Long noncoding RNAs: the search for function

Monya Baker

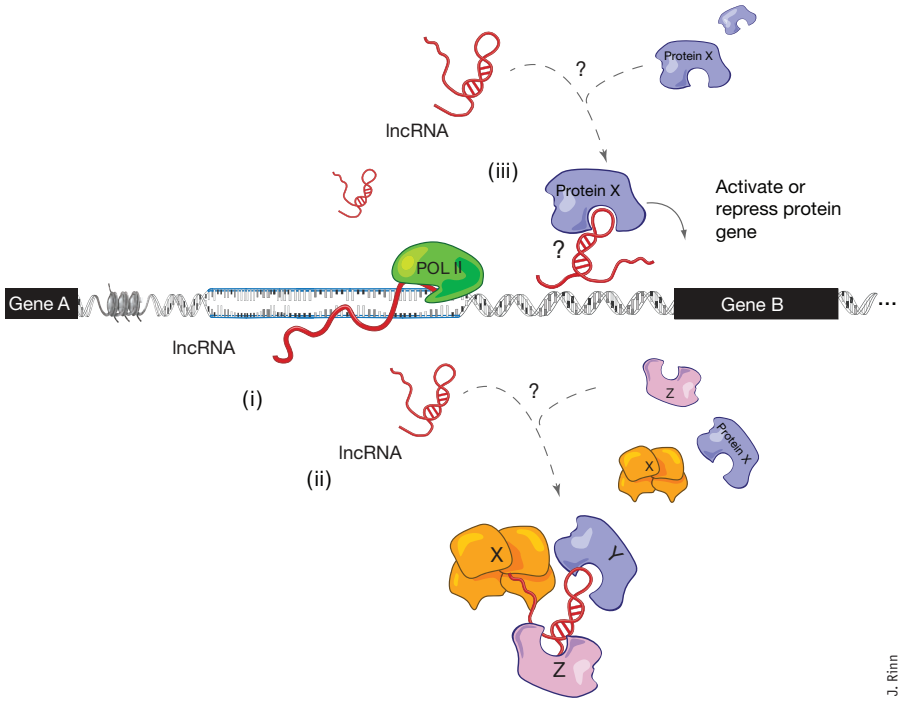
Transcripts are easy to find: sorting out what they do is a challenge.

In the late 1990s, before the publication of the human genome, John Rinn, then at Yale University, was hunting for protein genes on chromosome 22 with his graduate student adviser Michael Snyder. The only genes they found were ones that had already been discovered, but their arrays identified a steady stream of transcribed regions with no apparent purpose. These long noncoding RNAs (lncRNAs) came from genome regions that were known to lack protein genes. The transcripts also lacked open reading frames and other properties necessary for them to be translated into proteins.

Most scientists at the time dismissed such transcripts as noise, but Rinn kept doing experiments. "I started cloning them," he recalls, "and I realized that if I could clone them, they must be stable." And if they were stable, he thought, perhaps they were functional, too. In 2004, Rinn took a postdoctoral position with Howard Chang at Stanford University, after the two came up with a scheme to learn what, if anything, these mysterious transcripts were doing.

This work led eventually to the discovery of a noncoding RNA they named HOTAIR[1]. This 2.2-kilobase spliced RNA transcript interacts with the protein complex polycomb to modify chromatin and repress transcription of the human *HOX* genes, which regulate development. How exactly it does so is still unclear.

What is clear is that HOTAIR is just one of thousands of lncRNAs. Although less than 2% of a mammalian genome codes for protein, studies consistently show that half or even more of the genome is transcribed. Partly because suitable research tools are in their infancy, scientists are only beginning to uncover the functions of these transcripts (**Box 1**).



Long noncoding RNAs (lncRNAs) could be a byproduct of transcription (i), a scaffold linking proteins (ii) or a guide bringing proteins to specified parts of the genome (iii). The same lncRNA can function simultaneously as a scaffold and a guide. POL II, RNA polymerase II.

Function focus

Everyone has favorite analogies for how lncRNAs might function. Chang recently showed that HOTAIR serves as a 'modular scaffold', assembling a molecular cargo of specific combinations of enzymes that are equipped to regulate target genes[2]. Rinn likens some of these scaffolds to an air traffic controller, guiding regulatory machinery to the appropriate spots in the genome. His work has shown that hundreds of lncRNAs are physically associated with polycomb and other chromatin-modifying complexes[3]. That, he says, would explain why the same protein complexes act on different sequences in different cells. Other researchers have suggested an effector component: lncRNA binds to a protein, changing its structure and activating it[4]. Tom Cech, a scientist at the University of Colorado at Boulder who won the Nobel Prize for his work on RNA, believes that each of these mechanisms and more may be in play.

