Gene regulation by long non-coding RNAs and its biological functions

In the format provided by the authors and unedited
Supplementary Box 1. Long non-coding RNA stoichiometry

The abundance of a particular lncRNA species could be a key indicator of the mechanism by which it exerts its function, thereby requiring careful evaluation.

Different mechanisms, different stoichiometry requirements

In principle, a small number of molecules of cis-acting lncRNAs should be sufficient to efficiently act on a single target-gene locus. On the other hand, trans acting lncRNAs, including lncRNAs with structural roles, or lncRNAs that regulate the activity of proteins or of other RNA species through direct interactions, may require a higher number of molecules to achieve a significant biological outcome, as shown in several studies mentioned in the main text\(^1\)\(^-\)\(^3\) (see also Fig. 5a, 5c). For instance, in HeLa cells, there are \(\sim36\) copies of \(PNCTR\) and \(\sim286,000\) copies of the protein PTBP1. Each \(PNCTR\) contains a total of 2,178 YUCUY or YYUCUY sequences for PTBP1 binding\(^4\)\(^-\)\(^5\). Thus, \(PNCTR\) can theoretically sequester 7.12% to 27.44% of cellular PTBP1\(^4\). Similar stoichiometry analyses were also performed for \(h\)\(FAST\)\(^4\)\(^-\)\(^6\), PWS region \(s\)no\(\)-lncRNAs\(^7\) and \(\)SPAs\(^2\). The stoichiometry has also been carefully examined for competitive endogenous RNAs (ceRNAs). Experimental evidence for such ceRNA–mRNA crosstalk was initially described for the tumor-suppressor \(PTEN\), which is regulated by the abundance of its pseudogene (\(PTENP1\))\(^8\). Given the pervasive transcription and processing of the mammalian genome, which gives rise to different type of ncRNAs, including circular RNAs, the ceRNA theory involves the crosstalk between complicated regulatory networks containing different types of ncRNAs\(^9\). Evaluating the ceRNA model using the quantitative study of the miR-122 network as example\(^10\), the authors concluded that the modulation of a single miRNA target is not sufficient to cause detectable changes in the level of expression of the other targets of the same miRNA, indicating that simple RNA-to-RNA competition may not account for the phenotypic result of the perturbation of a ceRNA, at least with the expression levels observed in tissues in homeostasis\(^10\).

Between the highly expressed trans-acting lncRNAs, and the lowly expressed cis-acting lncRNAs, an intermediate scenario of stoichiometric requirement is that of cis-acting lncRNAs that spread along large regions of a chromosome, such as \(Xist\)\(^11\). Estimates of \(Xist\) expression are in the range of \(\sim100\) to \(\sim1000\) copies per cell, which is consistent with the contact of several \(Xist\) molecules with multiple loci and subsequent coating of the transcriptionally inactive X chromosome. Nuclear organization has been shown to play a critical role in facilitating the interactions between \(Xist\) and several loci on the X chromosome. Interestingly, HiChIRP analysis (a method that determines the chromatin
interactions of an RNA of interest in conditions that enable the interrogation of chromatin 3D architecture) uncovered that the lowly expressed immune regulatory lincRNA-EPS (~10 copies/cell) is able to contact and regulate hundreds of loci by taking advantage of particular 3D chromatin structure and epigenetic complexes that amplify lncRNA regulatory effects\textsuperscript{12}. These studies highlight the importance of spatial proximity as a key determinant of interaction specificity.

**Average vs. local lncRNA expression**

The occasional apparent discrepancy between the molecular stoichiometry and the biological outcome caused by lncRNA perturbations underlies interaction dynamics and regulatory loops that are still not well understood. Nevertheless, it is clear that lncRNA abundance measured as average expression in a population of cells does not necessarily correlate with the lncRNA concentration, which is the determinant of the probability of interaction between lncRNA and its partners. The average quantification does not consider cell-to-cell variability, which can be high for individual lncRNAs, especially in\textit{ vivo}\textsuperscript{13}. Moreover, proteins, chromatin and RNA can form local environments such as in membrane-less organelles, which limit the free diffusion of molecules and favour biochemical interactions. Accordingly, phase separation has been proposed to be utilized by \textit{Xist} to perform the silencing of X chromosomes\textsuperscript{14}, and has been involved in the assembly of paraspeckles by key domains of \textit{NEAT1-long}\textsuperscript{15,16}. A more detailed study of subcellular lncRNA interactions and localization can be performed with high-resolution microscopy and with biochemical approaches such as APEX-seq (a method of RNA sequencing based on direct-proximity labeling of RNA using the peroxidase APEX2 fused to a cellular protein of interest), which allow the spatial study of the transcriptome in different subcellular compartments\textsuperscript{17}. A deeper understanding of lncRNA interaction dynamics will shed more light onto the stoichiometric requirements of lncRNA function.

