A blooming genomic desert

Vivien Marx

Identifying functional regions in genomes takes scientists beyond protein-coding regions and into stretches formerly known as genomic deserts.

Name-calling is impolite. The genomic regions between protein-coding genes have been pelted with terms such as 'barren', 'junk' or 'deserts'. This name-calling has reflected a certain level of scientific ignorance, says Chris Ponting, a computational biologist at the University of Oxford who works on genome analysis, including noncoding RNAs. The regions may well just be "barren in terms of anything we might recognize."

According to members of the Encyclopedia of DNA Elements (ENCODE) Project Consortium, around 80% of the human genome is functional. Some call this number an overestimation. To others the figure is a symptom of apophenia, a human tendency to see patterns where there are none[1]. Scientists see that gene deserts bloom with transcripts. What is less clear is what the blooming means: for example, which areas of the genome have loci for noncoding RNAs that fulfill functional roles in a cell.

Among the many types of discovered RNAs are long noncoding RNAs (lncRNAs), which extend beyond 200 nucleotides. And within this group are intergenic lncRNAs, which grab attention because their function is more likely to be independent of known protein-coding genes. But opinions differ as to which lncRNAs—or subsets of them—are functional.

To date, around 8,000 intergenic lncRNA loci have been found in the human genome[2]. As the evidence emerges and methods evolve, scientists see different ways forward in this field.

Gene deserts bloom, but researchers disagree on how much of the terrain is functionally important.

A founder

Calling these noncoding regions 'deserts' is an "unjustified preconception," says John Mattick, who directs the Garvan Institute of Medical Research near Sydney, Australia. He has been called the grandfather of the noncoding RNA field—name-calling that makes him chuckle. He is glad the era is over during which scientists entirely denied the existence of noncoding RNAs. The field has moved beyond a handful of groups and is becoming fashionable. Noncoding RNAs are "the new black," he says.

To probe the noncoding transcriptome for functionality, Mattick and colleagues, along with Harvard University researcher John Rinn and a team at Roche NimbleGen, developed and applied RNA CaptureSeq. This method combines oligonucleotide capture arrays with RNA sequencing using second-generation instruments[3]. The team achieved an over 4,600-fold sequencing depth, which defines the average number of times a base has been sequenced. The arrays allow scientists to target select intergenic regions of the genome where transcription may be rare or where transcripts are produced at low abundance.

Sequencing at high coverage gives the team the ability to focus on a subset of the genome and get high resolution. And the so-called deserts were, Mattick says, shown to be "alive with transcripts."

Mattick continues this approach and is discovering patterns of noncoding RNA related to regions associated with disease from genome-wide association studies. One of his projects is about interrogating loci associated with neurodegenerative diseases.

The human genome is "a ZIP file extraordinaire," he says, with the transcriptome offering much complexity. Mattick does not believe that noncoding RNAs are evolutionary debris, as some do.


Garvan Institute of Medical Research

The human genome is "a ZIP file extraordinaire," says John Mattick.

Just like protein-coding genes, they are subject to evolutionary selection, albeit with "different constraints," he says.

Transcriptome analysis is complex because the transcriptome can differ from one cell to the next and from one developmental phase to the next. Capturing these dynamics calls for new types of approaches in RNA sequencing to analyze genomic subsets quantitatively and qualitatively, Mattick says. He and his team are working on ways to standardize these transcriptome analyses to allow data comparison from one study to the next.

Noncoding RNAs bring to light viewpoints from the early days of molecular biology. Viewing enzymes as components and the genes as entities that encode the components was a reflection of a "deeply mechanical age," Mattick says. But not everyone shared this view: for example, biologist and Nobel laureate Barbara McClintock was suspicious of thinking about genes only as units that are protein-coding, says Mattick. "It turns out she was dead right on that." In the late 1960s, biologists Eric Davidson and Roy Britten speculated about RNA's regulatory function beyond its role as a shuttle between genes and proteins. At the time, the notions were unfashionable.

Mattick was intrigued when he first heard about introns in the genomes of eukaryotes and decided to explore noncoding RNAs, including when he went on a sabbatical at the University of Cambridge in the early 1990s. He encountered raised eyebrows and curiosity, which encouraged him to keep working to understand DNA stretches that are transcribed but not translated.

As researchers have explored the chemical versatility of proteins, they have found that RNA can also fold into structures, interact with proteins and recognize other RNAs or DNA. Mattick's hunch has been that noncoding RNAs fulfill a regulatory role, but he knew he needed data and experimental approaches that stand up to scrutiny.

Knockout experiments might show function of one or two noncoding RNAs, but he wanted a broader approach. Sequencing combined with expression analysis and targeted knockdown experiments has been a way to begin assessing functionality of these RNAs, he says.

