

Some roads ahead for noncoding RNAs



Researchers roll up their sleeves for the work ahead in the noncoding RNA field.

By Vivien Marx

Scientists who study noncoding RNAs (ncRNAs) are mapping out how they want to explore them and their functions in cells and tissues in health and disease. Some teams outline future plans in papers. A perspective on best practice standards for circular RNA research¹ includes recommendations for purifying, profiling, quantifying and validating these ncRNAs and approaches to determining regulatory mechanisms.

A community-driven publication forthcoming in *Nature Reviews Molecular Cell Biology*² addresses ways to explore long noncoding RNAs (lncRNAs) in development, cell biology and disease. The authors note that GENCODE³ has identified more than around 20,000 lncRNAs and the FANTOM⁴ consortium 30,000. Among other aspects, they point out that lncRNAs often have low expression levels, which can lead to undersampling. Some ways that researchers explore to address this include targeted capture, imaging, spatial transcriptomics and single-cell sequencing approaches. The authors note it's important to develop and apply methods to identify and understand the roles of lncRNAs and RNA networks, such as by exploring lncRNA localization and structure–function relationships. Methods such as sequencing, chemical probing and imaging can be brought to bear to gain better understanding of lncRNA roles in cell and developmental biology, neurological disorders and cancer, among others. This paper is not a to-do list for any one individual lab, says Caltech researcher Mitch Guttman, a co-author, but it's “what we think of as sort of the next stages for us as a community.”

In his view, says co-author and University of Queensland researcher Tim Mercer, functional analysis of ncRNAs such as *Xist*, *H19* or *HOTAIR* and other ncRNAs has become well-established. Characterizations of yet other ncRNAs may not have yielded the clear-cut results experimenters would like, but methods for characterizing ncRNAs keep emerging and data have amassed.

As matters shift from sweeping statements about junk and transcriptional noise, tasks shift to the practicalities of exploring functionality



There's plenty to do in the noncoding RNA field. Matters are shifting from statements about 'junk' and 'noise' to the study of ncRNA function, such as their roles in health and disease.

of ncRNAs to uncover their roles in differentiation, development and disease, says Mercer. He sees a new generation of scientists settling in to do the “hard work” of building on the field's accomplishments, in which technology development and application have mattered. It will matter, for example, to combine methods – existing ones and new ones still to be developed. And it means being an explorer. Just as the Amazonian rain forest is sculpted on many levels by evolution, these forces have sculpted the human genome, including ncRNAs. Both rain forest and genome are abuzz with activity. The genome is constantly being transcribed; isoforms emerge; there's splicing; genes interleave with other genes. The genome is far from what was once called “islands of genes among intergenic deserts.”

Beneficial tools

“RNA-seq is an amazing technology to profile noncoding RNAs,” says Maite Huarte, a researcher at Cima Universidad de Navarra in Spain who studies lncRNAs and gene

regulation in cancer⁵. She also co-authored the community-driven paper² on lncRNAs. RNA-seq is versatile and can be adapted to detect a specific type of transcript. For example, one can select for criteria such as size or RNA with 3' or 5' ends. The technique can be combined with various RNA purification protocols, immunoprecipitation, subcellular fractionation and proximal labeling, among others. It's this method's range, she says, that “is giving us a vast view of noncoding RNA expression.” This view will be expanded through third-generation RNA-seq using long reads and direct RNA sequencing, she says. What is still lacking in her view is a full understanding of how lncRNAs work and knowledge of whether their mode of action might shape subclasses of ncRNAs. Key questions ahead involve determining ncRNA structures and dynamic behavior inside the cell. “What are their interactors? What are the specificity determinants?”

To study ncRNA function, different methods are used to assay loss and gain of function because, she says, “each method has its own



“It totally fascinated me,” says Maite Huarte (center) when John Rinn told her about his early noncoding RNA results. She joined his lab as a postdoctoral fellow. She is now a principal investigator at Cima Universidad de Navarra, where she works on ncRNAs and cancer.

limitations.” For example, gene knockouts can affect regulatory DNA sequences and therefore mask the effect of deleting the RNA. Fusions with guided dead Cas9 for gene activation (CRISPRa) or inhibition (CRISPRi) “may interfere with the chromatin environment of the noncoding gene,” she says. In principle, RNA interference (RNAi) can deplete ncRNA without affecting the chromatin environment, but it can have off-target effects. “That’s why it is important to use orthogonal methods and interpret the results rigorously,” she says.