**Supplementary Box 2. lncRNAs in the cancer genomes**

The unprecedented accumulation of sequencing data has opened the doors for a systemic study of the cancer genome and its highly dynamic plasticity. With the development different genome scale projects such as The Cancer Genome Atlas (TCGA), the Cancer Cell Line Encyclopedia (CCLE), or the International Cancer Genome Consortium, it has become possible to integrate genetic, epigenetic, and transcriptional alterations associated to cancer progression as well as clinical and therapeutic outcomes from thousands of cancer biopsies. This explosion of available source data has allowed to accumulate evidences for lncRNA dysregulation at several levels, included alteration of gene
expression, copy number, chromatin organization and 3D structure, ultimately linked to the acquisition of malignant phenotypes.

Transcriptomic studies have unveiled the particularly high number of IncRNAs expressed in cancer cells, indicating that the expression of IncRNAs is defining of distinct cancer subtypes. Although the numbers may differ between studies depending on the methodology applied for IncRNA annotation, it is clear that cancer genomes underlie a high transcriptomic complexity, perception that grows with increasing sequence depth and number of samples analyzed\(^\text{18}\). The analysis of poly A+ RNA-seq from more than 7,000 samples, including tumors, tumor cell lines and normal samples, identified 58,000 distinct IncRNA genes expressed in tumors. The expression of 8,000 of these IncRNAs is associated with a lineage or cancer type, underlining the specific linkage of some IncRNAs to cancers of different tissue origin\(^\text{19}\). Moreover, specific IncRNA expression profiles are linked to clinical parameters such as the metastatic potential, patient survival or immune infiltration\(^\text{20}\). These observations put forward the value of IncRNAs as prognostic markers.

At the DNA level, several oncogenic IncRNAs have been identified due to their gene copy number alteration in the tumors of specific cancer types, such as SAMMSON (*survival associated mitochondrial melanoma specific oncogenic non-coding RNA*)\(^\text{21}\), ALAL1 (*amplified in lung adenocarcinoma long noncoding RNA 1*) or FAL1 (*focally amplified IncRNA on chromosome 1*)\(^\text{22}\), frequently amplified and with cancer-promoting roles in melanoma, lung adenocarcinoma and epithelial tumors respectively. Similarly, the significant number of mutations that lie outside protein coding regions has motivated the search for frequent mutations within IncRNAs that may account for cancer hallmarks, now possible thanks to the availability of whole genome sequences of cancer samples. Unexpectedly, the analysis of 2658 cancer whole genomes by the PCAWG (*Pan-Cancer Analysis of Whole Genomes*) Network only predicted a small number of IncRNA drivers -defined as genes whose mutations cause tumor growth, thus subject of positive selection\(^\text{23}\). This may be explained by the fact that this analysis only has the power to detect mutations that occur relatively early in tumorigenesis, as well as to the lack of knowledge to relate IncRNA primary sequence with function, which hampers the adaptation of the analysis methods for IncRNAs. Nevertheless, the observation strongly contrast with the frequent distribution of germline polymorphisms associated with heritability of complex traits outside of protein coding genes\(^\text{24}\). The cited PCAWG study pointed to RMRP (*RNA component of mitochondrial RNA processing endonuclease*), frequently amplified and mutated in sites predicted to affect the secondary structure of the RNA\(^\text{23}\) among the most
frequently mutated genes. In contrast, the extensively studied pro-oncogenic MALAT1\textsuperscript{25}, despite concentrating many indels, did not meet the criteria to be classified as a driver, since its mutations do not correlate with changes in its expression, and are not associated with high cancer cell fractions or biallelic loss\textsuperscript{13}. Interestingly, another study that screened 16,000 lncRNAs for effects on growth of six different human cell lines using CRISPRi-mediated downregulation, showed that out of the 499 lncRNAs affecting growth, 89% showed this effect in exclusively one cell type\textsuperscript{26}. This indicates that, in contrast to protein-coding genes, the functions of lncRNAs are highly cell type-specific and dependent on the lncRNA expression levels, suggesting that the studies based on genomic analysis of few thousands of samples could still lack the statistical power to detect most of the functionally relevant lncRNAs. It is expected that the implementation of single cell RNA-seq techniques in patient-derived samples will further exhibit the exquisitely specific lncRNA expression across the heterogeneity of tumor cells.

