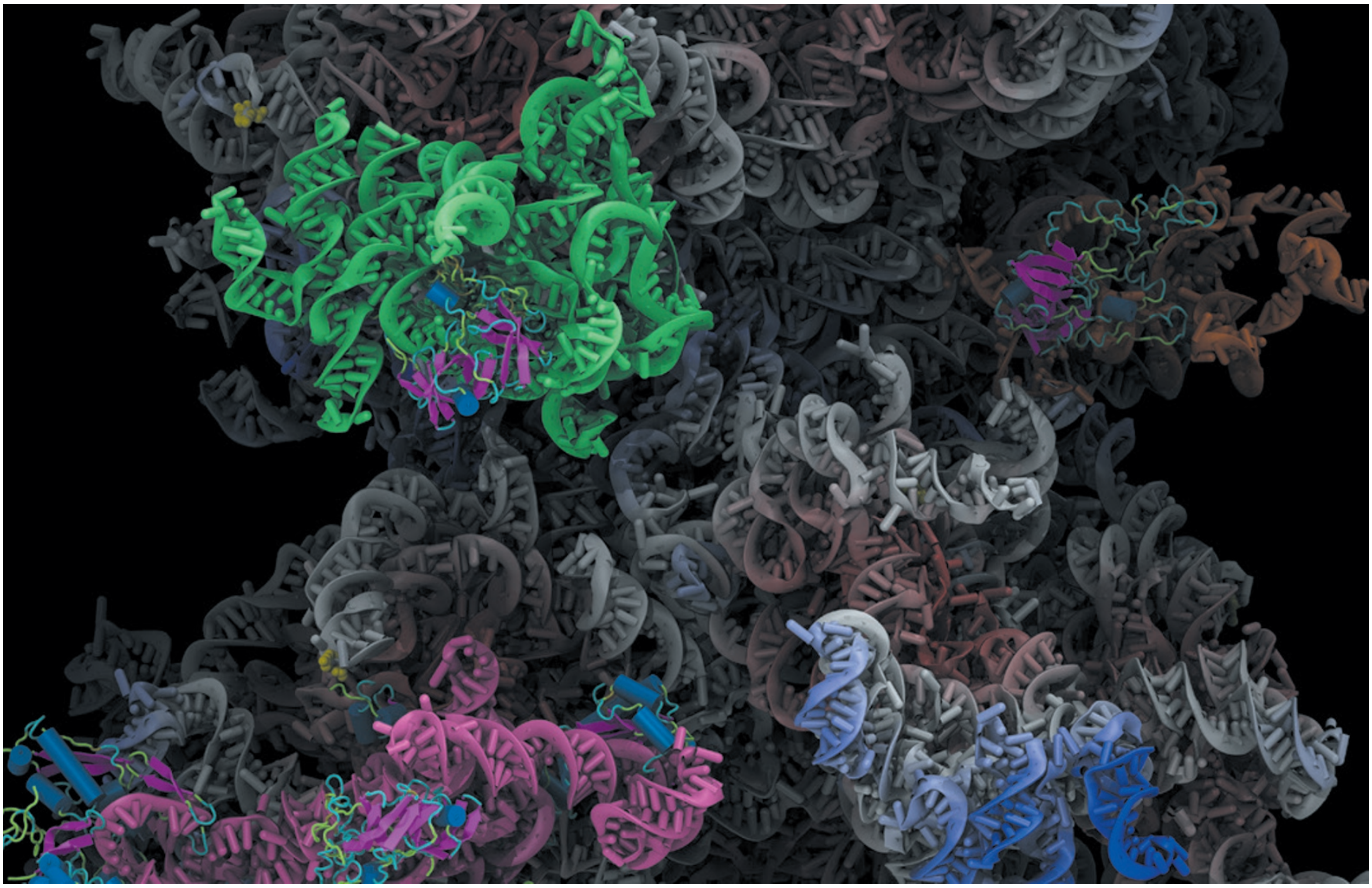
TECHNOLOGY FEATURE

FINDING FUNCTION IN MYSTERY TRANSCRIPTS

Little is known about the function of most long non-coding RNAs. But a suite of new tools might change that.