John Rinn of the University of Colorado Boulder agrees with this orthogonal methods approach. A forthcoming paper from his lab and colleagues at Harvard University “has every kind of ‘seq in it,’” he says, such as RNA-seq, ATAC-seq and hi-C, which is used to study long-range interactions in a given genome. They assessed the mechanism of the lncRNA *Firre* and find it activates gene expression from a distance⁶.

Papers on lncRNAs are numerous, says Rinn. There may be as many as 1,500 papers just on *HOTAIR*, which lacks a phenotype in mice. With lncRNAs what is most important to him is: “What are we learning and why?” All lncRNAs may have a role, but he hopes labs in this field clarify for themselves why they care about a particular ncRNA. When they study it, the focus should be on what it does in an organism and how it might connect to human health or disease.

In his lab, he strives for “irrefutable evidence” about ncRNA functions that he believes play a role in human biology and genetic disease. *Firre*, for example, plays a role in hematopoiesis. He thinks highly of a therapeutic that targets RNA to treat spinal muscular atrophy. Rare diseases are an important research area and useful for exploring lncRNA function. One can also perturb cell circuits with RNAs. “If we know what genes an RNA regulates, we can just use that RNA itself to turn on genes or turn off genes that are important in disease,” he says. “That’s what we’re really trying to figure out now.”

Megan Linscott and her colleagues use an array of techniques to probe ncRNAs and their regulatory role across the human lifespan; this is the focus of Toni Pak’s lab at Loyola University, where Linscott is a postdoctoral fellow. Linscott studies regulatory switches that play a role in puberty and that regulate hormones. The scientists amplify miRNA through cDNA synthesis; use RT-qPCR to assess primary, precursor and mature versions of microRNAs (miRNAs); apply northern blotting to separate smaller from larger RNA fragments; and use miRNA immunoprecipitation. The lab has developed an miRNA degradation technique to assess half-life as well as the degradation products of specific miRNAs. Says Linscott, her favorite technique is poly-some profiling. It looks at actively translating

mRNAs and their associated miRNAs. It’s a way to “catch the miRNA in the action of mRNA silencing.” She would find useful to have a more robust way to track movement of miRNA within the cell in real time and without using a large tag.

Numbers challenge

To explore questions of function, there is no dearth of ncRNAs. The National Institutes of Health National Human Genome Research Institute launched the Encyclopedia of DNA Elements (ENCODE) to identify functional elements in the human genome. The project was scaled up to annotate the entire human genome and also include the mouse genome in the GENCODE project, the Encyclopedia of Genes and Gene Variants. GENCODE release 41 (GRCh38.p13) lists for the human genome 19,095 lncRNA genes, 7,566 small ncRNA genes and 54,291 lncRNA locus transcripts. GENCODE release M30 (GRCm39) for mouse includes 14,525 lncRNA genes, 6,105 small ncRNA genes and 25,419 lncRNA locus transcripts. By contrast, according to NONCODE, one of the databases on ncRNAs, the human genome has 96,411 lncRNA genes and 173,112 lncRNA transcripts. The mouse genome has 87,890 lncRNA genes and 131,974 lncRNA transcripts. The database also keeps track of ncRNAs in animals and in plants.

Overall, says Mercer, the difference between these numbers connects to the fact that a gene as a “single discrete unit” is not what genes in the genome are revealing themselves to be. The genome is a big network of transcribed, interleaved, overlapping units. Counting individual genes is thus tough, he says: protein-coding genes, alternatively spliced genes and noncoding genes. The varying numbers in online resources about ncRNA transcripts and genes might in some cases be rather arbitrary. In general, a transcript tends to mean a unique transcript, says Mercer. A protein-coding gene can comprise different transcripts given the different splice variants. “That also goes for long noncoding RNAs,” he says. *Xist*, for example is a lncRNA that comprises a number of different alternatively spliced transcripts. As Rinn explains, both with coding and noncoding genes, “a gene is the compression of all the transcripts inside that gene.” When analyzing RNA-seq data, a researcher will likely start with the gene of interest and “then you dig into which isoforms,” he says.

