Are there any *HOTTIP*s for defining coding potential of lncRNAs, or just a lot of *HOTAIR*?

Journal Item

Are there any *HOTTIP*s for defining coding potential of lncRNAs, or just a lot of *HOTAIR*?

Rebecca L Mather[*,1], Mark Hirst[1] & Francesco Crea[1]

[1]School of Life, Health & Chemical Sciences, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK

[*] Author for correspondence: Rebecca.mather@open.ac.uk

In the era of next generation sequencing, we have discovered that the human genome encodes up to 60,000 long non-coding RNAs (lncRNAs) [1]. Despite the abundance of these transcripts, few have been characterised. Some studies have demonstrated that lncRNAs are diverse, three-dimensional species that have functions including interactions with DNA, microRNAs, and epigenetic factors, allowing them to regulate many cellular functions [2]. Another function of lncRNAs is transcriptional regulation of mRNAs via interaction with regulatory sequence elements (e.g. eQTLs) [3]. As well as their functional diversity, expression of lncRNAs is tightly controlled in a cell-type and disease-specific manner, making these transcripts an exciting frontier for potential pharmacological intervention [4]. LncRNAs are of particular interest in oncology research as they represent some of the most differentially expressed transcripts between primary and metastatic tumours [2]. It is therefore not surprising that there are several examples of lncRNAs that have clinical relevance, including HOTTIP and HOTAIR, which show promise as biomarkers for hepatocellular carcinoma [5] and breast cancer respectively [6].

By definition, lncRNAs should not be translated, but predicting their protein coding potential remains a challenge. Evidence has emerged that some lncRNAs contain short open reading frames (sORFs) which may encode functional micropeptides (e.g. myoregulin) [7]. One of the reasons for this has been highlighted by studies demonstrating that bioinformatic tools are unable to fully distinguish between coding and non-coding transcripts [8]. This problem is mostly due to a commonly used lower limit of 100 nucleotides, within a transcript's sequence, for an open reading frame to be identified. This means that sORFs, and indeed the encoded micropeptide, may be overlooked [9]. Thus far, much of the identification of the sORFs has relied on either evolutionary conservation, known patterns of codon occurrence, or mass spectrometry. However, these methods are limited by the transcript's size, abundance, and amino acid composition [10]. There are nevertheless several ways that we can address this problem experimentally; one method of which is demonstrated by a study on *Medicago truncatula*, a small legume native of the Mediterranean region. This study demonstrated that the fusion of reporter genes to the sORF promoter regions of pre-miR-171b resulted in transcription of a detectable peptide product [11].

Despite mounting evidence that some lncRNAs are translated nomenclature does not reflect this as the transcripts remain classified as "non-coding". It is not yet clear where we should draw the line as to which transcripts should be re-defined as coding, and which should remain as non-coding. This is further complicated by emerging evidence indicating that some lncRNAs have dual coding and non-coding roles [12]; and that some of these genes can code for protein, but only in specific cell types as recently demonstrated in *Drosophilia sechellia;* [13].

LncRNAs may also overlap with enhancers, giving rise to non-coding enhancer RNAs (eRNAs), which have been shown to be functional despite being disregarded by many as 'genomic junk' [14]. In a study by Bose *et al.* eRNAs were shown to have major, locus-specific roles in enhancer activity as evidenced by CBP interacting with eRNAs as they are transcribed [15]. However it is not yet clear how many eRNAs are also lncRNAs.

Current HUGO guidelines state that lncRNAs should be named using the suffix *-functional* if they are shown to have function after their original naming, but should not be renamed. HUGO guidelines also state that lncRNAs should be named based on genomic context e.g. adding the suffix *–AS* for antisense lncRNAs [16]. As we move forward in assigning functions to each of the lncRNAs, it may be pertinent to consider generating classes and definitions that are also based on coding potential, in order to produce a clear definition for the lncRNAs. To do this we must consider the function, cellular localisation, and genomic context of the transcript [17]. These parameters will help to determine if efficient translation is likely to occur and may also implicate a likely role.

We believe that by using a standard methodology to determine classification of lncRNAs, this highly complex family would be less hindered by inaccurate annotation. If employed, these methods should be based on a combination of bioinformatics and wet lab techniques. For instance, ribosome profiling could be utilised in combination with computational prediction, in order to validate if ribosome association and movement along the transcript occurs (for a review of techniques see [10]).