The possibilities seem endless (**Box 2**). Some lncRNAs may even enhance transcription through chromosome looping or other means[5,6]. "I don't know why people think that lncRNAs are all doing one thing," says Rinn. "They are just new types of genes, and their repertoire of functions I think will rival the proteome."

BOX 1 LONG NONCODING RNAs REGULATING GENES

Although researchers are still struggling to uncover the mechanisms and protein partners of long noncoding RNAs (lncRNAs), evidence of their importance in basic biology is pouring in.

Last year, John Rinn at the Broad Institute and colleagues used a specially designed microarray to test the expression of about 900 lncRNAs in human fibroblasts and pluripotent stem cells. Just over 100 lncRNAs were induced when fibroblasts were reprogrammed to pluripotency; a similar number of lncRNAs were repressed. The team focused on a couple of dozen lncRNAs that were more upregulated in induced pluripotent stem cells than in embryonic stem cells, reasoning that these would be particularly important for reprogramming. Previously published data indicated that Oct4, an essential transcription factor for pluripotency, was binding sites in the genome identified as coding for lncRNAs. Gain- and loss-of-function experiments showed that at least one of these, called lincRNA-RoR, for long intergenic noncoding RNA and regulator of reprogramming, was essential for a variety of functions, including reprogramming as well as modulating genes known to respond to oxidative stress, DNA damage and p53, a protein that regulates the cell cycle and is implicated in about half of all human cancers[12].

Another set of experiments hinted at extensive regulatory systems featuring lncRNAs. Several lncRNAs were found to be regulated by p53; one transcript, lncRNA-21, binds a protein known as heterogeneous nuclear ribonucleoprotein K (hnRNP-K). Almost 600 genes are affected in common by all three—p53, lncRNA-21 and hnRNP-K (ref. 13).

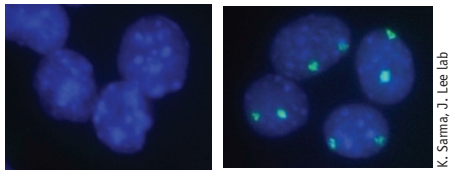
This March, Howard Chang at Stanford University described how a lncRNA called HOTTIP can help to activate the transcription of several *HOXA* genes *in vivo*. The molecule is transcribed from the tip of the *HOXA@* locus, where it binds adaptor proteins and sets nearby chromatin marks to drive transcription. When HOTTIP was knocked down, the proteins MLL1 and WDR5 were not observed on the transcription start sites of *HOX5* genes as they normally are. However, these effects could not be rescued by expressing HOTTIP from another region on the genome, indicating that the lncRNA must work from the chromosome on which it is transcribed. Indeed, chromosome conformation capture techniques have indicated that chromatin looping brings the RNA into close proximity to the activated genes[3].

Because their functions are difficult to study, long noncoding RNAs are generally classified by their origins. They seem to come from everywhere in the genome: the noncoding side of genes, alongside and between the protein-coding regions and, especially, the long stretches in genomes where no protein-coding genes are thought to exist at all. Noncoding transcripts are traditionally classified as long at around 200 nucleotides, an arbitrary distinction based on RNA purification technologies. Most are thousands of nucleotides long.

Nowadays, new putative lncRNAs are generally identified by an RNA sequencing technique called RNA-seq, which uses high-throughput sequencing to profile cell transcripts. Chromatin analysis also helps. Work by Rinn and others showed that histone methylation patterns that are characteristic of transcribed protein genes also apply to lncRNAs, and resulting 'chromatin-state maps' have now been used to flag thousands of putative lncRNAs[6].

So many lncRNAs are being identified that it is difficult to know what to study in depth, says John Mattick, a genome biologist at the University of Queensland. His tactic is to use microarray data to find transcripts with the greatest change in expression between tissues; differences of more than 20-fold are not uncommon, he says. "When you have a mountain of things to look at, you just want the ones that are sticking well above the pack." And though the work is still in early stages, it seems to be paying off. Knocking down or ectopically expressing these transcripts changes cells' phenotypes for about half of the cases he has investigated, he says.