Chris Ponting from the University of Edinburgh and Wilfried Haerty from Earlham Institute in Norwich, UK, point out that some catalogs contain up to 270,044 lncRNA transcripts⁷. Mattick says there are hundreds of thousands of cataloged lncRNAs, dozens of databases and databases of databases. Well over



The genome is far from what was once called “islands of genes among intergenic deserts,” says Tim Mercer of the University of Queensland.

100,000 human lncRNAs have been recorded. Overall, Mattick thinks the scientific community underestimates how much of the human genome is functional. He has a long-running bet over this with Ewan Birney, co-director of the European Bioinformatics Institute. At the time the bet was made, says Mattick, Birney said he thinks less than 20% is functional, while Mattick believes the number to be higher.

In their work, says Mattick, co-author of a new book on RNA⁸, scientists might identify a ncRNA with expression that is altered under certain conditions, such as during particular stages of development or in cancer. Then they “poke at it” with small interfering RNA (siRNA) to knock down the effect that is consistent with the role it seems to play in development or cancer. With such an assay, however, determining the ncRNA’s function is mechanistically fuzzy. That’s a challenge going forward. The gold standard, he says, is to knock down an ncRNA gene and also, for example, ectopically express it and rescue the effect. This is quite difficult with lncRNAs, but has been done.

When scientists seek to knock out lncRNAs, they can perturb DNA regulatory elements. Methods exist to bypass this, says Mattick. For instance, one can knock down enhancers post-transcriptionally using siRNA. This targets only the RNA, and the RNA knockout effect becomes clear. In the ncRNA field and in science generally, people can be tempted to generalize from observations too quickly. “That explanation becomes the orthodox explanation,” he says. “And yet, it’s actually just the first interpretation.” This has, for example, happened with enhancers, which are genomic loci that control development.

In Mattick’s assessment, existing studies suggest there are around 400,000 enhancers in the human genome. “And that’s my rough estimate of how many lncRNAs there are,” he says. The major function of lncRNAs are, in his view, to guide “divide-or-differentiate decisions” during development. These genes are not producing proteins but, for instance, organizing chromatin domains in a cell and making the conformational shifts needed for transcription and splicing.

Genomes, says Mattick, are “zip files” of transcription, with many layers of information. “The human genome is incredibly information dense,” he says. ncRNAs are, for example, involved in brain development in ways yet to be deciphered. “There’s just a whole world of these things that are being produced in different stages of differentiation and development, and we’ve hardly scratched the surface of which ones do what.”

Being an RNA person

“I see myself as an RNA person,” says Gene Yeo, a researcher at the University of California, San Diego. He focuses on the way RNAs are synthesized, processed and ultimately destroyed. The principles in that realm apply to coding and noncoding RNAs.

Questions of function, says Yeo, are “when the coding and noncoding crowd or groups divide,” he says. Some labs study the proteins RNAs make and others look at those RNAs that do not make protein. But it’s also important to keep in mind, he says, a number of ncRNAs have sections that are translated into open reading frames and peptides.

He looks forward to RNA research, both coding and noncoding, to track birth, decay and destruction of RNAs, which “is still difficult to achieve at scale, at resolution and live.” Both with coding and noncoding RNAs, it has become easier to see indications of what an expressed gene could be. But measurements at scale are mainly in fixed tissues and thus cannot capture change over time. In live tissue “we can do one or two RNAs at a time,” he says. It’s also not easy to achieve subcellular resolution nor assess complicated tissues.

Yeo advises keeping in mind that beyond the around 25,000 human genes there are hundreds and thousands of alternative isoforms. “When I think about RNA, I think really isoforms,” he says. Because isoforms cannot be distinguished by in situ hybridization, “I would say we’re missing 80% of the picture,” such as subcellular effects. This adds to the live cell measurements that aren’t readily possible.

Yeo’s bias, he says, is to study neuronal RNAs. And they are on the move, which is

best tracked by imaging rather than a stack of many snapshots⁹. Such ‘stills’ can’t be collected for all transcripts at a time in both the nucleus and cytoplasm. For example, an RNA might travel one meter from its birthplace in the central nervous system to a synapse in a neuromuscular junction in a person’s leg. “The most interesting RNA localization problems are actually in these long-range cell types.”