We thereby suggest three new distinct classes of lncRNAs, which are compatible with current HUGO guidelines, namely; “putative long non-coding RNAs” (plncRNAs), for those that are predicted to be non-coding solely based on bionformatic analyses; “dual coding/non-coding RNAs” (dcncRNAs), for those that have been experimentally confirmed as having both coding and non-coding function, and finally; “confirmed long non-coding RNAs” (clncRNAs), for those that have been functionally identified as pure lncRNAs (e.g. HOTAIR).

In conclusion, the line separating non-coding and coding lncRNAs is still somewhat blurred and a standardised method for classification may be useful in determining the potential functions of these increasingly interesting and complex entities.

1. Iyer MK, Niknafs YS, Malik R *et al*. The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* 47(3), 199-208 (2015). **
Demonstrates the abundance of lncRNAs in the human transcriptome and annotates almost 50,000 previously unannotated lncRNAs, including many associated with cancer.
2. Crea F, Watahiki A, Quagliata L *et al*. Identification of a long non-coding RNA as a novel biomarker and potential therapeutic target for metastatic prostate cancer. *Oncotarget* 5(3), 764-774 (2014).**
Demonstrates funtional significance and expression specificicity of the lncRNA PCAT18. Also describes PCAT18 as a potential therapeutic target and biomarker for prostate cancer.
3. Hon CC, Ramilowski JA, Harshbarger J *et al*. An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* 543(7644), 199-204 (2017).*
This study outlines a method for obtaining accurate 5' ends of lncRNAs using RNASeq data allowing more accurate annotations.
4. Liu SJ, Nowakowski TJ, Pollen AA *et al*. Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biol* 17 67 (2016).
5. Quagliata L, Matter MS, Piscuoglio S *et al*. Long noncoding RNA HOTTIP/HOXA13 expression is associated with disease progression and predicts outcome in hepatocellular carcinoma patients. *Hepatology* 59(3), 911-923 (2014).
6. Gupta RA, Shah N, Wang KC *et al*. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 464(7291), 1071-1076 (2010).
7. Anderson DM, Anderson KM, Chang CL *et al*. A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell* 160(4), 595-606 (2015).**
Study outlines the ability of a lncRNA to encode for a functional and detectable micropeptide.
8. Cech TR, Steitz JA. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* 157(1), 77-94 (2014).
9. Cabrera-Quio LE, Herberg S, Pauli A. Decoding sORF translation - from small proteins to gene regulation. *RNA Biol* 13(11), 1051-1059 (2016).
10. Bazzini AA, Johnstone TG, Christiano R *et al*. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J* 33(9), 981-993 (2014).**
Study descibes methodology to confirm coding potential of transcripts experimentally.
11. Lauressergues D, Couzigou JM, Clemente HS *et al*. Primary transcripts of microRNAs encode regulatory peptides. *Nature* 520(7545), 90-93 (2015).
12. Nam JW, Choi SW, You BH. Incredible RNA: Dual Functions of Coding and Noncoding. *Mol Cells* 39(5), 367-374 (2016).
13. Prieto-Godino LL, Rytz R, Bargeton B *et al*. Olfactory receptor pseudo-pseudogenes. *Nature* 539(7627), 93-97 (2016).**
This study demonstrates the specificity of lncRNA expression by cell type and how stop codons may be 'skipped through'
14. Adelman K, Egan E. Non-coding RNA: More uses for genomic junk. *Nature* 543(7644), 183-185 (2017).
15. Bose DA, Donahue G, Reinberg D, Shiekhattar R, Bonasio R, Berger SL. RNA Binding to CBP Stimulates Histone Acetylation and Transcription. *Cell* 168(1-2), 135-149.e122 (2017).
16. Wright MW. A short guide to long non-coding RNA gene nomenclature. *Hum Genomics* 8 7 (2014).
17. Niazi F, Valadkhan S. Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3' UTRs. *RNA* 18(4), 825-843 (2012).

Financial & competing interests disclosure
Research associated with this article is currently funded by Cancer Research UK (grant 22592). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed. No writing assistance was utilized in the production of this manuscript.