Still, the debate over just what proportion of all noncoding RNAs are functional is surprisingly fierce. "It is difficult to discriminate functional transcripts from those that may be byproducts of other processes," says Tim Hughes, a genome biologist at the University of Toronto, "but many transcripts that come from intergenic regions are starting to look like real signals. They show up relatively consistently in different experiments, contain splice junctions and are present in high numbers," says Hughes. Still, he advocates caution. Differential expression could be explained by activity in nearby protein-coding genes, for example. "Higher abundance presumably increases the likelihood that a transcript is functional, but it's not really proof," he says. "Ultimately we have to go in and do experiments to demonstrate that things have function."



K. Sarma, J. Lee lab

RNA FISH probes (green) that bind the lncRNA Xist (right) can be displaced using specially designed sequences of locked nucleic acids (left).

Advances in uncovering the function of lncRNAs could be self-reinforcing, says Chang. Standard approaches often neglect noncoding genes. Whole-exon capture used in disease association studies, for example, restricts sequencing to regions that code for proteins. And when researchers find that a transcription factor binds to an intergenic region, their first instinct is often to investigate the nearest 'protein gene'. "The knowledge that there are long noncoding RNA genes could change someone's strategy," says Chang.

Tool box

As more researchers begin to investigate lncRNAs, there is a greater need to annotate them so that researchers know what to look for, know if they find something new and know how to name what they find. In general, microarrays cannot distinguish between different forms of a transcript, and sequencing often indicates multiple 'isoforms' of a transcript without indicating which is the most biologically relevant. "Protein-coding RNA has been studied for long enough that most major forms of the transcripts are known, but lncRNAs are too new for that," says Rinn. Right now, he says, it is not always clear where a lncRNA gene starts or stops.

BOX 2 POSSIBLE ROLES FOR LONG NONCODING RNA

Junk. Sloppy machinery means some sequences are transcribed unnecessarily.

Byproduct. The act of transcription may help to prepare the genome for future transcripts or open the DNA to activate nearby genes.

Scaffold. Various protein complexes may need to work in unique combinations; noncoding RNA keeps the proteins together.

Guide. Noncoding RNA may guide complexes to the right spots in the genome.

Effector. An intimate collaboration of RNA and protein allows a protein to modify chromatin or otherwise regulate gene expression.

Enhancer or activator. The promoters of some protein genes may get a boost from noncoding RNAs.

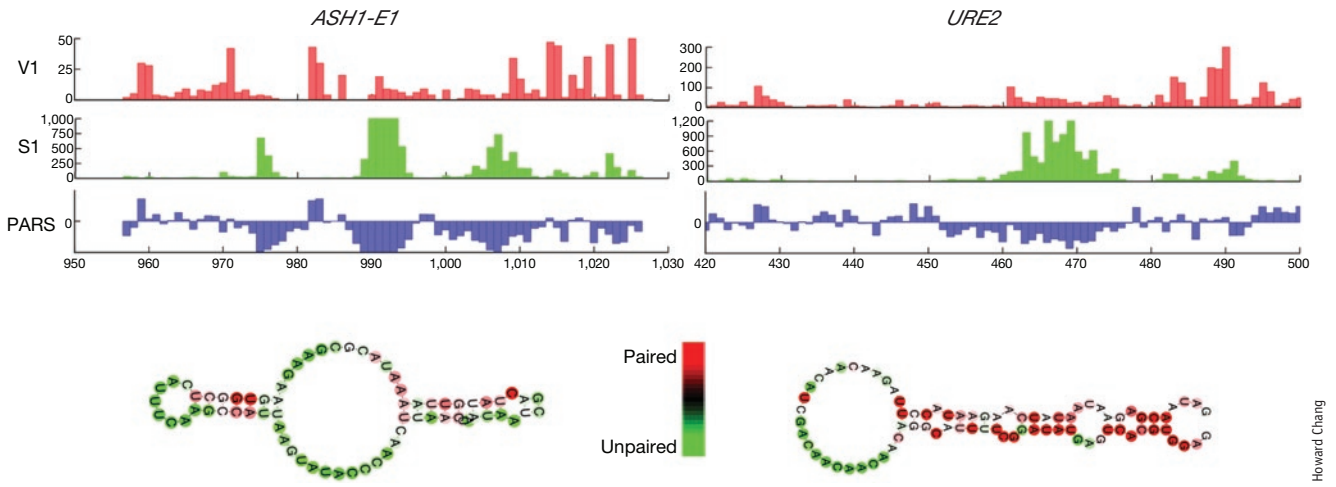
Several cataloguing and annotation efforts are underway. Early this year, Mattick established a database (http://www.lncrnadb.org/) especially for long noncoding RNAs backed by experimental data. Entries are manually curated from the research literature and linked to the University of California Santa Cruz Genome Browser and Noncoding RNA Expression database[7]. FANTOM (functional annotation of the mammalian genome), a large international consortium led by scientists at the RIKEN Yokohama Institute, has documented tens of thousands of noncoding RNA transcripts in mouse tissues at several stages of development.