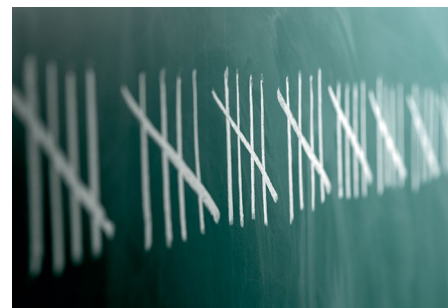
It’s a good career choice to study RNA, says Yeo. Genes are transcribed as multiple different isoforms, which vary by cell type, developmental stage or disease. For now, what these isoforms do is not yet clear. He, too, is hopeful about RNA therapeutics. RNA can be the drug substrate a drug can act upon; as in the case of mRNA vaccines, it might be the drug itself; it might involve engineered CAR-T cells in cancer. When studying the genetic and genomic basis of disease, researchers need to consider multiple types of downstream RNA expression that may or may not involve protein production.

Unlike therapeutic approaches with proteins or small molecule inhibitors of proteins such as kinases, RNA delivers sequence specificity for a target. One can potentially hit many isoforms before they are made. And unlike with proteins, the study of the 3D structure of RNAs is “just in its infancy,” he says. “I think that there’s much more room to grow with, for new ideas, in this area.”

Critical voices

Ponting and Haerty’s paper on genome-wide analysis of human lncRNAs⁷ includes “A Provocative Review” in its title. They elaborate on the pitfalls they see with experimental and computational methods on lncRNAs. They point to “questionable evidence” and “unsubstantiated claims” in the scientific literature about lncRNAs and contamination of the literature with low-quality evidence from paper-mill publications.

Some studies, says Ponting, “overstate the extent of lncRNA function.” Some papers



Estimates of the total number of ncRNAs, such as those in the human genome, vary.

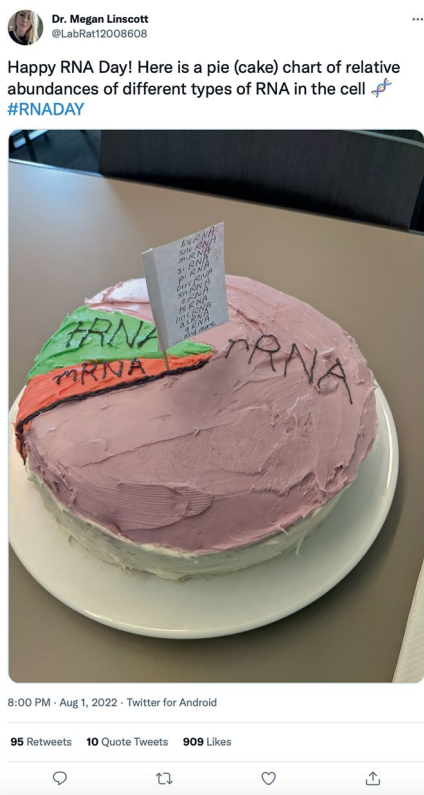
argue that lncRNAs are important as a “class” because there are tens or hundreds of thousands of them and large genomic stretches are transcribed to make them. But actually, says Ponting, human lncRNA exons span only about 2.3% of the human genome, “and their number collapses to about 20,000” when only those that are observed frequently are kept in the tally. “We want to provoke skepticism of a claim when the evidence, or its lack, demands it,” says Ponting. Just because one lncRNA has been shown to be functional doesn’t mean that all lncRNAs are functional, he says, which is the ‘lonely fact’ fallacy, one of the several “logical fallacies” they see in the lncRNA literature.

Measuring the abundance, length, complexity and localization of a lncRNA, says Ponting, is no substitute for discovering what happens to cells or organisms when one neatly perturbs it without affecting anything else. “This is extremely difficult – I’m not pretending otherwise – but it’s what we have to aspire to,” he says. When it’s not yet possible to get there, “we have to be honest and explain our experiment’s limitations and why its data might be explicable in different ways, not always in terms of lncRNA function.” Changes to an RNA in a very localized way can tweak what goes on in cells and thus help with characterizing function. But confidence in this “sequence–function” relationship is only gained by reversing the cellular effect by ‘rescue experiments,’ he says. “We didn’t want to critique the lncRNA field without saying where it could go from here,” says Ponting. The end of the paper presents suggestions on prioritizing human lncRNAs for experimentation, such as a targeted search for human disease mutations given that promoters and splice sites of evolutionarily conserved lncRNAs might have mutations that contribute to developmental disorders.

