a few naked holes.

All this transcriptional accounting has hastened an already heady RNA rush. Even before the pervasive nature of transcription became clearer, molecular biologists had begun to trot out new classes of RNA molecules that are responsible for important happenings in cells. Thrust farthest into the limelight are the microRNAs (miRNAs), which stop the production of certain proteins, but they have been joined

by a growing number of other RNA families, such as small nucleolar RNAs (snoRNAs) and Piwi-interacting RNAs (piRNAs), with vital roles in cellular and developmental processes — vital enough to earn the DNA that encodes them the label 'RNA genes'.

## The long and the short of it

On the whole, the established classes of RNAs are short molecules, around 20 or 30 nucleotides in length. The non-coding RNAs that Rinn has been championing run to 200 or even 10,000 bases apiece. The issue at the moment is whether, among this bounty of long RNAs, researchers will find anything as biologically meaningful as the shorter RNAs have proved to be. HOTAIR shows that some such molecules have function — but is it the exception or the rule? "It's controversial whether these are mostly just noise or regulatory function," says Jürg Bähler of the Wellcome Trust Sanger Institute in Cambridge, UK, who led one of the yeast RNA sequencing projects.

Those who doubt the importance of RNA bemoan their logical problem: it is impossible to prove lack of function. Even when an important cellular job does get pinned on a long RNA, as it did for HOTAIR, the doubters worry that it is too tempting to extrapolate across the board.

**"Many transcripts are made that we don't understand. We still don't know what those transcripts do, if anything."**

— Ewan Birney

by a growing number of other RNA families, such as small nucleolar RNAs (snoRNAs) and Piwi-interacting RNAs (piRNAs), with vital roles in cellular and developmental processes — vital enough to earn the DNA that encodes them the label 'RNA genes'.



John Rinn may have found a new class of long RNA genes.

K. MAR

Ewan Birney, a bioinformatician at the European Bioinformatics Institute and one of the leading scientists in ENCODE, says that the debate now is about what proportion of long RNAs serve a purpose. "I used to be a much stronger sceptic three to four years ago," he says. "Now I'm accepting that transcription is pretty complicated and that many transcripts are made that we don't understand. Where I still have some scepticism — what we still don't know — is what those transcripts do, if anything."

John Mattick, the director of the Centre for Molecular Biology and Biotechnology at the University of Queensland in Brisbane, Australia, has no such qualms. He is a long-time advocate of non-coding RNA's importance. The doubters, he says, "keep regressing to the most orthodox explanation [that the long RNAs are junk]. But they can't just sit on their intellectual backsides and tell us to prove it." But prove it is just what researchers are starting to do, with a growing number of examples that showcase these molecules' capabilities.

The idea of long non-coding RNAs is not new. Xist, the most famous example, was discovered in 1991. Its 17,000 nucleotides can be found in almost every cell of mice and humans, where it obviates gene expression along an entire X chromosome. Because females have two Xs to their male (XY) counterparts' one, they use Xist to switch off the extra X and compensate for the disparity.

### Varied roles

Xist RNA is transcribed from the chromosome it mutes, and coats it along its length. No one really knows exactly how it attaches and what makes it so effective at gene silencing. What is clear, however, is that part of the molecule attracts chromatin remodelling complexes — enzymes that turn genes on and off by tinkering with DNA's packaging. Get enough of these complexes together, and it seems that you can turn off a whole chromosome.

Over the past few years, the RNA field has compiled a brief list of other long non-coding RNAs. Many of those that have been studied control the activity of protein-coding genes. As the pace of these discoveries has picked up, they have revealed that long RNAs can control genes in a surprising variety of ways, from both near and far, and that their function is not necessarily dependent on the exact sequence of the RNA, as it is when RNA is coding for proteins. This suggests that scientists have only begun to appreciate what RNA is capable of.