The Havana team of the ENCODE (Encyclopedia of DNA elements) consortium is manually annotating lncRNAs in the human genome. The field is so new that just naming the transcripts can be difficult, says Jennifer Harrow at the Wellcome Trust Sanger Institute, who coordinates the Havana team. They use a combination of RNA-seq data, chromatin-state maps and computer algorithms to identify lncRNAs, but there is much more work to be done, Harrow says. Transcriptional evidence alone cannot show what a lncRNA is doing.

Meanwhile, experimental tools designed for other applications are being extended to lncRNAs. Companies such as Active Motif and Millipore sell RNA immunoprecipitation kits to purify proteins and to identify bound lncRNAs in ribonucleic protein complexes. Many lncRNAs are now represented on microarrays from Agilent and Life Technologies. Life Technologies also sells TaqMan quantitative PCR (qPCR) assays to precisely evaluate the expression of certain lncRNAs. The RNAi (RNA

Members of J. Rinn laboratory

lncRNAs allow proteins to regulate different genes in different cells, says John Rinn, of the Broad Institute.

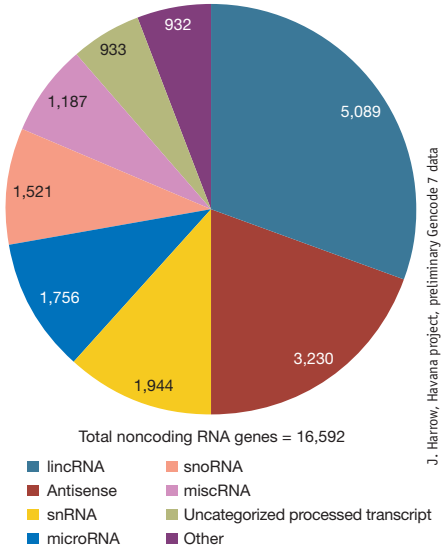


Parallel analysis of RNA structure (PARS) sequence fragments produced by nucleases to simultaneously identify the structures of many long RNA transcripts.

interference) Consortium maintains a reference set of RNA sequences for knocking down lncRNAs. These tools are useful, but more are needed, says Jeannie Lee, who studies noncoding RNA at Harvard Medical School. For example, the machinery for knocking down RNA is mostly in the cytoplasm, but many lncRNAs are in the nucleus, a fact that makes loss-of-function experiments inefficient.

High on scientists' wish lists are techniques that can be used to identify both the genome regions and the proteins with which lncRNAs interact. Long noncoding RNAs do not always undergo canonical base-pairing, so the sequence of a transcript yields few clues about how it interacts with the genome. In December 2010, Lee and scientists from Harvard University and reagents company Exiqon showed that non-natural nucleic acids could be used to watch how Xist, a 17-kilobase lncRNA and one of the first discovered, interacts with the X chromosome. Her team used locked nucleic acids that were complementary to two sections of Xist, then used fluorescent probes to observe how the lncRNA and associated proteins disassociated and reassociated with the genome[8]. Though Xist is an unusual lncRNA—it coats most of the inactive X chromosome—Lee believes the technique will be generally applicable. In fact, she says, as most lncRNAs are smaller than Xist, the search for disruptive sequences might be easier and less expensive.

Lee has also described a genome-wide technique for identifying lncRNAs bound to particular proteins[9]. In chromatin immunoprecipitation–sequencing (ChIP-seq) assays, researchers use antibodies to pull transcription factors from cell lysates, then wash away and analyze bound DNA to learn where transcription factors bind on the genome. RNA immunoprecipitation followed by sequencing (RIP-seq) exploits a similar idea, but instead uses antibodies to pull ribonucleoproteins from cell lysates and determines which RNA molecules are associated with them. Lee used the technique to identify more than 9,000 lncRNAs that interact with the polycomb complex in embryonic stem cells. Getting the system to work required a lot of optimization, says Lee. Her team had to try several batches of antibodies from several vendors before finding one with sufficiently high affinity and specificity. Even with a good antibody, she says, data are inherently noisy, making high-quality controls particularly important. "Your pulldowns won't mean a thing unless you have something to compare them to," she says.



