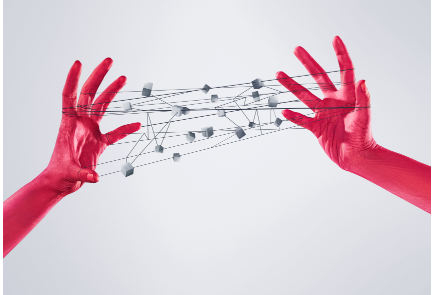
Meet some code-breakers of noncoding RNAs

Vivien Marx

The regulome—the part of the genome that regulates function—includes noncoding RNAs with varied functions yet to be deciphered.

Sometimes dogma has to go—for example, this one: DNA makes RNA; RNA makes protein. Only around 1.5% of the human genome codes for proteins, says Rory Johnson, a researcher at the University of Bern in Switzerland. Many loci generate RNAs, but not all transcribed DNA is translated into protein. Some transcripts, collectively called noncoding RNAs (ncRNAs), play regulatory roles, many of which are undeciphered. Among ncRNAs, microRNAs (miRNAs) are probably the best studied and long noncoding RNAs (lncRNAs) are the least-well understood, he says, and "of course, other classes might exist that we don't even know about."

Estimates of ncRNA numbers and species vary, says Erin Marshall, a researcher at British Columbia Cancer Research Centre, "but we can all agree, there's a lot." There are, she says, over 2,500 annotated miRNAs, more than 20,000 identified lncRNAs. The ncRNA species include lncRNAs, miRNAs, Piwi-interacting RNAs (piRNAs), ribosomal RNAs (rRNAs), Ro-associated Y, cytoplasmic RNAs, small nuclear RNAs (snRNAs), small nucleolar RNAs, transfer RNAs (tRNAs), and vault RNAs. The number of categories, and subtypes, for identified sequences might exceed the 100 mark.



Paper Boat Creative/ Stone/ Getty

There are many noncoding RNAs. Many of their regulatory roles are still undeciphered.

My name can be long

What motivates Johnson to explore ncRNAs is that "they are a huge reservoir of potential new disease genes." Beyond the numbers loom larger questions: what percentage of ncRNAs have evolved under selection, and which are functional?

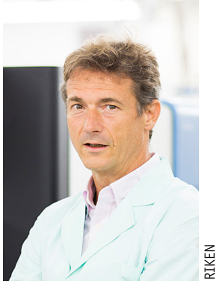
Sequences longer than 200 nucleotides are called lncRNAs but it's a name without functional implication, says Martin Turner, a Babraham Institute researcher. It's less clear how many ncRNA types there are, and classifying ncRNAs into functional subtypes gets challenging: rRNAs and tRNAs are ncRNAs; the ribonuclease angiogenin can cleave tRNA to yield biologically active fragments. The abundance of tRNA is regulated and it can affect the rate of protein synthesis's elongation phase; mRNA has noncoding sequences, 5′ and 3′ untranslated regions (UTRs), and these are invariably regulatory. These sequences can interact with RNA-binding proteins and other ncRNAs. "By extrapolation I think this is likely how many lncRNAs will also work," he says, as hubs that aggregate other molecules.

Uwe Ohler at the Max Delbrück Center for Molecular Medicine prefers three broad classes over existing ncRNA categories: (A) RNAs processed from distinct primary transcripts and with specific functions as RNAs, which include miRNAs, some lncRNAs such as Xist, which silences one of the two X chromosomes in mammalian females, and some 'competing' endogenous RNAs that might siphon a specific miRNA away from its target; (B) alternative transcript structures arising from mRNAs and ncRNAs, which are sometimes called processed transcripts and might include incompletely spliced mRNA isoforms or alternative isoforms that compete with protein-coding isoforms; and (C) unstable transcripts generated from transcriptional regulatory regions sometimes called enhancer RNAs or promoter upstream transcripts. These may have broad functions, such as helping to establish active chromatin environments.

I. Rigoutsos, Jefferson University/E. Dewalt, Springer Nature

Health and disease are likely shaped by the push-and-pull of numerous noncoding RNA regulators.