In one example published last year, molecular biologist Igor Martianov and his colleagues at the University of Oxford, UK, studied the human gene for dihydrofolate reductase, an enzyme involved in biochemical syntheses that has two 'on' switches for protein production. They discovered that the first of these switches actually triggers the manufacture of a 583-nucleotide-long RNA molecule, and that this RNA directly interferes with the second switch. When this happens, the enzyme is no longer made<sup>4</sup>.

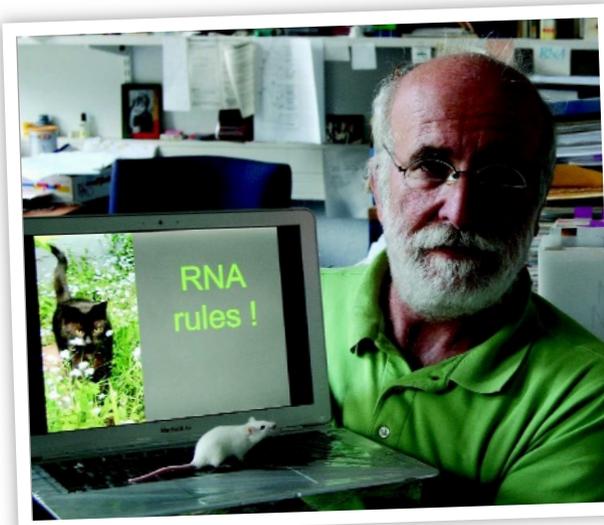
Working in a very different way, a long RNA called NRON seems to travel to the cytoplasm in order to influence the expression of protein-coding genes. Several thousand nucleotides long, NRON polices the trafficking of a transcription factor from the cytoplasm into the nucleus of the cells where it is active<sup>5</sup>. By doing so, it seems to control the transcription factor's activities, which include regulating T cells' immune response.

When Rinn discovered HOTAIR, it reinforced the idea that RNAs could be shuttling around the genome doing important jobs. Rinn was studying skin-cell lines cultured from the finger, foot, foreskin and eight other sites on the human body, trying to find out how these cells know their position.

HOTAIR, which stretches for nearly 2,200 nucleotides, is produced from within a cluster of the *HOX* genes that specify an early embryo's head end and foot end, as well as the order of the body segments in between. When Rinn found that this RNA affects the output of genes on chromosome 2, it was the first time such a cross-chromosome influence had been found. When he lowered levels of the RNA molecule, the activity of *HOX* genes on chromosome 2 jumped, and foreskin cells started behaving in an unusual way<sup>1</sup>.

Rinn initially wanted to name the molecule STAR1. The acronym for 'Suz-Twelve Associated RNA' refers to the enzyme that ferries this molecule from one chromosome to another, and the number one reflected Rinn's optimism that there are likely to be more STARS. But Rinn's lab partner, Howard Chang, wanted a "more humbling" name, Rinn says, and they settled on HOTAIR instead (for *HOX* antisense intergenic RNA). "Howard was right, but I think we are still both in search of more stars, not hot air," says Rinn, now at the Broad Institute in Cambridge, Massachusetts.

As Rinn has said, there is a vast sea of



Jürgen Brosius: an advocate of non-coding RNA.

long RNAs out there. The ones with functions already ascribed to them comprise just a minuscule fraction, and those seem to be regulating genes by very diverse means. To many, this lack of common function infers that science has only scratched the surface of the diversity of long RNAs. The massive scale on which transcription is taking place could be the least of biologists' problems compared with its mind-boggling functional complexity. What is needed, researchers say, is more data to show that RNAs do something useful on the genomic scale — but those data are proving remarkably difficult to collect.

One problem, when it comes to surveying RNA's usefulness, is that sequence does not provide any simple indicator of function. The sequence of non-coding RNA is not conserved between species in the same way that it is for protein-coding genes. If a sequence is doing something important for an organism because of the protein it codes for, then evolution is likely to have kept that region more constant across related species compared with any average stretch. But the same isn't true of RNA, which does not necessarily pair up with a complementary nucleotide sequence at all. Xist is not conserved in this way, nor are any of the other non-coding RNA stars along their full lengths.

