The Role of RNA Research in Community Health: Know the Fundamentals when Seeking the Future
By Omar Hedaya

It wasn't so long ago when the consensus around the human genome was that around ninety percent of it is "junk" DNA – DNA whose fate led to either useless transcripts or no transcripts at all and, therefore, was of no function. Of course, we now know this to be false, and research into "junk" DNA has since added another plentiful list of biology acronyms for us to remember. Add to this the recent developments in sequencing technologies and data science, and now phrases like "personalized medicine" and "precision therapy" seem stuck to the tip of everyone's tongue.

At the center of much of this excitement is the promise of therapeutics that either target and/or utilize RNA. An extensively reported hope in these times of COVID-19 is an RNA-based vaccine against SARS-CoV-2 currently in clinical development by Moderna and Pfizer, to name a few companies. Another breakthrough therapeutic that is already FDA-approved is Spinraza, an antisense oligonucleotide that targets the precursor to messenger RNA (pre-mRNA) deriving from the *SMN2* gene to improve spinal muscular atrophy disease outcomes in affected children. The economy has seen a dramatic increase in RNA-related research investments, totaling several billions of dollars invested in the last five years alone [1].

It seems inevitable that research into RNA will explode as investments increase and the technologies used to study RNA improve. What will this new frontier of RNA research look like, and what types of questions should we be asking about RNA so that we can maximize its therapeutic potential not only as a target but also as a tool? To answer this, it's important not only to look towards the newest breakthroughs in RNA research, but we have to remember the fundamentals, and we must think critically about all that we don't yet know.

One fundamental paradigm is the realization that as organisms became more complex through evolution, the number of protein-coding genes did not always scale with complexity, possibly reflecting a cost to having more genes [2]. For example, the simple nematode *Caenorhabditis elegans* has around 20,000 protein-encoding genes – almost the same number of protein-coding genes as in the human genome. Many of these genes are even orthologous to human genes, and some examples relevant to my own research include those encoding the cardiogenic transcription factors GATA4, MEF2C, and TBX5. These factors are essential to turn a stem cell into a beating cardiomyocyte. While they are highly conserved between us and *C. elegans*, the difference between our four-chambered heart and the nematode tube-like heart is significant. By eye, we can clearly discern the difference between a human and a worm, but that difference is less clear when we look at the genes that we have in common. The question that arises is: what makes us and a worm visually and functionally different?

Certainly, the answer lies partly in the distinct transcribed regions of the human and worm genomes that reside upstream, downstream, or within protein-encoding genes and the ways in which these regions are regulated and/or regulatory. For instance, *C. elegans* has a 100 million basepair (bp) genome, around 25% of which codes for proteins. In contrast, humans have a 3 billion bp genome, of which only approximately 1% encodes proteins. It seems that mammals have evolved more non-coding RNA sequences that are less conserved across species than their coding counterparts. For example, the median length of a 3′-untranslated region in *C. elegans* is 140 nucleotides, whereas in humans, it is 1200 nucleotides [3]. The worm has a median number of around three introns per gene with a median length of 76 bps, while we have around eight introns per gene with a median length of 938 bps [4,5]. These differences, which result in greater transcriptome diversity in humans, offer fertile ground to give rise to novel protein variants with specialized functions alongside novel RNA-mediated regulatory mechanisms exclusive to us.

One prime example is nonsense-mediated mRNA decay (NMD), an RNA surveillance mechanism whereby mRNA harboring a premature stop codon due to a genetic mutation or altered pre-mRNA processing is marked for degradation. Although highly conserved across eukaryotic evolution, loss of a key NMD regulatory protein is lethal in human and mouse, whereas loss is tolerated in organisms such as *C. elegans* and the yeast

Saccharomyces cerevisiae [6,7]. Data indicate that increased complexity in organisms such as humans is a consequence of the evolution of new biological processes. Understanding how molecular mechanisms have become interwoven in highly complex organisms such as ourselves will give us better resolution when seeking to understand diseases, ultimately helping us design novel therapeutics with improved efficacies.

