

MICROPEPTIDES ENCODED BY NON-CODING RNA

by

Joshua Forstedt

Copyright © Joshua Forstedt 2020

A Thesis Submitted to the Faculty of the

GRADUATE INTERDISCIPLINARY PROGRAM IN GENETICS

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

In the Graduate College

THE UNIVERSITY OF ARIZONA

2020

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Master’s Committee, we certify that we have read the thesis prepared by: **Joshua James Forstedt**
titled: **Micropeptides Encoded by Non-Coding RNA**

and recommend that it be accepted as fulfilling the thesis requirement for the Master’s Degree.

Fiona McCarthy

Fiona McCarthy

Date: Dec 14, 2020

Rebecca A. Mosher

Rebecca A Mosher

Date: Dec 16, 2020

[Signature]

Darren Hagen

Date: Dec 16, 2020

Final approval and acceptance of this thesis is contingent upon the candidate’s submission of the final copies of the thesis to the Graduate College.

I hereby certify that I have read this thesis prepared under my direction and recommend that it be accepted as fulfilling the Master’s requirement.

Fiona McCarthy

Fiona McCarthy
Committee Chair
School of Animal Comparative Biomedical Sciences

Date: Dec 14, 2020



Table of Contents

1. Abstract	4
2. Introduction	4
3. Non-Coding RNA & Micropeptides	7
3.1. Non-coding RNA with sORFs	7
3.1.1. Bifunctional RNA	8
4. NcRNA Translation	9
4.1. Micropeptides Encoded by Long Non-Coding RNAs	11
4.2. Micropeptides Encoded by Circular RNAs	12
4.3. Micropeptides Encoded by MicroRNA	14
5. Challenges of Non-Coding RNA classification	16
5.1. OMICs resolution	17
6. Advancements in Identification of Micropeptides	23
6.1. Small Open Reading Frames	23
7. Industrial Applications of Non-Coding RNA data	25
7.1. Diagnostic & Pharmaceutical Applications	26
7.2. Agricultural Applications	28
7.3. Bifunctional mechanism	29
7.4. Identification of targets	30
7.5. Functional prediction	31
8. Perspective	32
9. References	33

Abstract

Advancements in high throughput sequencing techniques have revealed different classes of non-coding RNA contain short open reading frames (sORF) with coding potential. Similarly a combination of bioinformatic tools and experimental data have identified that sORFs are in fact translated to small proteins called micropeptides. Thus non-coding RNA previously annotated as non-protein forming, seem to have newly identified coding ability and therefore need to be reclassified as bifunctional RNA. This review discusses the changing paradigms of ncRNA classification to include its coding functions, methods to identify sORFs within these classes, and the functional relevance of micropeptides. It also aims to highlight the potential application of micropeptides in therapeutic and agronomic interventions. Finally, it provides future research perspectives in the promising field of micropeptide biology.

Introduction

The well accepted “central dogma” of molecular biology oversimplifies the flow of genetic information and is almost unidirectional from DNA to RNA to protein (Li & Liu, 2019) . Since the “RNA-world” hypothesis, many regulatory non-coding RNA (ncRNA) have been identified (Nam et al., 2016) , and challenged this view. Moreover, the application of high-throughput RNA sequencing techniques has broadened the effector functions of RNA in genome regulation (Ruiz-Orera et al., 2020) . We know now that genomes of higher organisms are extensively transcribed than previously anticipated (Makarewich & Olson, 2017). Much of this pervasive transcription accounts for ncRNA, the proportion of which varies among species and is much higher than the protein coding component in higher eukaryotes (Zheng et al., 2019). Projects such as ENCODEv7 and FANTOM5 have systematically catalogued different classes of ncRNA providing a comprehensive repository for future analyses (Derrien et al., 2012) (Hon et al.,

2019). Well annotated examples within the ncRNA domain have been accredited with housekeeping or regulatory functions (Fig. 1) without being translated and adhere in their definition as non-coding. A novel class within this paradigm is long non-coding RNA (lncRNA) with newly identified coding potential (Nam et al., 2016) (Li & Liu, 2019). Over the last few years data has accumulated showing the translation of short open reading frames (sORFs) within lncRNA and can account for pervasive transcription (Ruiz-Orera et al., 2020). Such a concept redefines classical conformations of coding versus non-coding and delineates a rather overlapping function for RNA. In doing that it provides a new perspective about genome organization including what it means to define a gene (Li & Liu, 2019). The following sections will define some of these concepts that will be covered in detail within this review.

