A Conversation with Igor Ulitsky

Interviewer: Anke Sparmann

Senior Editor, *Nature Structural & Molecular Biology*

Igor Ulitsky is the Sygnet Career Development Chair for Bioinformatics and a Senior Scientist in the Department of Biological Regulation at the Weizmann Institute of Science.

Anke Sparmann: Your research focuses on discovering the functionalities of long noncoding RNAs [lncRNAs] and how these functions are encoded in the sequence, especially what determines their cellular localization. What got you interested in noncoding RNAs in the first place and what are the challenges that you find working with them?

Dr. Ulitsky: I first became interested in noncoding RNAs towards the end of my PhD studies. We were looking at regulatory networks governing transcriptional regulation for protein-coding genes. Then we got interested in microRNAs, so I went for my postdoc to Dave Bartel's lab at the Whitehead Institute. This was around the same time that people began doing large-scale maps and seeing that there is a lot of transcription outside the boundaries of protein-coding genes producing RNAs that seemed to be very similar to mRNAs, but it wasn't clear whether these were functional or what they might be doing.

Together with another postdoc in the lab, Alena Shkumatava, we became interested in studying to what extent long noncoding RNAs are found in different species. We compared zebrafish and human and mouse and we found that there is a lot of turnover, but there is a subset of lncRNAs that is deeply conserved. They're found throughout vertebrates and we could show that two of them were actually functionally important in development of zebrafish and that this functionality was conserved. When I started my own lab about 6 years ago, we decided to focus on understanding what kind of things these long noncoding RNAs are doing and how they're carrying out these functions.

Anke Sparmann: You're using screens, but also computational methods and evolutionary analyses to determine these functions. How do these two methods differ in what you learn from them?

Dr. Ulitsky: My background is in computer science; I did my PhD in computational biology. A lot of what we were doing at the time was based on taking a lot of knowledge that we had on proteins and trying to use networks and similarities and so on to predict protein function. lncRNAs are much more complicated because we don't have much to start with. We don't have any sort of clear "gold standard" where we can say, "We understand *these* lncRNAs, so now let's use that information to learn something about the other lncRNAs," because we don't have enough knowledge to begin with. That's why we have to do more experimental biology to try to build some initial understanding of a few select examples, so we can then go back to the more computational side and generalize it. We've been taking five or six favorite genes and really trying to nail down what they're doing and how they're doing it, while in parallel always trying to think how we can generalize from that.

We screen for fragments of the lncRNA genes or use CRISPR to screen for functionalities of genes, because eventually there are maybe 50,000 different lncRNA genes in the human genome. We don't think that all 50,000 are functional. Maybe there are a few thousand that are functional, maybe just a few hundreds, but that's still a huge diversity. While it's important to study them one at a time, we're not going to get very far very fast if we keep doing that. We always try to understand these examples, but, on the other hand, we're always thinking how we can experimentally or computationally—and ideally, both—try to generalize that and say, "Okay, we're learning something about this gene, but it's also applicable to others." We've made some progress on this, but it still remains the main challenge both for us and for the field in general.

Anke Sparmann: Proteins have functional domains that always look the same and you can predict the function just by the protein having that specific domain. In lncRNA that's not quite how it works.

Dr. Ulitsky: That's still the blueprint, though. In the last 5 years we've tried to find as many of these domains and as many lncRNAs that are behaving like that. I still think that there are some, but today we realize that this is likely a minority. If we think about a "beads-on-a-string" model where you have a long RNA that's built from these functional domains—from "beads"—the way that we like to represent proteins, this would yield certain expectations.

Published by Cold Spring Harbor Laboratory Press; doi:10.1101/sqb.2019.84.039529

We would expect it to have multiple conserved regions. We'd expect that if we look at its evolution, we're going to see preservation of much of the sequence. Even if the sequence is more flexible because structure is more important than the primary sequence, we would still expect to see that the RNA would maintain at least the general boundaries between, say, human and mouse. There are some lncRNAs that are like that; *Cyrano* and *NORAD* [noncoding RNA activated by DNA damage] behave this way. But if we look at the typical lncRNA and how it evolves and how abundantly it is expressed and what its sequence looks like, it's quite rare that we see such genes.

Again, if we go back to maybe a couple of thousand lncRNAs that are functional, how many of them are actually these RNA "machines" with multiple domains resembling ribosomal RNA? At the extreme, my guess would be that this is a relatively small percentage: maybe 5%–10%. Many of the others might have some functional RNA sequence, but a lot of the functionality is about taking that functional region and expressing it in a particular context. Where the lncRNA is expressed, how it is expressed, how much time it spends on chromatin or in the nucleus, what is happening around it, all seem to be more important than any particular combination of sequences. Still, the field's thinking about this is evolving all the time. If you ask me in another 5 years, maybe everything I just said is wrong.

Anke Sparmann: Is it more the structure or the sequence that determines whether these kinds of the domains might be functional?

Dr. Ulitsky: Thinking about noncoding RNAs is very much shaped by this idea that their structure has to be very important. Other noncoding RNAs—ribosomal RNAs, microRNAs, snoRNAs [small nucleolar RNAs], tRNAs [transfer RNAs]—all act through adopting a particular structure and through interacting with proteins that are recognizing these specific structures. If we look at how these other RNAs evolved, in most cases we can really see that there is a lot of pressure to preserve these particular structures.

If you look at conservation of lcRNAs between different species, there are some cases where we see evidence that evolution has preserved a particular lncRNA structure while changing the rest of the sequence, but these cases are quite rare. For a typical lncRNA, evolutionarily we don't really see evidence for conservation for any particular structure. For the very few that have been interrogated experimentally, there is not a lot of evidence that structure is very important. Of course, our view is very biased: One can say that structure must be important, and if we're not seeing it it's probably because we don't have the right tools.