Category A, says Ohler, might number in the thousands, but there are likely tens of thousands in B and C. Categories B and C contain most of the lncRNAs, but Ohler is quite skeptical of whether they have specific functions as RNA. He and his team have observed that of around 20,000 transcripts in a cell line, only approximately 2,000 are lncRNAs, of which around one-third have a reasonable half-life. This might look different in primary cells, especially brain and testis, given



FANTOM6 is getting under way with large-scale ncRNA perturbation experiments, says Piero Carninci.

their distinct transcript repertoires. "However, it definitely makes sense to annotate all of these RNAs, and better than they currently are," he says.

NcRNAs have intrigued Ohler since his postdoc days working on alternative splicing and miRNAs. Computational biologists need to look beyond a single regulatory mechanism and gaze across different data sets, he says. NcRNAs have been, he says, "constantly good for a surprise," such as how the low transcription levels of enhancers and promoters connect transcriptional and post-transcriptional control.

Turner sees value in understanding ncRNA location in the cell, which is "a big clue to its potential function." The existing wealth of information about lncRNAs in transcriptional regulation and nuclear processes partly reflects the fact that more labs work on transcription and that many lncRNAs are found in the nucleus, he says. The roles of new types of ncRNA in translation, RNA localization, signal transduction and other cytoplasmic processes are particularly appealing to him. Understanding function will require genetic approaches and some "good old-fashioned biochemistry of protein–RNA complexes," he says.

The regulome, applied

The regulome is "the new frontier" and sometimes called "the living genome," says Howard Chang, a Stanford University School of Medicine researcher. "The regulome is where nature and nurture come together to impact health and diseases." Chang directs Stanford's Center for Personal Dynamic Regulomes (CPDR), where, according to its description, researchers seek to bridge "a deep technological chasm" between the accumulated knowledge about cells gathered in lab settings and the "inability to learn the regulatory landscapes of diseases" from clinical samples. A personalized genome is insufficient for personalized medicine; informed interventions require a "personalized understanding of the regulatory landscape of disease."

On the to-do list at the CPDR are methods, interpretative frameworks and data collection so scientists and physicians can interpret personal regulomes. To navigate the regulome will take a "GPS system" and tools for analyzing the regulome with sensitivity, speed and comprehensiveness. For example, says Chang, the tools will offer ways to find DNA switches that turn genes on and off, infer which protein or RNA factors act on those switches, or read out the variation in gene control from individual cells.

Potentially, ncRNAs represent new types of health or disease indicators. For example, by including profiles of miRNA isoforms in their analysis, Jefferson University researcher Isidore Rigoutsos and colleagues distinguished different breast cancer subtypes and, in later work, 32 cancer subtypes[1]. MiRNAs, which are around 20 nucleotides long, regulate a variety of processes, with more than one miRNA per locus, says Rigoutsos. Each miRNA locus makes multiple isoforms, and their composition and abundance will differ between tissues and people. There are tRNAs and a multitude of tRNA fragments. One person's disease subtype can, he says, be shaped by the push-and-pull of many ncRNA regulators. He and his team linked triple-negative breast cancer to two ncRNA categories, miRNA isoforms and tRNA fragments. When profiling ncRNAs in the clinic or in basic research, labs will want differentiated tallies, he says. But they need to keep in mind that current databases do not contain all of the isoforms encountered in experiments.

Annotated lncRNAs show highly specific expression patterns, says Ohler. That's also because alternative splicing and enhancers as well as transcription-associated ncRNAs are more specific than the gene they regulate. These lncRNAs indicate which regulatory regions are active. And they can be markers for aberrant cell states such as in cancer, when genomic rearrangements can lead to unique transcripts.

Marshall and colleagues profiled the small ncRNAs of the NCI-60 Human Tumor Cell Lines established at the National Institutes of Health[2]. Some ncRNAs may be tissue specific or present at certain developmental stages. The understudied piRNAs appear to play a role in cancer. The team found tissue-specific piRNA expression, an observation Marshall hopes can feed into studies of piRNA expression in cancer. More generally, when labs come across gene deregulation in cancer, she says, they should take ncRNAs into account. Therapies cannot target every protein but, she says, ncRNAs might be a way to change proteins indirectly.