KARISSA Y. SANBONMATSU



Conceptual model of a long non-coding RNA that provides structural scaffolding (grey) while binding to proteins (magenta) using highly structured domains (green, pink and orange).

BY KELLY RAE CHI

In 2013, a group of researchers decided to dig deeper into a human embryonic stem-cell line called H1 — and uncovered some surprises. H1 is one of the best known stem-cell lines, yet the team managed to unearth more than 2,000 previously uncharacterized stretches of RNA[1]. What is more, 146 of those were exclusive to human embryonic stem cells, offering tantalizing leads into pluripotency — the ability to become any cell type in the body.

These transcripts had gone unnoticed because they contain repetitive stretches of code that sequence analysers had tended to filter out.

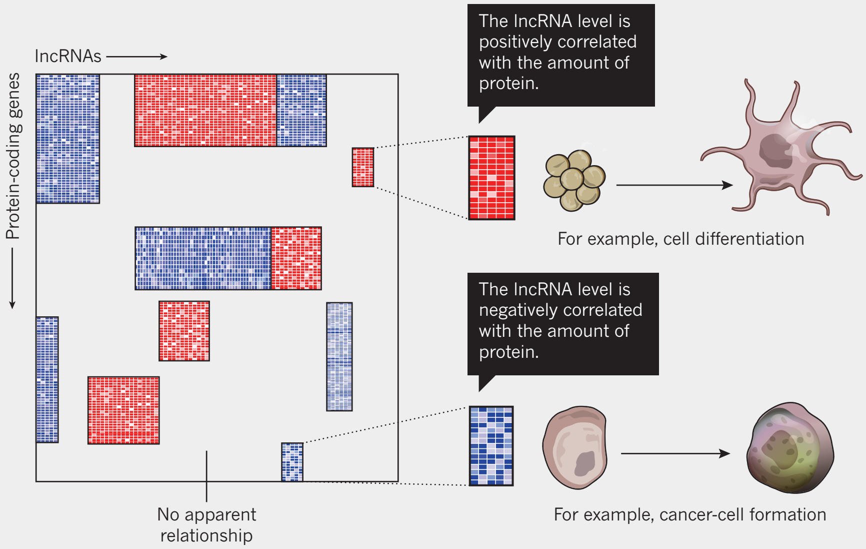
It was a big blind spot. Other labs had uncovered early evidence of RNAs that are rich in repetitive codes and important in human stem cells. As the researchers, who were based mostly at California's Stanford University, examined their haul, they realized that they had hit on exactly these kinds of RNA. Among their list of 146 RNA sequences, says team member Vittorio Sebastiano, three of the most abundant — which they named HPAT2, HPAT3 and HPAT5 — seemed to be necessary for establishing the pluripotent cells that develop into a human fetus: those that comprise the 'inner cell mass' of an embryo that has yet to implant in the uterus[2].

These RNA stretches are examples of long non-coding RNAs (lncRNAs) — sequences at least 200 bases long that do not encode proteins. lncRNAs are present in many different kinds of tissue and are often found in specific spots inside a cell. But most lack a defined function and, until recently, were thought to be little more than transcriptional noise.

That view began to shift as more data rolled in showing that the genomic regions from ▶

GUILT BY ASSOCIATION

Grouping long non-coding RNAs (lncRNAs) according to how their presence correlates with the expression of certain proteins can offer clues to what cell processes these mysterious transcripts are involved in.


lncRNAs
Protein-coding genes
The lncRNA level is positively correlated with the amount of protein.
For example, cell differentiation
The lncRNA level is negatively correlated with the amount of protein.
For example, cancer-cell formation
No apparent relationship

ADAPTED FROM M. GUTTMAN & J. L. RINN *NATURE* **482**, 339–346 (2012)

which lncRNAs are transcribed are more highly conserved through evolution than was imagined, implying that they had some function. But, to this day, a neat and sensible classification system for lncRNAs remains out of reach. They are still 'each their own snowflake', says John Rinn, who discovered lncRNAs as a graduate student about 15 years ago and now runs a lab specializing in the molecules at the Broad Institute of MIT and Harvard in Cambridge, Massachusetts.

Now, revolutionary tools such as the genome-editing platform CRISPR–Cas9 are making the task of discovering what individual lncRNAs do much easier.

