CellPress

Capturing a Long Look at Our Genetic Library

Julien Lagarde[1,2] and Rory Johnson[3,4,*]
[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain
[2]Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain
[3]Department of Medical Oncology, Inselspital, University Hospital and University of Bern, 3010 Bern, Switzerland
[4]Department of Biomedical Research (DBMR), University of Bern, 3008 Bern, Switzerland
*Correspondence: rory.johnson@dbmr.unibe.ch
https://doi.org/10.1016/j.cels.2018.02.003

Long-read sequencing, coupled to cDNA capture, provides an unrivaled view of the transcriptome of chromosome 21, revealing surprises about the splicing of long noncoding RNAs.

In his story "The Library of Babel," Jorge Luis Borges imagines a library of books with every conceivable permutation of letters. Every story told, or to be told, is found there. Similarly, our genome contains every gene and transcript, coding and noncoding, to be expressed during the human lifetime. But our catalog of this genetic library remains unsatisfactory—our books miss entire chapters, and many are completely unaccounted for.

In this issue of *Cell Systems*, Mercer, Mattick, and colleagues mark an important step in overcoming this by reporting a deep survey of the transcriptome of long noncoding RNAs (lncRNAs) and mRNAs on human chromosome 21 (Chr21) (Deveson et al., 2018). This is made possible by coupling two powerful techniques: cDNA capture and sequencing on a Pacific Biosciences (PacBio) instrument. This work joins several recent studies harnessing the unrivaled power of third-generation single-molecule sequencing to accurately survey the transcriptome (Sharon et al., 2013; Lagarde et al., 2017). This technology frees us from our dependence on short-read technology and opens fundamental questions about lncRNA biology.

Until now, researchers have relied heavily on assembly of short reads from Illumina-based RNA sequencing (RNA-seq) experiments to map lncRNAs. Programs such as Cufflinks have enabled labs to create catalogs in their favorite cell type or organism (Trapnell et al., 2010). But accurately assembling the exons of long transcripts from much shorter reads is a daunting algorithmic challenge. Sensitivity is low (entire genes are missed), false positives are frequent (i.e., nonexistent transcripts are assembled), and almost all transcripts fall short of the true 5′ and 3′ ends (Lagarde et al., 2017; Steijger et al., 2013). This is particularly acute for lncRNAs due to their low expression and sparse read coverage. Nevertheless, the canon of lncRNA knowledge rests largely upon these catalogs.

Deveson et al. and others have realized that long-read technology can overcome these issues. It can confidently report exon connectivity, while 3′ ends are identified by encoded polyA tails, and 5′ are also frequently reached (Lagarde et al., 2017; Sharon et al., 2013). However, the low sequencing depth of PacBio (~50,000 reads per lane versus ~300 million for Illumina), coupled to lncRNAs' low expression, introduces a new challenge (Sharon et al., 2013).

To address this challenge, Deveson et al. coupled long-read sequencing to cDNA capture. The latter method, pioneered by the authors themselves, focuses sequencing firepower onto known or suspected RNA-producing loci – particularly valuable for low-expressed lncRNAs (Mercer et al., 2014). Using oligonucleotide capture, cDNA libraries are first enriched for regions of interest—here, the entire human Chr21, representing 1.5% of the genome (Deveson et al., 2018). Captured cDNA, of which >70% originates from Chr21, is then sequenced by both long- and short-read technologies. In this way, the length of PacBio is harnessed to map exon connectivity, while deeper short reads can accurately quantify expression and splicing. A similar approach was recently employed by the GENCODE consortium to improve annotation of known lncRNAs (Lagarde et al., 2017).

By coupling third-generation sequencing to cDNA capture, the authors produce one of the deepest ever transcriptional maps of a human chromosome. They report a dataset of 387,029 reads from K562 cells and testis, revealing altogether 1,589 lncRNA transcript models. Approximately half of identified exons are novel. Simulation experiments suggest that, at least in the tissue panel sampled, the number of cataloged exons is approaching saturation.