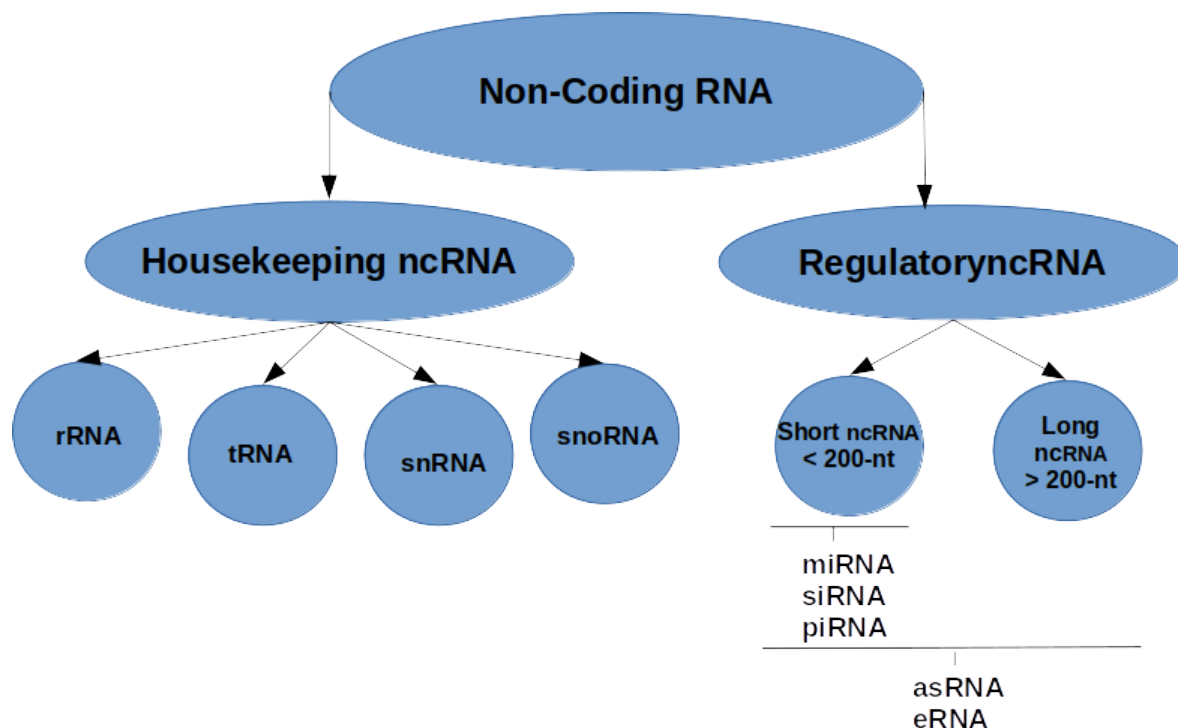


Figure 1: Types of ncRNA. NcRNA can be grouped based on housekeeping functions or regulatory functions. Housekeeping ncRNA include ribosomal (rRNA), transfer (tRNA), small

nuclear (snRNA), and small nucleolar RNAs (snoRNAs). Regulatory ncRNA consist of short ncRNA (<200nt) and long (>200nt) and include microRNAs(miRNAs), small interfering RNAs (siRNAs), Piwi-associated RNAs (piRNAs), antisense RNAs (asRNAs), and enhancer RNAs (eRNAs). (Adapted from Losko et al., 2016)

Long non-coding RNA (lncRNA) are greater than 200-nt in length and make up one of the largest classes of ncRNA. They are known to abrogate miRNA function by competing with miRNA target binding sites as well as act as scaffolds for proteins (Yao et al., 2019). Large scale transcriptome analyses have revealed that lncRNA accounts for a significant proportion of all transcripts and may therefore contain large stores of potential coding sequences. (Andrews & Rothnagel, 2014). Similarly large scale genome analysis studies done in *E. coli*, *Drosophila melanogaster*, *Danio rerio*, *Arabidopsis thaliana* and humans and mice also provide sufficient evidence supporting the presence of translated sORFS within lncRNA (Zheng et al., 2019), (Bazzini et al., 2014)(Kong et al., 2020), (Ingolia et al., 2011). lncRNA transcripts are similar to their protein coding mRNA counterparts as they are both transcribed by the same machinery and are known to contain post-transcriptional signatures resulting from capping, splicing, and even some epigenetic markers (Makarewich & Olson, 2017). But lncRNA transcripts differ from coding transcripts in that they lack codon conservation and are free of selection pressures. Studies done to identify lncRNA containing sORFs have primarily focussed on cross-species conservation and have found that functional sORFs and not their host lncRNA are highly conserved indicating a functional role different from that of lncRNA in the cell. (Makarewich & Olson, 2017).