Some lncRNAs are thought to act as scaffolds for proteins to hang off while they manipulate the packaging of DNA. The functions of others are merely hinted at by their co-occurence with proteins (see 'Guilt by association') or by the effects of their absence — cancers start to spread to other parts of the body, and developmental disorders such as autism spectrum disorder arise.

Today, it is broadly assumed that the molecules do have biological functions, says geneticist John Mattick, who heads the Garvan Institute of Medical Research in Sydney, Australia. "The evidence is on the table. That's been the big change — and it's palpable," he says.

There are still doubters, but arguments now tend to caution against assuming function rather than dismiss its likelihood altogether. In the cautious camp is Michael McManus at the University of California, San Francisco. In 2013, his group mined published data sets of RNA sequences and found tens of thousands of new human lncRNAs, although many were present in cells only in very low amounts[3]. Most of these still need to be followed up, McManus says. And even with the latest tools, figuring out what the myriad lncRNAs do will be a slog requiring an army of experts.

FUNCTION FIRST

Because the list of lncRNAs is so long, a key step is deciding which ones to prioritize for study. Rinn advocates starting with those from regions of the genome that have been already linked to disease. Another idea is to look at where the lncRNAs are located — finding one near a transcription start site might mean that it is involved in regulating the nearby gene, for example. These days, researchers can track the location of molecules inside cells with relative ease. Rinn, along with others at the Broad Institute and at the University of Pennsylvania in Philadelphia, has managed to discern the positions of 61 lncRNA molecules within skin, lung and cervical tumour cells using a technique known as single-molecule RNA-FISH (RNA fluorescence *in situ* hybridization)[4].

Scientists can also now test lncRNA function by using CRISPR–Cas9 and other gene-editing techniques to interfere with part of the DNA sequence from which it is transcribed or with the promoter that directs its transcription. Some of these technologies allow labs to quickly screen huge numbers of lncRNAs. The logic is the same as when CRISPR–Cas9 is used to look at the function of a protein-coding gene: introduce single-base deletions or substitutions into the DNA and watch the effects of the altered transcript.

The only problem is that lncRNAs are less likely than proteins to be disabled by subtle alterations, says cancer biologist Howard Chang of Stanford School of Medicine, who develops and applies new methods for studying how lncRNAs work. The sequence alterations often need to be more drastic.

This is where RNA researchers have made CRISPR–Cas9 their own. They have expanded the CRISPR toolbox to include ways to block or prompt the transcription of a specific lncRNA. Rinn and his team have developed yet another approach: a tool known as CRISPR-Display (or 'CRISP-Disp'). Rinn compares it to a drone that can deliver an item — in this case, a specific lncRNA — anywhere in a cell. If a role in gene activation is suspected because a lncRNA normally lies alongside a certain part of the genome, then that role can be tested by moving the lncRNA to a different genomic location and watching for gene activation in the new spot.

His group had been trying for years to make this happen using older genome-editing methods. Then, once CRISPR–Cas9's crystal structure was published in 2014 (ref. 5), the team was able to tweak CRISPR's machinery to carry large packages, and had CRISP-Disp up and running within months. "It's very high-throughput: we can put 100 different lncRNAs at different sites and ask what they do at once," says Rinn.

But figuring out function using this and other CRISPR techniques that block lncRNAs can be more complicated than it seems. For some lncRNAs, the DNA code overlaps with regions that are important for protein-coding genes, so destroying those makes the effects tricky to interpret, says Andrew Bassett, a genome-editing specialist at the University of Oxford, UK. And just because a functional change isn't seen doesn't mean that there is no function; the effect may be very subtle, or perhaps revealed only when the cell is faced with a particular threat.

Two often-cited examples of this involve lncRNAs known as NEAT1 and MALAT1. They are abundant in cells and are highly conserved across mammals. Researchers know that they bind DNA to protein, but deleting the DNA stretches has no observable effects in mice. It is a familiar story to researchers in the field. "There are mysteries everywhere," says Mattick.