Long noncoding RNAs are just one of many noncoding transcripts being annotated. lincRNA, long intergenic noncoding RNA; snRNA, small nuclear RNA; snoRNA, small nucleolar RNA; and miscRNA, miscellaneous RNA.

Getting structure

Figuring out what precisely lncRNAs are doing requires more than the identification of a transcript's protein partners, says Tom Cech, of Colorado University: "Even for the most well-studied noncoding RNAs, the field is still grappling with the question of what are the relevant RNA structures." Computational tools are often used to assess the structures of smaller RNAs, but lncRNAs represent a more difficult challenge.

Unlike protein-coding genes and shorter RNAs, lncRNAs are poorly conserved between species, which makes it harder to translate results from one organism to another and also lends an additional degree of uncertainty about whether a given lncRNA is functional. Many researchers suspect that although sequences are not well conserved, the structures they form may well be. If this is true, structural information could provide a more meaningful way to classify lncRNAs. With enough data, structures might become reliable indicators of function, and so guide researchers toward better follow-up experiments.

Not surprisingly, several researchers are working on the problem. Together with Eran Segal at the Weizmann Institute of Science in Israel, Chang recently described

Mark Yamaguma, Stanford University

Howard Chang of Stanford University believes that understanding the structure of long noncoding RNA could reveal much about its function.

a technique that uses high-throughput sequencing to assess the structure of the lncRNAs in an entire yeast transcriptome[10]. In the technique, called parallel analysis of RNA structure (PARS), transcripts are first digested with a set of two nucleases that cleave RNA in certain single-stranded or stacked-base conformations; the digested fragments are then sequenced and used to determine which sections of the RNA exist in single-stranded and other conformations. Chang is currently working out ways to use PARS in living cells, the better to compare the lncRNA transcriptome under different conditions.

Jeannie Lee at Harvard Medical School is characterizing Xist, one of the first long noncoding RNAs to have been discovered.

Separately, a team of researchers from the University of California Santa Cruz published a similar technique called Frag-seq for fragmentation sequencing, using only one nuclease[11]. Analysis of digested fragments over the entire mouse transcriptome successfully mapped single-stranded regions in multiple ncRNAs whose structures are known.

Using sequencing to identify structure is far more challenging than using it to identify transcripts, says Chang. Even simple things could make a big improvement. For example, says Cech, a nuclease that precisely cuts only double-stranded RNA could make for more accurate analysis.

These genome-wide approaches are useful, says Cech. But he suspects that before functional classes can be definitively assigned, more details will need to be carefully worked out for a few systems. What we need, he says, are "multiple examples that can be taken down to the structural and mechanistic level so that we have the same sort of understanding as we do for transcription and RNA splicing and translation and other cellular processes. Until we drill down to that level, we don't have much understanding."

1. Rinn, J.L. *et al. Cell* **129**, 1311–1323 (2007).
2. Tsai, M.C. *et al. Science* **329**, 689–693 (2010).
3. Khalil, A.M. *Proc. Natl. Acad. Sci. USA* **106**, 11667–11672 (2009).
4. Wang, X., Song, X., Glass, C.K. & Rosenfeld, M.G. *Cold Spring Harb. Perspect. Biol.* **3**, a003756 (2011).
5. Ørom, U.A. *et al. Cell* **143**, 46–57 (2010).
6. Guttmann, M. *et al. Nature* **458**, 129–140 (2009).
7. Amaral, P.P., Clark, M.B., Gascoigne, D.K., Dinger, M.E. & Mattick, J.S. *Nucleic Acids Res.* **39**, D146–D151 (2011).
8. Sarma, K., Levasseur, A., Aristarkhov, A. & Lee, J.T. *Proc. Natl. Acad. Sci. USA* **107**, 22196–22201 (2010).
9. Zhao, J. *et al. Mol. Cell* **40**, 939–953 (2010).
10. Kertesz, M. *et al. Nature* **467**, 103–107 (2010).
11. Underwood, J.G. *et al. Nat. Methods* **7**, 995–1001 (2010).
12. Loewer, S. *et al. Nat. Genet.* **42,** 1113–1117 (2010).
13. Huarte, M. *et al. Cell* **142,** 409–419 (2010).

Monya Baker is technology editor for *Nature* and *Nature Methods* (m.baker@us.nature.com).