Scope of the Review

This review aims to redefine RNA as a regulatory element and a source for micropeptides. It will focus on mechanistic details on sORF translation and describe the various classes of sORF encoded micropeptides. Further it presents ways for micropeptide identification and delineate

micropeptide functions via examples reported in the literature. Functional micropeptides could have large implications for future pharmacological targets and leveraging these elements in plants could dramatically impact the need for pesticides and fertilizers, and potentially make crops more resistant to higher average temperatures. Finally, this review attempts to highlight some of the challenges that may impede the progress of micropeptide biology and bridge the gaps in our understanding of this emerging field.

Non-Coding RNA & Micropeptides

Non-coding RNA with sORFs

Andrews and Rothnagel define sORFs as a string of 2 to nearly 100 codons that may be translated (Andrews & Rothnagel, 2014). While some sORFs that are translated are present within overlapping regions of known coding sequences in their 5' or 3' UTRs, others are located within previously annotated as non-coding RNA such as, intergenic regions, lncRNA or antisense RNA (Andrews & Rothnagel, 2014). Finally, circRNA which are covalently closed looped RNA (Wu et al., 2020), as well as pri-miRNA which are precursor molecules to miRNA also contain sORFs (Couzigou et al., 2015) (Fig. 2). It is possible that sORFs in any given stretch of nucleotide sequence can occur by chance and have no real function (Mackowiak et al., 2015). However recent reports suggest a close association with ribosomes and may indicate that they are translated (Slavoff et al., 2013). The sORFs translated into short peptides (<100aa) as identified by ribosome profiling (Ingolia et al., 2011) are implicated in regulating various cellular processes such as proliferation, metabolic homeostasis, myogenic differentiation, and translation inhibition (Wang et al., 2019). Similarly, studies done in flies to humans have uncovered a handful translated sORFs in genes previously annotated as lncRNA (Couso & Patraquim, 2017). The real challenge faced by researchers is to discern whether the

translated sORFs produce stable peptides, in significant amounts and whether they are functionally relevant (Nam et al., 2016). But before that, the presence of these sORFs hidden within lncRNA highlights the converging and overlapping functions for RNA (Li & Liu, 2019), and as such needs further investigation.

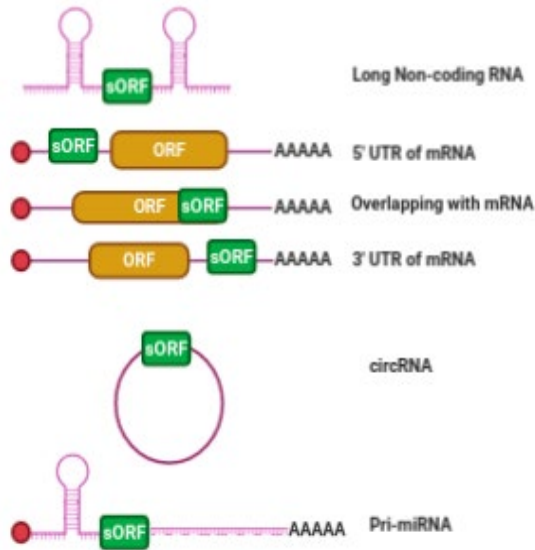


Figure 2: Distribution of sORF in various transcripts. (Adapted from Yin et al., 2018).