THE RNA INTERACTOME

An entirely different approach is to find out what the lncRNAs are interacting with. "It's still believed, despite the importance of lncRNAs, that, ultimately, they can't carry out their functions without accessory factors," says molecular biologist Jeannie Lee of Massachusetts General Hospital in Boston and co-founder of the company RaNA Therapeutics in Cambridge, Massachusetts. These accessory factors are almost always proteins.

Lee and others have set about unveiling lncRNA interactions using a lncRNA — called Xist— that is known to be necessary for silencing one of the two X chromosomes in the cells of female mammals to stop females from having twice as many X-chromosome gene products as males. The proteins that bind to Xist silence gene expression through multiple mechanisms.

LEFT: WWW.MICHAELWEST.ORG; RIGHT: REF. 2

But in the past year, scientists have finally made inroads into pinning down the identities of these partners. Now, the proteins are known to not only pull in other molecules that silence transcription, but also to repel some that promote it.

A host of techniques are available for probing a lncRNA's crowd of protein partners: broadly, researchers link RNA and protein together using agents such as formaldehyde or ultraviolet (UV) light, then use mass spectrometry to parse what is bound to what and come up with the 'interactome'. Often, these analyses have many steps, and therefore require the scientist to make many strategic choices. How should the RNA and protein be linked? How can real signals of interactions be distinguished from artefacts? And hanging over all such studies is the problem that RNA often behaves differently *in vitro* from how it does inside a cell.

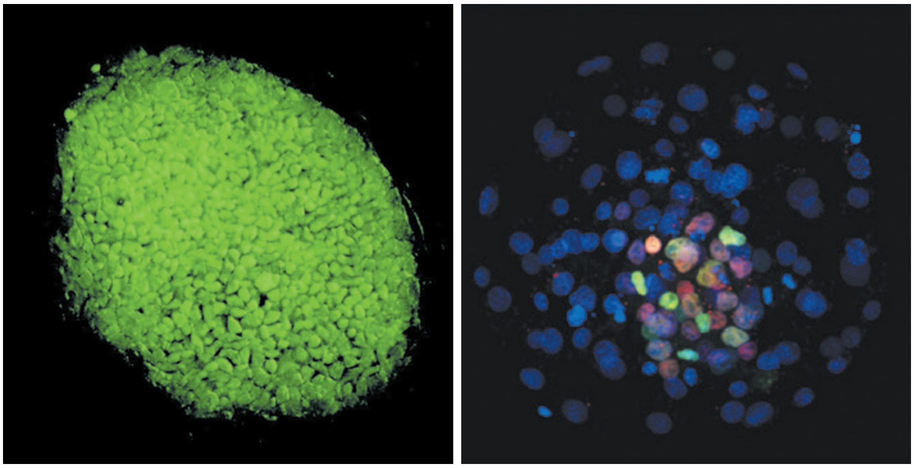
This is why those working on Xist interactions have focused on techniques for identifying RNAs bound to proteins inside living cells. Helping them are recent improvements in the sensitivity of mass spectrometry. In the past year, Chang's group has combined an assay that uses formaldehyde as a linking agent with the latest mass-spectrometry techniques to demonstrate that Xist binds to 81 proteins *in vivo*[6]. Guttman's group used UV light instead — and revealed ten partners, including one not found by Chang's group[7].

Lee has been working on another UV light method called iDRiP (identification of direct RNA-interacting proteins) and has revealed an Xist interactome of about 100 proteins[8]. That is a complex that rivals the ribosome for size. Lee thinks that iDRiP could be used to look at other lncRNAs but that the protocol would probably need tweaking for each lncRNA. The chosen linking method, she adds, will depend on a number of factors, including where the lncRNA is in the cell and how abundant it is.

Even for the well-studied Xist, the full interactome is far from settled. One of the central debates concerns whether, and how, Xist interacts with a structure called polycomb repressive complex 2 (PRC2), which influences gene expression by modifying proteins called histones. Some groups, including Guttman's, have found little evidence that Xist binds to PRC2; others insist that it does. Guttman thinks that the reason could be down to different experimental protocols: "Xist may bind more strongly *in vitro* than *in vivo*. It's also a question of how one separates the strongest binding interactions from background interactions," he says.