These confident transcript models enable us to revisit old questions about lncRNAs, as well as formulate completely new ones. For example, we can ask to what extent lncRNA genes or products differ from protein-coding genes—the answer seems to be surprisingly little (Lagarde et al., 2017). In terms of mature length, or promoter chromatin, previously observed differences don't hold up to scrutiny afforded by full-length structures. In addition are myriad instances where incorrect gene annotations are extended to their full length or when separate fragmentary annotations are united to form a single, correct gene model (Lagarde et al., 2017; Deveson et al., 2018).

Deveson et al. extend these observations to splicing. It is known that lncRNA splicing exhibits some distinct properties compared to coding genes: it is less efficient (Tilgner et al., 2012), and their splice sites are less conserved (Nitsche et al., 2015). But now, Deveson et al. have identified another potentially more interesting hallmark of lncRNA: high rate of alternative splicing. Comparing the "percent spliced in" (PSI), a measure of frequency with

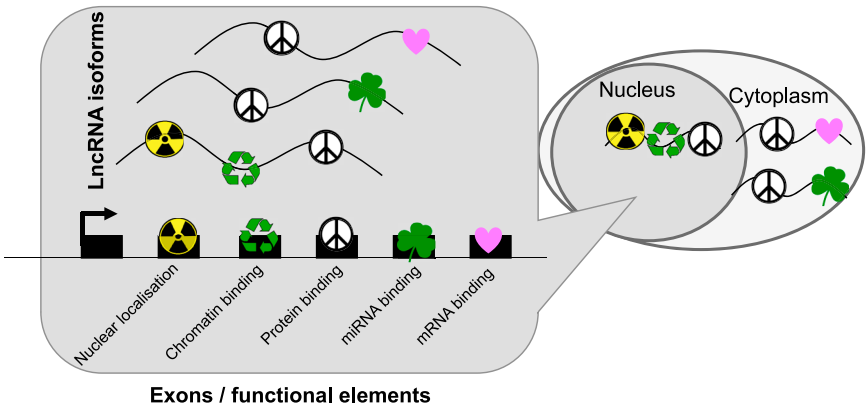

Figure 1. Generating Long Noncoding RNA Diversity
Deveson et al. propose that extensive alternative splicing may generate long noncoding RNAs (lncRNAs) with diverse functions through the differential inclusion of modular elements mediating nuclear localization or chromatin-, protein-, miRNA-, or mRNA-binding.

which exons are included in transcripts, they show that lncRNA exons tend to be alternative far more than those of coding genes. In other words, a given lncRNA transcript tends to choose just a subset of available exons. The authors name this "universal alternative splicing."

Universal alternative splicing, if validated, could have profound implications for our understanding of lncRNA functions. LncRNAs are thought to be modular: composed of combinations of functional elements and analogous to protein domains (Guttman and Rinn, 2012). There is growing interest in identifying such elements, but so far, relatively few are known (Marín-Béjar et al., 2017). The differential inclusion of such elements through exon splicing could be a mechanism of producing lncRNAs with diverse functions (Figure 1). Indeed, without the constraint to maintain an open reading frame, lncRNAs could exploit this mechanism more freely than mRNAs.

Of course, as is often the case for lncRNAs, one can interpret most observations equally as evidence for function or the lack of function. Does widespread alternative splicing reflect modularity, or simply relaxed constraint? This joins other features of lncRNAs—such as tissue-specific expression, nuclear localization, and lower evolutionary conservation—as features that can be interpreted in polar opposite ways. While Deveson et al. articulate an attractive argument for a "functionalist" interpretation, we would argue that one should adopt non-functionality as a null hypothesis to be falsified. The authors also showed that low PSI is a property of other transcribed noncoding sequences, namely untranslated regions, suggesting that splicing constraint is relaxed when an open reading frame is not present. Nevertheless, natural selection tends to make good use of available biological variation. It is entirely likely that, once better maps of lncRNA elements become available, compelling examples of alternatively spliced isoforms with demonstrably different functions will be uncovered.