As sequencing and structural data come together, his emerging view of lncRNAs is that they perform many functions: the molecules recognize particular sequences in DNA and possess structural elements that can recruit and target cargo proteins, such as chromatin-modifying enzymes, to precise locations, he says.

Mattick is interested in understanding the structure of the genetic information systems. The protein repertoire of organisms has remained relatively stable, but what has "blown out" with evolution, he says, is the regulatory genomic superstructure, in which much activity is transacted by RNAs.

ENCODE data reveal millions of locations in the human genome where different epigenetic marks are made in different cells and at various developmental stages. An "army" of noncoding RNAs may well be guiding the epigenetic proteins that supervise development, he says.

Epigenetics is revealing plasticity of the genome. Mattick believes that the plasticity may be superimposed on a more hardwired regulatory system that controls development. For example, various forms of RNA editing and retroposon activity enable physiological adaptation and brain function, he says. These facets of plasticity have evolved in tandem, lending organisms the ability to react to environmental changes.

Being practical

Snowboarders show a certain renegade streak when they whoop as they go down mountainsides or stairway railings in city parks. Harvard University researcher John Rinn, who is an avid snowboarder, appreciates the speed of progress in genomics. Every year feels "like a light year" with changing techniques and concepts, including a heightened interest in noncoding RNAs.


The Pop!Tech Institute

John Rinn decided to expand on the mere handful of existing lncRNA knockout mouse models.

The field might have been "hyped," which increases skepticism, says Rinn, who also has appointments at the Broad Institute of Harvard and MIT and at Beth Israel Deaconess Medical Center. When microarrays began yielding results about noncoding transcripts, he saw how distrustful colleagues were of the technology and how wary they were of artifacts. "The burden of proof needed to be higher," says Rinn.

In recent work, he and his colleagues completed a mouse knockout experiment to characterize the function of select noncoding RNAs[4]. The study ends a personal chapter in this field, says Rinn, because it delivers evidence that certain noncoding RNAs play important physiological roles. He hopes the work also ends the doubting about noncoding RNAs and their function more generally.

He had been seeking a type of experiment that could not be criticized as experiments relying on high-tech tools with artifacts have been. He decided on a lengthy and expensive approach—"Let's do something 'old school' and do some genetics," he recalls thinking. He set out to expand the mere handful of existing lncRNA knockout mouse models.

He and his colleagues at the Massachusetts Institute of Technology (MIT), Rutgers, and Regeneron Pharmaceuticals made it a multimillion-dollar, five-year project. They first selected the noncoding RNA loci, and Regeneron produced knockout mice. Rinn praises the company scientists as "risk-takers," as this work does not offer sure-fire returns on investment.

The effort focused on intergenic lncRNAs. The researchers applied a filtering algorithm to whittle down their list of lncRNA candidates. They excluded transcripts with identifiable domains indicative of certain protein families as well as those overlapping with known protein-coding genes, microRNAs, transfer RNAs and pseudogenes. They put in place assays as well as RNA-seq and computational pipelines and picked 18 loci of intergenic lncRNAs.

When the team made gene knockout mice for these 18 lncRNA loci, they found that three mutant strains did not survive, others had growth defects and suffered from issues with internal organs notably heart, lung and gastrointestinal tract, and still others had brain defects. Rinn and his colleagues believe the results show that these intergenic lncRNAs play "critical roles" in these animals' physiology. Rinn says he hopes this work will encourage other scientists to do more experiments of this type. "We now know that this work is relevant, it's not perfect, we have more work to do," he says.

One necessary task is to proceed as with protein-coding genes and look for sequence motifs and other facets to group noncoding RNAs into families. Another step, with an eye to applications, is about finding the intergenic lncRNAs most relevant to human disease. This search could include efforts to explore this "new chemical space," he says, which might lead to therapies targeting noncoding RNAs.

Rinn says scientists have been working on noncoding RNAs with hesitation. In snowboarding, he says, hesitating before a jump is bad news; instead, "you have to let go." Drawing a parallel to the noncoding RNA field, he hopes his mouse knockout experiment will remove hesitation, "allowing people to start doing cooler tricks off of jumps."

Skeptical eye

Rinn's mouse knockout experiment is "exactly what needs to be done," says computational biologist Sean Eddy at the Howard Hughes Medical Institute, Janelia Farm Research Campus. These types of experiments are challenging because functional loci do not necessarily show an obvious phenotype in a first-pass screen.

Also, a phenotype may result from a deletion that affects the organism in surprising ways. After all, he says "conserved regions are conserved because they're doing something, even if it's not making the lncRNA you think you're studying," he says. With Rinn's experimental approach, "he's got a handle on a phenotype, and he can go about showing in future work that the lncRNA is responsible for a function that, when disrupted, leads to that phenotype."