Another worthwhile paradigm is exemplified by our non-coding transcriptome that is yet to be functionally understood. The human ENCODE project estimates that around 80% of the genome is transcribed, most of which is *non* protein-coding [8]. The human genome is also replete with interpersonal variations within these non-protein-coding regions that could potentially affect our health and thus, require our scrutiny. This is demonstrated by a study that identified 2516 cancer-associated single nucleotide polymorphisms (SNPs), many of which were located within intergenic sequences as well as sequences upstream and downstream of genes, none of which were known to be regulatory [9]. Transcripts arising from genomic regions that do not encode proteins can potentially participate in cell-specific biological processes. A study characterizing 849 non-coding RNAs in mouse brains using *in situ* hybridization showed that many RNAs are found in specific cell-types. In fact, RNAs of different sizes that are transcribed from the same DNA locus, and in some cases from one or the other strand of the DNA locus, can end up being expressed in completely different cells. In one case, localization of the sense and its antisense counterpart manifested clearly two distinct streaks of cells within the cerebellum [10]. Some of these RNAs have been shown in other studies to function in critical cellular roles, such as influencing chromatin modifications or playing a structural role through phase separation [11,12].

It is also important to recognize that many primate-specific non-coding RNAs have been identified, but have yet to be studied [13]. Notably, the relationship between these RNAs and diseases remains largely unknown: most studies perform RNA-sequencing using whole tissues, which precludes detecting differences manifested by different cell-types because of averaging transcript abundance across multiple tissue-types and/or dilution of transcripts that are not widely expressed. Thus, we are unable to determine if abnormalities in, for example, cell-type-specific RNA expression patterns in the brain are responsible for neurological disease.

The speed at which RNA research has grown is remarkable, and it will be exciting to see how the field develops in the next few years and decades to come. Nevertheless, amidst the shiny and impressive imagery of genetic therapies and novel medicines, it is key to remember the importance of exploring the literature from decades ago when applying the tools of today. As one example that pertains to the utility of gene or RNA editing using guide RNAs, there is still much to learn about the rules for engaging CRISPR-Cas methodology before it can be useful as a therapeutic. As another example, we will not cure new diseases simply by sequencing the genomes of patients and sitting on millions of terabytes of data: efforts in genome sequence will be futile if sequencing changes cannot be ascribed a biological consequence. In other words, it is one thing to identify the sequence variations in a patient's genome, but yet another to recognize which variations have pathological relevance and could present a therapeutic target. Mining the uncharted territories of non-coding RNAs for function has the potential to yield many insights into how humans came to be, the origin of human diseases, and new disease therapies. There is still much to learn.

References

1 Wang, F., Zuroske, T. & Watts, J. K. RNA therapeutics on the rise. *Nat Rev Drug Discov* **19**, 441-442, doi:10.1038/d41573-020-00078-0 (2020).
2 Mattick, J. S. RNA regulation: a new genetics? *Nat Rev Genet* **5**, 316-323, doi:10.1038/nrg1321 (2004).
3 Mayr, C. Evolution and biological roles of alternative 3'UTRs. *Trends Cell Biol* **26**, 227-237, doi:10.1016/j.tcb.2015.10.012 (2016).
4 Ragle, J. M. *et al*. Coordinated tissue-specific regulation of adjacent alternative 3' splice sites in *C. elegans*. *Genome Res* **25**, 982-994, doi:10.1101/gr.186783.114 (2015).
5 Sakharkar, M. K. *et al.* Distributions of exons and introns in the human genome. *In Silico Biol* **4**, 387-393 (2004).
6 Johns, L. *et al*. *Caenorhabditis elegans* SMG-2 selectively marks mRNAs containing premature translation termination codons. *Mol Cell Biol* **27**, 5630-5638, doi:10.1128/MCB.00410-07 (2007).
7 Azzalin, C. M. & Lingner, J. The human RNA surveillance factor UPF1 is required for S phase progression and genome stability. *Curr Biol* **16**, 433-439, doi:10.1016/j.cub.2006.01.018 (2006).

8 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).
9 Freedman, M. L. *et al.* Principles for the post-GWAS functional characterization of cancer risk loci. *Nat Genet* **43**, 513-518, doi:10.1038/ng.840 (2011).
10 Mercer, T. R. *et al*. Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci U S A* **105**, 716-721, doi:10.1073/pnas.0706729105 (2008).
11 Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106**, 11667-11672, doi:10.1073/pnas.0904715106 (2009).
12 Polymenidou, M. The RNA face of phase separation. *Science* **360**, 859-860, doi:10.1126/science.aat8028 (2018).
13 Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **22**, 1775-1789, doi:10.1101/gr.132159.111 (2012).