References:

1. Guo, C. J. \textit{et al.} Distinct Processing of lncRNAs Contributes to Non-conserved Functions in Stem Cells. \textit{Cell} \textbf{181}, 621-636 e622, doi:10.1016/j.cell.2020.03.006 (2020).
2. Wu, H. \textit{et al.} Unusual Processing Generates SPA LncRNAs that Sequester Multiple RNA Binding Proteins. \textit{Mol Cell} \textbf{64}, 534-548, doi:10.1016/j.molcel.2016.10.007 (2016).
3. Lee, S. \textit{et al.} Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. \textit{Cell} \textbf{164}, 69-80, doi:10.1016/j.cell.2015.12.017 (2016).
4. Yap, K. \textit{et al.} A Short Tandem Repeat-Enriched RNA Assembles a Nuclear Compartment to Control Alternative Splicing and Promote Cell Survival. \textit{Mol Cell} \textbf{72}, 525-540 e513, doi:10.1016/j.molcel.2018.08.041 (2018).
5. Llorian, M. \textit{et al.} Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. \textit{Nat Struct Mol Biol} \textbf{17}, 1114-1123, doi:10.1038/nsmb.1881 (2010).
6. Tichon, A., Perry, R. B., Stojic, L. & Ulitsky, I. SAM68 is required for regulation of Pumilio by the NORAD long noncoding RNA. \textit{Genes Dev} \textbf{32}, 70-78, doi:10.1101/gad.309138.117 (2018).
7. Yin, Q. F. \textit{et al.} Long noncoding RNAs with snoRNA ends. \textit{Mol Cell} \textbf{48}, 219-230, doi:10.1016/j.molcel.2012.07.033 (2012).
8. Poliseno, L. \textit{et al.} A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. \textit{Nature} \textbf{465}, 1033-1038, doi:10.1038/nature09144 (2010).
9. Kleaveland, B., Shi, C. Y., Stefano, J. & Bartel, D. P. A Network of Noncoding Regulatory RNAs Acts in the Mammalian Brain. \textit{Cell} \textbf{174}, 350-362 e317, doi:10.1016/j.cell.2018.05.022 (2018).
Denzler, R., Agarwal, V., Stefano, J., Bartel, D. P. & Stoffel, M. Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance. *Mol Cell* **54**, 766-776, doi:10.1016/j.molcel.2014.03.045 (2014).

Schertzer, M. D. et al. IncRNA-Induced Spread of Polycomb Controlled by Genome Architecture, RNA Abundance, and CpG Island DNA. *Mol Cell* **75**, 523-537 e510, doi:10.1016/j.molcel.2019.05.028 (2019).

Mumbach, M. R. et al. HiChIRP reveals RNA-associated chromosome conformation. *Nat Methods* **16**, 489-492, doi:10.1038/s41592-019-0407-x (2019).

Liu, S. J. et al. Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biol* **17**, 67, doi:10.1186/s13059-016-0932-1 (2016).

Cerase, A. et al. Phase separation drives X-chromosome inactivation: a hypothesis. *Nat Struct Mol Biol* **26**, 331-334, doi:10.1038/s41594-019-0223-0 (2019).

Clemson, C. M. et al. An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Mol Cell* **33**, 717-726, doi:10.1016/j.molcel.2009.01.026 (2009).

Yamazaki, T. et al. Functional Domains of NEAT1 Architectural IncRNA Induce Paraspeckle Assembly through Phase Separation. *Mol Cell* **70**, 1038-1053 e1037, doi:10.1016/j.molcel.2018.05.019 (2018).

Fazal, F. M. et al. Atlas of Subcellular RNA Localization Revealed by APEX-Seq. *Cell* **178**, 473-490 e426, doi:10.1016/j.cell.2019.05.027 (2019).

Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigo, R. & Johnson, R. Towards a complete map of the human long non-coding RNA transcriptome. *Nat Rev Genet* **19**, 535-548, doi:10.1038/s41576-018-0017-y (2018).

Iyer, M. K. et al. The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet*, doi:10.1038/ng.3192 (2015).

Li, Y. et al. Pan-cancer characterization of immune-related lncRNAs identifies potential oncogenic biomarkers. *Nat Commun* **11**, 1000, doi:10.1038/s41467-020-14802-2 (2020).

Leucci, E. et al. Melanoma addiction to the long non-coding RNA SAMMSON. *Nature* **531**, 518-522, doi:10.1038/nature17161 (2016).

Hu, X. et al. A functional genomic approach identifies FAL1 as an oncogenic long noncoding RNA that associates with BMI1 and represses p21 expression in cancer. *Cancer cell* **26**, 344-357, doi:10.1016/j.ccr.2014.07.009 (2014).

Rheinbay, E. et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102-111, doi:10.1038/s41586-020-1965-x (2020).

Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-1235, doi:10.1038/ng.3404 (2015).

Gutschner, T., Hammerle, M. & Diederichs, S. MALAT1 -- a paradigm for long noncoding RNA function in cancer. *J Mol Med (Berl)* **91**, 791-801, doi:10.1007/s00109-013-1028-y (2013).

Liu, S. J. et al. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355**, doi:10.1126/science.aah7111 (2017).