Eddy cautions scientists to not overinterpret or extrapolate from Rinn's subset of loci to the functionality of all lncRNA. There are thousands, possibly millions of lncRNAs, and claiming function for all with data from only some is problematic. For that kind of extrapolation, scientists would need to define loci and take a random subsample to uncover the proportion of "all lncRNAs" that have knockout phenotypes. All of which is a "crazy" experiment, he says, and a waste of time and money.

In his view, most annotated lncRNAs are indeed "noise and artifact." Therefore, Rinn's approach to cherry-pick the most interesting lncRNAs "is a great way to go," says Eddy. The hope is that if he finds a number of lncRNAs with similar biogenesis paths and functional properties, he will be able to define a biologically meaningful class of noncoding RNAs. Those properties can be used to sweep systematically through the genome to find others, as has been the practice for other functionally defined classes of noncoding RNAs such as microRNAs or small nucleolar RNAs, he says.

For now, lncRNAs are not yet a single, biologically meaningful class but rather a "big heterogeneous mix" sharing the traits of length and a noncoding identity. "And all self-respecting RNA zealots, like me, expect that there's functional noncoding RNAs with a variety of different functions and biogenesis pathways," Eddy says.

A colleague's catalog of lncRNAs might, upon a hard gaze, turn into a large assortment of quite different entities with varying properties, "most of which is likely noise and artifact—but not all," he says.

Among Eddy's set of computational tools are those that rely on conserved RNA secondary structure as well as sequence conservation. But RNAs do not have to function via a conserved secondary or tertiary structure. In the case of the functional lncRNAs studied to date, such as XIST, HOTAIR and others, "it's unclear yet that their functions depend on any conserved structure," he says. Claims to that effect have been made "with varying levels of evidence that aren't convincing to me yet."

Tuning, tweaking

"My view is that noncoding RNAs in general are more tweakers or tuners than they are fundamental," says Ponting. He agrees with others on the importance of working out the functional roles of noncoding RNAs, and he is pleased to see Rinn and colleagues' mouse knockout study. A challenge, however, is that the work involved deleting much DNA to create the knockouts removing "a lot of real estate" that binds transcription factors or regulates neighboring coding genes. Ponting knows that Rinn wants to show phenotypes shaped by noncoding RNA. "In my view he hasn't proved that," Ponting says.

Emily Charles Photography

Less than 5% of lncRNA sequence is functional, says Chris Ponting.

The deleted DNA will have included noncoding RNA stretches but also other regulatory sites. "At least that's a possibility," he says. Ponting has more confidence about the results from experiments that involve knockout followed by rescue, such as the work from the lab of MIT molecular biologist David Bartel. That team knocked out noncoding genes in zebrafish and then 'rescued' them by inserting human RNA loci. "That's the right experiment," Ponting says, acknowledging the difficulty of such an experiment in mice. He says that some groups have successfully performed a rescue experiment using mouse embryos derived from mutant cells containing a modified rescue bacterial artificial chromosome transgene[5].

Recapitulating functionality of gene loci in the lab is challenging, Ponting says. For example, in a knockout experiment an animal's limbs might be slightly shortened, but that might be in line with a common variation in limb length.

Ponting agrees that even some protein-coding genes do not lead to discernible phenotypes. "And yet they have survived the vagaries of time, and they stay part of the genome," he says. In his view, the criterion for functionality that all genes must pass is the test of evolution.

"Everyone wishes for the human genome to be of exceptional design," he says. But instead it remains the result of forces acting on random mutations. Unlike fruit flies, humans do not have large population sizes and cannot readily dismiss deleterious DNA changes. The amount of noncoding sequence in genomes "has ballooned over evolutionary time," but that increase is not necessarily a function of organism complexity, Ponting says.

The ENCODE data show that much sequence is functional in the sense that it is biochemically active. The entire genome is read out through transcription, replicated or touched by enzymes at some point in the cell's life cycle. "So all of it is, in some sense, biochemically functional," he says. When studying noncoding RNAs, he believes it is helpful to apply the concept of "true" functionality, which takes into account whether a noncoding RNA is of "considerable importance to the organism."

According to this criterion, not 80% of the human genome is functional, as the ENCODE Project Consortium has indicated, but tenfold less: somewhere around 8–10%, he says. These numbers are derived from evolutionary models, which are "agnostic" about whether a region codes for proteins. When it comes to lncRNAs, his view is that "less than 5%" of lncRNA

sequence is functional, a quantity that is derived from evolutionary models.

The evolutionary argument, he says, is one that most people do not believe. Instead they seek experiments that demonstrate functionality. In his own work on lncRNAs, he has found six noncoding RNAs with functions in the cell.