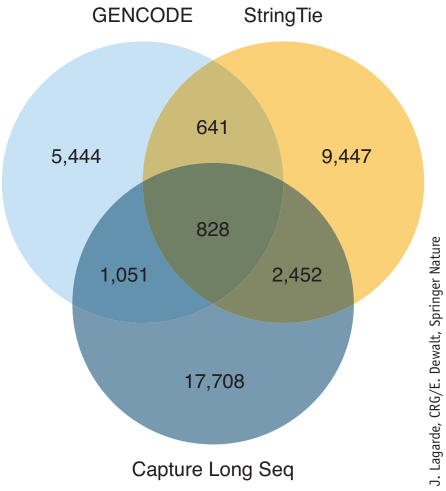
Beyond tissue-specific expression, it would help to know how ncRNAs affect their targets, to be able to consider each tissue individually and in terms of its lineage, says Marshall. Genes not typically expressed in adult cells are expressed in cancer, so identifying tissue-specific gene-regulatory mechanisms can deepen understanding of tissue biology and could uncover new disease-relevant functions. Sequencing specific tissues for ncRNAs, especially small ncRNAs, is helpful for determining tissue-specific expression, she says, pointing to other work from the Rigoutsos lab. The scientists profiled miRNAs in 13 human tissues and discovered more than 3,700 novel miRNAs. Almost all of the miRNA sequences they reported are primate specific, which means, says Rigoutsos, that experimenters "need to be doubly smart about how they are studied, because their functional impact cannot be captured by mouse models."

Annotation headache

A large variety of useful tools and annotations exist for studying ncRNAs, says Marshall. "However, consistent and frequent updates to annotation files, such as miRBase, which contains known miRNA annotations, are crucial," she says. One potentially useful resource would document tissue specificity of ncRNAs, both functionally and in annotation models. Increased sequencing depth, a view of ncRNAs across development, and tissue-specific sequencing would enhance these resources, she says. As of right now, ncRNA study is hampered by the incomplete state of ncRNA annotation.

Genome annotation "remains a headache," says Turner, also because genes have complex outputs and different cell types can have qualitatively different outputs from the same gene. For example, a gene can lead to an ncRNA; small RNAs can stem from one of the gene's introns. "If you accept that cell state will play a role on the qualitative transcriptome, then this is possibly our biggest challenge going forwards," he says.

Information on the specificity of a given ncRNA is a real bottleneck, says Ohler. Current annotations mainly provide overviews of all isoforms, but tissue-based information can be lacking. In some cases, libraries were pooled across different cell



J. Lagarde, CRG/E. Dewalt, Springer Nature

Methods find different numbers of transcript structures in the same set of GENCODE-annotated human gene loci[3]. StringTie is a short-read transcript assembler; Capture Long Seq avoids assembly.

types to annotate transcripts. A concerted effort would need to involve annotation of RNAs in a condition or tissue-specific manner and provide this information. That means "we badly need 'context-dependent' genome browsers," he says.

Annotation efforts face a "necessary compromise between throughput and quality," note the developers of a method to improve annotation, including that of ncRNAs. Johnson, Roderic Guigo at the Centre for Genomic Regulation at The Barcelona Institute of Science and Technology and colleagues developed RNA Capture Long Seq, or CLS, which links targeted RNA-capture with third-generation sequencing[3].

At issue is the challenge that manual annotation is slow and needs dedicated funding. Short-read sequencing transcriptome reconstruction delivers fast annotations, but it can lead to incomplete structures lacking terminal exons or splice junctions. This aspect hits lncRNA annotation especially hard because of the low expression levels of these ncRNAs and the few reads available for reconstruction. The community, the study authors note, is faced with a growing divergence of large-scale automated annotations of "uncertain quality," such as NONCODE's collected data, and the "highly curated, 'conservative' GENCODE collection." RACE-seq, a method involving amplification of cDNA ends followed by sequencing, helps but it's low-throughput. CaptureSeq, an RNA-capture sequencing method, helps to raise the concentration of low-abundance transcripts in cDNA libraries and is high-throughput. But, the authors note, the transcript structures "lack the confidence" for inclusion in GENCODE.

The CLS developers linked targeted RNA-capture and third-generation long-read sequencing with PacBio instruments. Their capture library was based on intergenic GENCODE lncRNAs in human and mouse tissue and also other elements such as tiled probes that targeted loci that may produce lncRNAs such as enhancers and ultraconserved elements. The scientists targeted oligonucleotide probes to these sequences. According to the team, CLS found new transcript structures in annotated lncRNA loci—for example, exons, splice sites and transcription termination sites in the SAMMSON oncogene. The method also yielded several thousand transcript models from unannotated regions that mapped to probed or unprobed regions.