Another way to seek evidence of function en masse is to get rid of long non-coding RNAs and watch how animals cope. But such an experiment may produce only subtle changes in an organism as a whole, and could still miss the importance of a transcript. "I think the cell will use these transcripts at very different times and in very different cell types and conditions," Gingeras says. "You may need to see them in a very specific context to see the function."

That is what Jürgen Brosius of the University of Münster, Germany, and his colleagues found when they removed a 150-nucleotide

**"Either there's a hell of a lot of developmentally regulated transcriptional noise, or these RNAs are sending signals into the system."**

— John Mattick

L. E. G. BROSIUS

RNA from mouse neurons, where it is normally transported down the cellular fingers that communicate with other cells<sup>6</sup>. The engineered animals looked and acted more or less the same as the control animals — but Brosius says that on close inspection they weren't as inquisitive and had unusual exploratory behaviours. Such activity might be lethal in the wild, Mattick says, "but it was affecting their behaviour in ways that were far too subtle to be assessed in a cage".

### In search of function

If slicing out non-coding RNA doesn't often reveal its function, then perhaps looking at its lifespan will. This vein of thinking brings a potentially bigger blow for RNA's believers than the knockout studies: the possibility that cells are destroying long RNAs almost as fast as they are making them. Studies in yeast have shown that many long RNAs seem to be so rapidly gobbled by the nuclear exosome — a protein complex that degrades RNA — that it is hard to imagine them having any function at all. Some are labelled for destruction as soon as they peel away from their DNA blueprint. David Tollervey, who studies RNA processing at the Wellcome Trust Centre for Cell Biology in Edinburgh, UK, says that long RNAs could have almost-instant effects or cells might be making many long RNAs merely to show that they've done so. In other words, the point of the exercise might be transcription itself, rather than the transcript.

There are already known examples in which RNA production seems more important than the actual product. In 2004, Fred Winston and his colleagues at Harvard Medical School in Boston, Massachusetts, studied a 551 nucleotide RNA called SRG1 that is made by brewer's yeast<sup>7</sup>. It switches on and off the adjacent gene *SER3*, which helps make serine (an amino acid that the yeast needs to be healthy). But in this case it is the process of making the non-coding RNA that regulates *SER3*, rather than the RNA itself. The trick here is that the DNA sequence from which SRG1 is transcribed runs through the on switch for *SER3*. So when a yeast cell is manufacturing a lot of RNA for SRG1, it blocks access to the *SER3* switch. This is what happens when the yeast sits happily in a flask of rich medium and has no need to generate its own serine.

In his transcriptional surveys of humans, Gingeras has shown that about three times as many transcripts carry a molecular label for rapid destruction than do not carry one. But Gingeras thinks that these apparently doomed RNAs still do more for cells than just getting made. When a map of pervasive transcription is overlaid with a map of short non-coding

RNAs, such as microRNAs, the two overlap<sup>8</sup>. Gingeras thinks that the short RNAs are frequently embedded within the longer transcripts, and then excised.

Over the past few years, Mattick has been gathering other circumstantial evidence that long RNAs have widespread function. In a paper published in January, he and his team examined 1,328 non-coding RNAs whose expression patterns had been mapped in the Allen Brain Atlas, but the functions of which were unknown. The team found that nearly two-thirds of these molecules were produced in specific regions of the mouse brain — in certain cell types or in specific parts of neurons<sup>9</sup>. More recently, Mattick's team identified 174 non-coding RNAs that are expressed in mouse embryonic stem cells in a decidedly selective manner, either correlating with the cells' capacity to develop into any other cell type or with particular events along the path to specialization<sup>10</sup>. "You've only got two alternatives," Mattick concludes. "Either there's a hell of a lot of developmentally regulated transcriptional noise, or these RNAs are sending signals into the system."