But it's important to keep in mind that even if you take a random long RNA, it's going to fold into a structure. It's going to be GC-rich; it's going to be a stable structure. That's not to say that lncRNAs don't have a secondary structure; they do, and in many cases it's going to be a stable structure. But the evidence that these structures are actually driving their function or their recognition by other proteins in a way similar to all these other classes of noncoding RNAs is relatively limited. While there is probably some subset where structure is very important, it's likely that this fraction is relatively limited. Evolution-wise, we don't see a lot of evidence for that conservation.

Anke Sparmann: You said earlier that localization of RNAs is very important, and then you found this element that is important for RNA localization overall. How did you find that, and how do you find that it works?

Dr. Ulitsky: We became interested in localization for two reasons. We think that the functionality of many of these RNAs, which in many cases act on chromatin or regulate gene expression in *cis*, will require that the RNA needs to stay in a particular place. There is not a lot known about where RNAs spend their lifetime and what determines the distribution of RNA in the cell. This is relevant for the lncRNAs, but it's increasingly appreciated that it's very important for mRNAs as well. The typical model used to be that RNA is made and then needs to be exported very quickly to be translated in the cytoplasm. Today, from a lot of RNA-seq studies in various different systems, it's clear that while most mRNAs do that, there's also a variety of mRNAs that for different reasons actually stall in the nucleus for quite some time, awaiting a certain signal or stimulus that then allows them to rapidly export an already-made RNA and translate it in the cytoplasm. There was a lot of interest in what determines, for a given RNA, how long it's going to stay on chromatin in the nucleus versus in the cytoplasm.

The other motivation is that, if you're thinking about screening for functional elements in lncRNAs, the ability to do subcellular fractionations efficiently and reproducibly allows for a very easy screening system where we can test different sequences and compare them to see which sequences carry out a certain function: that function being to either keep the RNA in the nucleus or export it to the cytoplasm. Once we've used this screen to find these elements, we can go and look at other more complex, and possibly more interesting, functions.

We decided to take nuclear lncRNAs—long RNAs that we know stay in the nucleus—and then take short fragments of these RNAs and place them in the context of an RNA that in regular conditions is very efficiently exported. For example, we take the long RNA of GFP [green fluorescent protein], which is typically efficiently exported to the cytoplasm, and we test which short fragments of about 100 bases can act as a kind of brake: to see when, once we insert a specific sequence, the RNA is no longer efficiently exported to the cytoplasm and spends more time in the nucleus.

Once we identify these elements, the real power of this approach comes not necessarily in identification of the sequence, but in the fact that you can take that sequence and systematically mutate it. DNA synthesis makes it easy for us to generate thousands of such fragments that are very similar to each other but that differ by only one base or two or 10. This way, we could really narrow it down and identify a specific region that is important, iden-

tify the protein that is binding that region, and study what else about the sequence is important.

Once we have identified a functional element, we can look at whether RNA structure is important for that functional element. In this case, we don't really have evidence that structure is important but we can ask questions like, "What happens if, instead of binding in this particular place, we now move the binding site three or four or five or 10 bases sideways," and so on. We can do this reproducibly across hundreds of thousands of different variants. We're also taking this approach to ask other questions. Instead of localization, what if we look at stability? What if we look at functionality? What if we measure an actual phenotype of lncRNA by comparing many different variants at the same time? This also got us interested in studying some other factors that are influencing the distribution of the RNA within the cell.

Anke Sparmann: If you move the element, do you really see differences?

Dr. Ulitsky: Yes. It appears to matter quite a lot, both where it is and its broader context. This particular element is about 40 bases long—and now we can narrow it down to about 20 bases—and appears to be much more effective when it's found in the internal exons of an RNA rather than in terminal exons, which is especially surprising. If you think about the classical models of mRNA localization, typically these elements are found in the 3′ UTR [untranslated region] that we tend to think of as a reservoir of various motifs that are going to influence things like the stability and the localization of the RNA. In this case, we see that this element is actually preferentially found in internal exons. In mRNAs it's typically part of the coding sequence, presumably because it interacts or somehow is influenced by the context of where it is acting.

We're also trying to see what happens if we take that element and put it in a different context altogether. If you take a different mRNA instead of the GFP that we've been using until now, we see some interesting differences. Context matters in this case, presumably because we have an HNRNPK [heterogeneous nuclear ribonucleoprotein K] protein that needs to bind a particular place, but other factors around it are helpful and they can quite dramatically influence the activity of it. We're seeing that within this short sequence element, HNRNPK also needs to bind to a very particular place. There is maybe one alternative place we can find where it binds, but in most other places —even if we take a strong binding site for that protein— it's not going to work. There are only two particular slots where it can be effective.

Anke Sparmann: What do you think will be the next wave of new discoveries in this field?

Dr. Ulitsky: It's hard to say. The field is maturing gradually. We have a large number of lncRNAs for which we have some evidence that they are functional. There are some phenotypes in cells, and an increasing number where there are phenotypes in model organisms. Also increasing, although slowly, is evidence that human diseases are affected. The big challenge now is to figure out how many different mechanisms these are adopting. To what extent are these mechanisms really similar one to another? We have a "box" of the current mechanisms and some of them have been around for quite some time, but I really think that some of these mechanisms are going to be outside of that box. The challenge is to figure out what else is out there in terms of what kinds of things, mechanistically, a long RNA can do within a human cell.