The CLS developers believe their method addresses the chasm between quality and throughput related to annotation and that the quality of transcript models it delivers approaches the quality levels achieved by human annotators. To date, annotation has been either slow and expensive, because it's done manually, or else quick and inaccurate, when using RNA-seq with assembly, says Johnson. CLS, with its combination of RNA capture and long-read sequencing, gets the best of both worlds. "We hope it will allow us to radically improve annotations for the community and answer some fundamental questions about the nature of lncRNA genes and how many there are," he says. Challenges remain, such as how to cover the diversity of all cell types over development. And the capture process involves guessing where a scientist expects to find new lncRNAs.

Projects depend on annotation, says Johnson, and labs tend to assume existing annotations are correct. "But they are not," he says. They are "highly incomplete" on lncRNAs. The research community has little idea how many lncRNAs have yet to be found, and many of the known ones are annotated fragments. "If we can get better annotations, it will help us tackle the really big questions: are lncRNAs playing important roles in human disease, human evolution?"

Tools down the line

Computational analysis has been crucial in the ncRNA field, says Ohler, for annotation of transcriptionally active regions of the genome, for the assessment of potential secondary structure, for quantifying conservation of primary and secondary structure. *In vivo* structure probing will become crucial, in his view. RNAs are not naked molecules and the current algorithms that look only at RNA in isolation cannot provide the "true picture," he says.

Ribo-seq has been helpful but should be used with caution, says Ohler. It is a clean option for determining whether an RNA is in the cytoplasm and bound to ribosomes. As with transcriptional activity leading to ncRNAs of unknown function, Ribo-seq reveals many locations with apparent translation. His lab and others have been separating genuine, complete, even if sometimes short, open reading frames (ORFs) scanned by ribosomes from background signal. "But even then, there are lots of places where ribosomes appear active but don't produce anything resembling known polypeptides," he says. Many such loci may not lead to functional proteins, but rather fill other roles such as protecting ncRNAs from cytoplasmic degradation or influencing expression of the main protein-producing ORFs.

Ohler says his lab believes whole-molecule sequencing will add possibilities: the matured PacBio instruments will help simplify context-specific annotation. Theoretically, Oxford Nanopore's sequencers that also process long reads can read out nucleotide modifications, "which also has potential for a real game changer." RNA–RNA and RNA–DNA interaction protocols will possibly shed light on ncRNAs, especially lncRNAs with nuclear functions such as chromatin effects.

Ohler says ncRNA labs have two types of software tools at their disposal. Software such as Infernal helps with finding RNA structures in DNA sequence and can be used for studying ncRNAs. Such tools originated in RNA research and focus on secondary structure. Other tool types focus on gene regulation: where ncRNAs come from, how their expression correlates with other

BC Cancer Research Centre

PiRNA expression is tissue specific, an observation that can hopefully help with studying piRNA expression in cancer, says Erin Marshall.

genes and post-transcriptional regulation. RNA workbench, a project between multiple labs and which includes some of his tools, too, has set out to combine these two tool types, he says.

Large-scale projects

The GENCODE project, a large-scale project that includes the Wellcome Trust Sanger Institute, MIT, Yale University and other institutions throughout the world, is part of the larger project Encyclopedia of DNA Elements, or ENCODE. Among GENCODE tasks is the identification of noncoding elements in the human genome. The identified ncRNAs are being annotated through Ensembl's automated process and manually curated by the Human and Vertebrate Analysis and Annotation, or HAVANA, group.

FANTOM, or Functional Annotation of the Mammalian Genome, another large-scale project based at RIKEN, has embarked on FANTOM6. The teams are finalizing the pipeline for large-scale ncRNA perturbation experiments, says Piero Carninci, the project's coordinator, a researcher at the RIKEN Center for Life Science Technologies. First is RNA extraction from human cells by robots, followed by quality control steps, large-scale RNA capture, sequencing, data analysis and annotation. The experiments will be with human cells. The plan is essentially to knock down ncRNAs and see what happens. Functional studies will pursue results. The work will help to further establish the field of work on ncRNAs and contribute to reconciling the varying numbers of ncRNAs in published studies. In a publication based on FANTOM5 data[4], Carninci and collaborators generated an atlas of 27,919 human lncRNAs; the data are here.