This approach should gain more steam as part of ENCODE. The next-generation sequencers have been chugging away since the end of last year, and in 2009 should lay out the sequences of all the RNA molecules manufactured by two types of human cell. When the project eventually delivers transcriptomes for all 44 cell types, it will allow a closer analysis of when different sorts of human cell make different long RNAs and help infer something about their function.

**"Transcripts will be used at very different times and in very different cell types and conditions."**

— Thomas Gingeras

As for Rinn, he already has evidence that non-coding RNAs are so much more than hot air. In May, at this year's Biology of Genomes meeting, he presented work suggesting that there are as many as 2,000 long non-coding RNAs in human cells that shoulder biological responsibilities on a par with those of HOTAIR and that may therefore earn the status of RNA genes.

To find these, Rinn and Manolis Kellis, a computational biologist also at the Broad Institute, searched for sequences that are conserved as might befit a working stretch of RNA. They assumed that much of an RNA molecule's function depends on the three-dimensional architecture that the single-stranded molecule folds into. This, rather than the precise sequence of nucleotides, is what evolution will have worked to preserve. This means that an A can become a T, for example, as long as the T

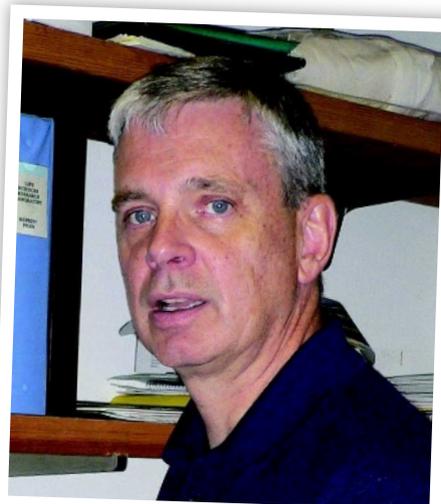
to which it anneals when the molecule folds switches in turn to an A and providing that the overall shape of the molecule is unchanged.

Using these types of bioinformatic rules, the team pulled out probable RNA genes. For a sample of these, they took a stab at predicting function and then tested whether the RNA's production was induced by certain cellular pathways. Many of them were. If their results hold up, Rinn and Kellis will have discovered the first large class of long RNA genes. "These RNAs could have functions as diverse as those of protein-coding genes," Rinn says. And it is not such a stretch to think that they could rival the 20,000-odd protein-coding genes in number, if there are other, as yet unidentified groups of long RNA genes out there.

That still leaves a lot of the transcriptional hairball unaccounted for, and it is possible that much of it is still noise. "With all this pervasive transcription," Rinn says, "the problem to working out whether most of it is functional or not has been that people simply haven't known where to start." Now, perhaps, they do. ■

**Anna Petherick is Nature's Research Highlights editor.**

1. Rinn, J. L. *et al. Cell* **129**, 1311-1323 (2007).
2. Nagalakshmi, U. *et al. Science* **320**, 1344-1349 (2008).
3. Wilhelm, B. T. *et al. Nature* **453**, 1239-1245 (2008).
4. Martianov, I., Ramadass, A., Barros, A. S., Chow, N. & Akoulitchev, A. *Nature* **445**, 666-670 (2008).
5. Willingham, A. T. *et al. Science* **309**, 1570-1573 (2005).
6. Skryabin, B. V. *et al. Mol. Cell. Biol.* **23**, 6435-6441 (2003).
7. Martens, J. A., Laprade, L. & Winston, F. *Nature* **429**, 571-574 (2004).
8. Kapranov, P. *et al. Science* **316**, 1484-1488 (2007).
9. Mercer, T. R., Dinger, M. E., Sunkin, S. M., Mehler, M. F. & Mattick, J. S. *Proc. Natl Acad. Sci. USA* **105**, 716-721 (2008).
10. Dinger, M. E. *et al. Genome Res.* doi: 10.1101/gr.078378.108 (2008).



Thomas Gingeras described 'pervasive transcription'.