On a more practical level, if lncRNA splicing really is as complex as suggested, it will have quite serious ramifications for how we annotate these genes. Is there value to users in individually annotating a vast assembly of splice variants? Or will we have to find more economical and abstract ways of annotating the splicing structure of a lncRNA?

Either way, long-read sequencing will likely lead to rapid improvements of lncRNA (and even protein-coding) gene annotations in coming years. Researchers will no doubt have to revisit more long-held assumptions about lncRNAs and re-quantify old short-read RNA-seq datasets using these new annotations. Differential gene expression studies can be carried out to find new targets in old data.

PacBio technology still suffers from several drawbacks that limit its usefulness in mapping lncRNAs. These include its high cost, low throughput, and cDNA read lengths that still do not exceed ~3 kb (Lagarde et al., 2017; Sharon et al., 2013). On the horizon is a technology that promises to resolve all these issues: direct RNA-seq by nanopore. The MinION from Oxford Nanopore Technologies offers direct RNA-seq with essentially no length limit (Garalde et al., 2018). Now, the race is on to apply this approach to lncRNAs.

By containing every possible book, Borges library held vastly more nonsense books than meaningful ones, including every possible error-containing version of any real book. The challenge for us now is to understand whether this applies to splicing of lncRNAs. Are they Borgian nonsense produced in an absence of selective pressure? Or a powerful mechanism for generating functional diversity through combinatorics?

ACKNOWLEDGMENTS

R.J. is supported by the Swiss National Science foundation through the National Centres for Competence in Research (NCCR) "RNA & Disease" and by the Medical Faculty of the University of Bern. J.L. is supported by the National Human Genome Research Institute of the US National Institutes of Health (grant U41HG007234).

REFERENCES

Deveson, I.W., Brunc, M.E., Blackburn, J., Tseng, E., Hon, T., Clark, T.A., Clark, M.B., Crawford, J., Dinger, M.E., Nielsen, L.K., et al. (2018). Universal alternative splicing of noncoding exons. Cell Syst. *6*, this issue, 245–255.

Garalde, D.R., Snell, E.A., Jachimowicz, D., Sipos, B., Lloyd, J.H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., et al. (2018). Highly parallel direct RNA sequencing on an array of nanopores. Nat. Methods. Published online January 15, 2018. https://doi.org/10.1038/nmeth.4577.

Guttman, M., and Rinn, J.L. (2012). Modular regulatory principles of large non-coding RNAs. Nature *482*, 339–346.

Lagarde, J., Uszczynska-Ratajczak, B., Carbonell, S., Pérez-Lluch, S., Abad, A., Davis, C., Gingeras, T.R., Frankish, A., Harrow, J., Guigo, R., and Johnson, R. (2017). High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. Nat. Genet. *49*, 1731–1740.

Marín-Béjar, O., Mas, A.M., González, J., Martinez, D., Athie, A., Morales, X., Galduroz, M., Raimondi, I., Grossi, E., Guo, S., et al. (2017). The human lncRNA LINC-PINT inhibits tumor cell invasion through a highly conserved sequence element. Genome Biol. *18*, 202.

Mercer, T.R., Clark, M.B., Crawford, J., Brunck, M.E., Gerhardt, D.J., Taft, R.J., Nielsen, L.K.,

Dinger, M.E., and Mattick, J.S. (2014). Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. Nat. Protoc. *9*, 989–1009.

Nitsche, A., Rose, D., Fasold, M., Reiche, K., and Stadler, P.F. (2015). Comparison of splice sites reveals that long noncoding RNAs are evolutionarily well conserved. RNA *21*, 801–812.

Sharon, D., Tilgner, H., Grubert, F., and Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. Nat. Biotechnol. *31*, 1009–1014.

Steijger, T., Abril, J.F., Engström, P.G., Kokocinski, F., Hubbard, T.J., Guigó, R., Harrow, J., and Bertone, P.; RGASP Consortium (2013). Assessment of transcript reconstruction methods for RNA-seq. Nat. Methods *10*, 1177–1184.

Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T.R., and Guigó, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res. *22*, 1616–1625.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.