T. Reut

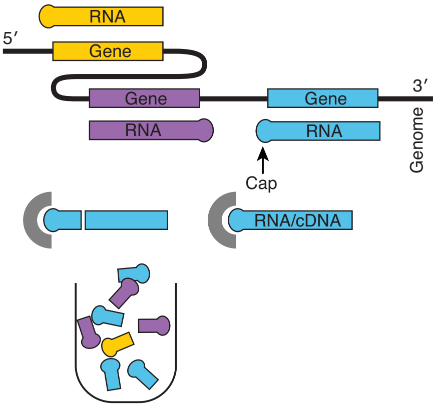
Projects rely on annotation, says Rory Johnson. But existing annotations are not always correct.

As FANTOM6 unfolds, data will be shared with the research community as quickly as possible, says Carninci. Once validated, they can be included in the GENCODE collection. The annotations need to tease out experimental observations; the teams want to link ncRNAs to tissue types, developmental stages and cell states.

FANTOM5 generated 3,000 libraries, says Carninci, and "there is a lot of diversity in these libraries. "The data are from primary cells, tissues, cancer cells, time courses of differentiated cells, induced pluripotent stem cells, immune cells. He and his colleagues believe these libraries will help researchers study ncRNA function. FANTOM applies CAGE, or cap analysis of gene expression[5], on a large scale. With ncRNAs, he says, CAGE delivers the expression profile at the transcriptional start sites; it reveals expression profiles for each promoter and shows transcription factor binding-site motifs. The method involves 'cap-trapping': the 5′ cap of mRNA is captured and avoids having plentiful rRNAs interfere with analysis. CAGE lets scientists "put a flag" on the sites that matter, he says, to focus the sequencing on that part of the transcript. It reaps information from near the promoter and enhancer sites, so researchers can look around that specific neighborhood for transcription factor binding sites and connect the results to possible mechanisms and function of transcriptional regulation.

The team has also developed and is now applying CAGEscan[6], which addresses one CAGE issue: the method helps with locating transcription start sites, but it can still be hard to connect that information to ncRNA findings. CAGE is also low throughput. CAGEscan involves paired-end sequencing at the 5′ end of cDNA-converted, capped RNAs, which makes it easier to associate promoter regions with the transcripts to which the promoter belongs.

The FANTOM teams have been using short-read technology and plan to use more PacBio sequencers to obtain longer reads, says Carninci, and increase sequencing throughput. Using PacBio instruments eases the coordination of experiments and results with the GENCODE teams, he says. FANTOM researchers are also developing a method for detecting RNA–chromatin interaction that resembles Hi-C, a tool for 4D-chromatin mapping. Also in the making: a collaboration with the Human Cell Atlas, a venture including the Wellcome Trust Sanger Institute, Karolinska Institute, the Broad Institute and RIKEN. His, and RIKEN's, goal in this collaboration is also to put ncRNAs on that map.



P. Carninci, Y. Hayashizaki, RIKEN/ E. Dewalt, Springer Nature

In FANTOM6, cap analysis of gene expression (CAGE) is being used on a large scale to study ncRNAs. With 'cap-trapping,' 5′ cap expression data are captured at transcriptional start sites.

FANTOM scientists are also exploring single-cell-level identification of ncRNAs, where the low abundance is particularly acute. "So far you cannot capture all the transcripts in each cell," says Carninci. "It's technically impossible." But it is possible to computationally infer information from data captured from some cells.

In a publication[7], the Human Cell Atlas scientists point out that the optimal amount of information to be collected from human cells, including about ncRNAs, will emerge as "a balance of technological feasibility and the biological insight provided by each layer."

With ncRNAs, scientists know they are not hunting transcriptional noise. Their research subject no longer exposes them to the doubt or ridicule of yesteryear. Many ncRNAs are likely regulatory but, says Carninci, "they are regulatory RNAs once we prove it."

Vivien Marx is technology editor for *Nature Methods* (v.marx@us.nature.com)

1. Telonis, A.A. *et al. Nucleic Acids Res.* **45**, 2973–2985 (2017).
2. Marshall, E.A. *et al. Sci. Data* **4**, 170157 (2017).
3. Lagarde, J. *et al. Nat. Genet.* **49**, 1731–1740 (2017).
4. Hon, C.-C. *et al. Nature* **543**, 199–204 (2017).
5. Shiraki, T. *et al. Proc. Natl. Acad. Sci. USA* **100**,15776–15781 (2003).
6. Plessy, C. *et al. Nat. Methods* **7**, 528–534 (2010).
7. Regev, A. *et al. eLife* **6**, e27041 (2017).