Ponting's interest in lncRNAs was originally piqued when he heard Mattick speak about these being among the genome's most important molecules. Ponting says, "I didn't believe it."

Intrigued, Ponting set out to look at lncRNAs using evolutionary models that compare species and, more recently, that look across the human population. He and his team have not seen evidence "for strong and pervasive selection" acting on noncoding RNA sequence in humans.

He sees several possible interpretations of his findings. Perhaps noncoding RNAs do not have strong functions. Alternatively, and he favors this option, their function is governed by a small portion of their sequence. Perhaps only 5% of a 1,000-nucleotide lncRNA may be functional. These 50 bases might act as binding sites for proteins or microRNAs. "We should not look to functionality going right across the transcript," says Ponting.

Clinical eye

Uwe Ohler, a computational biologist at Max Delbrück Center for Molecular Biology in Berlin who works on microRNA analysis, says his wife is a genetic counselor who can apply only established, clinically relevant data when helping patients. Some of this information is about noncoding mutations and mutations in introns. But the vast majority of noncoding information is not yet part of her consultation sessions.

D. Ausserhofer/MDC

With lncRNAs, functionality might be a spectrum, says Uwe Ohler.

Skepticism in RNA biology and about RNA-based gene regulation has given way to excitement about the field, he says. Ohler calls the ENCODE Consortium definition of function "very loose." He has a more conservative approach to defining function, even though he also believes that much of the human genome is transcribed.

"Transcription itself is a noisy and imperfect process," Ohler says. Fluorescent reporters and single-cell genomics are revealing more of this background. "Background doesn't mean really random," he says, because different cells and cell types react flexibly and different genes are expressed at different times. RNA production will unfold as long as it leaves the cell unharmed and could be background without its own function. "If evolution doesn't select for it," he says, "it will possibly go away again."

One challenge for the field of lncRNAs is that the RNAs themselves have a "quick and dirty definition," one that is not based on their function. And 'lncRNA' is also used as a loose term that can even refer to promoter-associated transcripts. With microRNAs, scientists discovered how the cell processes them, which helps when training computers to recognize them.

With lncRNAs, functionality might be a spectrum. For example: they might not leave the nucleus, they might just have jobs related to chromatin, or they might be expressed at low levels. Hypotheses advanced by Mattick and others are not unreasonable, and there are probably noncoding RNAs still to be detected, he says. The evidence for such hypotheses is "where it's still lacking."

He believes Rinn's mouse knockout experiment to tinker with these locations has a chance to show functional loci, but like Ponting, he is concerned over which DNA stretches may have been deleted, and he too would like to see a lncRNA-based rescue experiment to show functionality.

RNA-seq has changed the entire field, Ohler says. Beyond microarrays that let researchers capture the "steady state," RNA can be now be tracked throughout the cell, even the polymerase's movement along the genome during transcription. "You can do experiments every few minutes," he says, acknowledging that the technology also produces artifacts.

He would like to see more "sound" analyses of RNA-seq data. Too many studies perform "suboptimal analyses" with conclusions being drawn from low-quality sequence reads.

Another challenge for the field is to understand RNA-based interactions.

Thinkstock

In mouse knockout experiments, finding function is challenging because obvious phenotypes may be lacking.

Although they may be akin to a transcription factor targeting a gene, the "tricky thing is that every RNA is expressed at different levels," Ohler says. Some might be transcribed occasionally, and others can be present with hundreds of copies, making profiling difficult.

"We don't know yet at this point how strong these effects really can be," Ohler says, which renders him more cautious than Mattick. On the RNA level, plenty happens in the cell. For example, a microRNA might be sequestered from its usual target gene. New interpretations of RNA-seq data might show noncoding RNAs or might reveal isoforms of protein-coding genes.

Much clarification is needed to help the field progress, also with a view to medicine. The difference between healthy and diseased individuals might lie in noncoding regions. But variation must first be judged as clinically meaningful.

Ohler is in the process of setting up a consortium to profile gene regulation that affects human health. Right now, noncoding RNAs are not in that class. One day, he says, some of them might be shown to have clinical importance. Then they might become an everyday part of his wife's work and that of other genetic counselors.

1. Graur, D. *et al. Genome Biol. Evol.* **5**, 578–590 (2013).
2. Young, R.S. & Ponting, C.P. *Essays Biochem.* **54**, 113–126 (2013).
3. Mercer, T.R. *et al. Nat. Biotechnol.* **30**, 99–104 (2012).
4. Sauvageau, M. *et al. eLife* **2**, e01749 (2013).
5. Grote, P. *et al. Dev. Cell* **24**, 206–214 (2013).

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