> ***"There are so many different lncRNAs that there's going to be a large zoo of different classes and motifs."***

The PRC2 debate highlights the importance of following up interactome assays with tests of whether breaking a lncRNA–protein interaction changes the RNA's function. Guttman has found that PRC2 deletion does not seem to affect Xist's ability to silence the X chromosome. By contrast, he says, perturbing the interaction between Xist and another protein implicated in gene silencing — called SHARP — does.



Stem cells known as H1 cells (coloured green, left image) have the ability to develop into any cell type partly because they contain long, non-coding RNAs. These same RNAs are found in the inner cell mass (various colours, right image) of blastomeres.

SECRETS IN STRUCTURE

A third avenue for probing the function of lncRNAs is to study their structure. This doesn't predict function as directly as it often does for proteins, but knowing more about an RNA's arches and folds is likely to inform nonetheless. "It's a wide open field that needs a lot of work," says structural biologist Karissa Sanbonmatsu of the Los Alamos National Laboratory in New Mexico. "There are so many different lncRNAs that there's going to be a large zoo of different classes and motifs."

Methods for establishing the secondary structure of a lncRNA include chemical probing strategies, such as one called SHAPE. It involves attaching acetyl groups to the RNA, modifying its backbone only at flexible regions. The modified sites block the enzyme that 'reads' RNA to create a complementary DNA sequence so that short DNA fragments are generated rather than long strands. The fragments can then be sequenced or sized on a gel.

Sanbonmatsu's group was the first to describe, in 2012, the secondary structure of a human lncRNA: the steroid receptor RNA activator (SRA), which had been known for more than a decade to associate with oestrogen receptors[9].

By chemically probing the bound and unbound SRA, as well as its various domains, Sanbonmatsu revealed the lncRNA in its full glory, including all of its stems, loops and bulges. It looked a lot like the 16s ribosomal RNA, a highly conserved molecular machine. Sanbonmatsu's team has since found other strongly structured lncRNAs, but it is unclear whether most lncRNAs are like this or whether they are floppy, or somewhere in between, she says.

The structural approach, too, has to cope with the problem that RNA behaves differently in test tubes and cells. And as with binding studies, the latest techniques are being done *in vivo*. In 2012, a team that included Chang described a version of SHAPE that can work inside living cells[10] and has since improved it to characterize thousands of RNA structures simultaneously.

Structural studies, like the others, require a large time investment — so careful choices must be made to focus on the lncRNAs that are most likely to have functions. Luckily, researchers are getting better at such triage, Sanbonmatsu says. She suggests to determine the likelihood of functional significance, scientists should start with lncRNAs that have known phenotypes, then chemically probe them to obtain secondary structures and check the extent to which they are conserved across other species.

Sebastiano already has those boxes ticked for the three lncRNAs that seem to be key in establishing the pluripotent inner cell mass of human embryos: HPAT2, HPAT3 and HPAT5. Now he plans to delve further into the mechanistic details of these factors and has a raft of planned experiments on his list, including assays to ascertain their interactomes as well as structural analyses. "There's a ton of work to do, and this is just the beginning," he says. "But considering that these sequences may explain a lot of our unique features as humans and as primates, the effort is well worth it." ■

Kelly Rae Chi *is a freelance science writer based in Cary, North Carolina.*

1. Au, K. F. *et al. Proc. Natl Acad. Sci. USA* **110,** E4821–E4830 (2013).
2. Durruthy-Durruthy, J. *et al. Nature Genet.* **48,** 44–52 (2016).
3. Hangauer, M. J., Vaughn, I. W. & McManus, M. T. *PLoS Genet.* **9,** e1003569 (2013).
4. Cabili, M. N. *et al. Genome Biol.* **16,** 20 (2015).
5. Jinek, M. *et al. Science* **343,** 1247997 (2014).
6. Chu, C. *et al. Cell* **161,** 404–416 (2015).
7. McHugh, C. A. *et al. Nature* **521,** 232–236 (2015).
8. Minajigi, A. *et al. Science* **349,** aab2276 (2015).
9. Novikova, I. V., Hennelly, S. P. & Sanbonmatsu, K. Y. *Nucleic Acids Res.* **40,** 5034–5051 (2012).
10. Spitale, R. C. *et al. Nature Chem. Biol.* **9,** 18–20 (2013).