Nonlinear outlook

To those not following the ncRNA field closely, the work can appear messy, says Mattick. The genome does not have discrete linearity; it’s not one gene next to the next. That’s why he is, for example, working on computational ways to explore all of the genome’s evolutionarily conserved structures and RNA structure–function relationships. Says Mattick, “there’s a whole new world to explore here.”

Criticisms such as those from the Ponting are correct, says Guttman, in that many aspects related to ncRNAs call for a “neat theory.” Proteins are encoded by mRNAs; genes encode mRNAs; and they are highly conserved and, over long stretches of evolutionary history sequence, tend to not



To celebrate RNA Day, Megan Linscott made a cake-chart. She works on noncoding RNAs as a postdoctoral fellow in the lab of Toni Pak at Loyola University.

diverge. Almost all protein-coding genes present in humans are highly conserved across mammals – in some cases, all vertebrates. But lncRNAs aren’t conserved in this way, which can bring on the view of a lncRNA that’s “clearly not important because it’s not conserved,” he says. *Xist* is only present in placental mammals. Marsupials lack *Xist* but have X-chromosome inactivation. One big criticism of lncRNAs is the lack of evolutionary conservation. “It’s because we try to fit it into that mold,” he says.

A decade ago, says Guttman, many fundamental technologies to characterize ncRNAs, including lncRNAs, didn’t exist. Microarray experiments were used to profile conditions and perturbations. But, he says, with such experiments “you’ll never see anything that’s happening for things you are not probing.” With RNA-seq, all researchers, not just the labs focused on ncRNAs, get data inclusive of ncRNAs, which can help them to explore unanswered questions, he says. They might probe: what are the gene structures of ncRNAs? What are the context-specific isoforms,

context-specific RNAs? What’s regulated in what way and what may be important in which context, such as immunological context, stress conditions or disease?

This approach, in Guttman’s view, has shifted the challenges from data generation to data analysis. Most labs primarily focus their attention, their quantification and alignment efforts on protein-coding genes, and are not aligning noncoding transcripts. But a researcher with a given biological question – not just a person with an ncRNA focus – is bound to uncover a bundle of differentially expressed ncRNAs in his or her data and see expression patterns.

One criticism long leveled at the ncRNA field is that, given their low abundance and low expression, ncRNAs can’t be that important. But, says Guttman, his lab, Joshua Mendell’s at the University of Texas Southwestern Medical Center and others have shown is that lncRNAs “can punch above their weight,” and act in a nonstoichiometric way to amplify effects. For now, it remains challenging to study the functional roles of ncRNAs. As a community, he says, as more data are made available in public databases “we should continue to mine, extract and build out” gene models and expression patterns and other features to shine a light on function of these ncRNAs. The last 10–15 years have delivered much progress, and yet for someone not in the field it might seem hard to know how to approach ncRNAs they encounter in the course of their projects. Part of the reason to work on the community-driven paper in *Nature Reviews Molecular Cell Biology*, he says, was not to signal the community agrees on all. Rather, it’s a framework to help people think about and study ncRNAs in the future.

Vivien Marx ✉

Nature Methods.

✉ e-mail: v.marx@us.nature.com

Published online: 6 October 2022

References

- Nielsen, A. et al. *Nat. Methods* <https://doi.org/10.1038/s41592-022-01487-2> (2022).
- Mattick, J. S. et al. *Nat. Rev. Mol. Cell Biol.* (in the press).
- Frankish, A. et al. *Nucleic Acids Res.* **49**, D1, D916–D923 (2021).
- Hon, C.-C. et al. *Nature* **543**, 199–204 (2017).
- Statello, L., Guo, C. J., Chen, L. L. & Huarte, M. *Nat. Rev. Mol. Cell Biol.* **22**, 96–118 (2021).
- Smallegan, M. J. et al. Preprint at *bioRxiv* <https://doi.org/10.1101/2022.05.15.492001> (2022).
- Ponting, C. P. & Haerty, W. *Annu. Rev. Genomics Hum. Genet.* <https://doi.org/10.1146/annurev-genom-112921-123710> (2022).
- Mattick, J. & Amaral, P. *RNA: The Epicenter of Genetic Information* (CRC, 2022).
- Le, P., Ahmed, N. & Yeo, G. W. *Nat. Cell Biol.* **24**, 815–824 (2022).