Bifunctional RNA

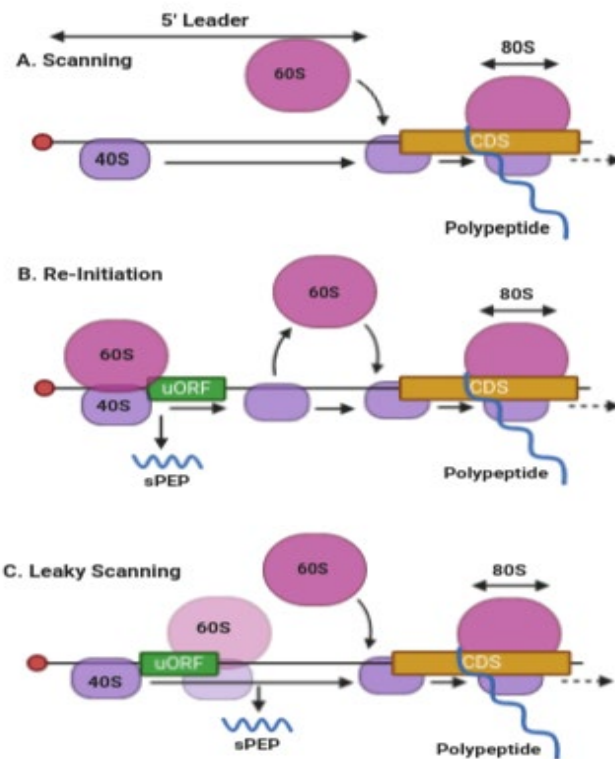
Bifunctional RNA are a class of RNA molecules that have both coding and non-coding functions. Bifunctionality is observed in numerous cases. One example is lncRNA with sORFs encoding short peptides in which the regions for coding and non-coding may overlap as they come from the same transcript. (Zheng et al., 2019) Another example is mRNA with non-coding isoforms. The presence of cis-acting elements in the UTRs of mRNA allow for self-regulation of the mRNA activity (Nam et al., 2016). Therefore, it is possible to demarcate coding and non-coding regions within the same transcript. An example is Oskar mRNA in *Drosophila melanogaster* which translates into a protein required for germline specification but can also retain a regulatory

function exerted via its 3'UTR in early oogenesis (Ruiz-Orera & Albà, 2019). Alternative splicing or the use of alternative promoters results in isoforms without or without coding potential. It is not the same molecule that has dual characteristics but a common loci which produces two different isoforms, indicating that the originating locus is bifunctional and not the RNA molecule (Hubé & Francastel, 2018). The bifunctional nature of RNA indicates that the cell uses a “switch” mechanism between coding and non-coding states (Zheng et al., 2019). But there is no clear consensus on how the cell regulates and maintains homeostasis between these two states, the dysregulation of which could be indicative of cellular pathophysiology. Additionally, the biological roles of functional peptides may be better understood when compared with the bifunctional RNA from which they are derived.

NcRNA Translation

Micropeptides are synthesized as small proteins that contain less than 100aa, in a way that is like larger proteins. However, they differ from peptide hormones and neuropeptides which are formed by enzymatic cleavage of larger precursors. (Makarewich & Olson, 2017). A study published in 2004 found that an AUG start site was absent in a glycyl-tRNA synthetase mRNA coding for the enzyme in yeast. The mRNA coded for two products, and whilst the shorter version used an AUG start site, the longer version utilized a UUG codon to begin translation, suggesting that cells use non-canonical in-frame codons for protein synthesis (Chang & Wang, 2004). Ingolia et al. showed that micropeptide translation also used upstream cognate non-AUG codons like CUG or GUG (Ingolia et al., 2011). Similarly, another study in 2013, used peptidomics to demonstrate that more than half the detected micropeptides were synthesized using cognate non-AUG codons. (Slavoff et al., 2013) Further, micropeptide translation can be explained through the presence of upstream ORFs (uORFs) (Andrews and Rothnagel, 2015). uORFs, a regulatory class of sORFs (Ruiz-Orera & Albà, 2019) are sites of translation

attenuation of associated downstream protein coding DNA sequences (CDS) that restrict ribosome access (Andrews and Rothnagel, 2014). They contain 5' leader sequences that can occasionally be transcribed into regulatory molecules which directly affect downstream coding sequences at the translational level. (Andrews & Rothnagel, 2014) Micropeptides may be translated by utilizing uORFs via one of two modes: By Ribosomal reinitiation or leaky scanning



(Fig 3.). For example, GCN4, a transcription factor in yeast, is inhibited by the translation of its uORF. But under nutrient starvation, translation is resumed from the canonical AUG codon, to restore basal level concentrations of the protein (Ji et al., 2015). sORFs enriched in lncRNA may be translated by similar mechanisms (Bazzini et al., 2014).

Figure 3: Re-initiation and scanning: A- A standard model of eukaryotic translation showing a 40S ribosomal subunit binding to the cap at the 5' end. The 40S scans the mRNA until the start codon, joins the 60S subunit to form the 80S elongation competent ribosome and translates the ORF. B- Re-initiation occurs when the 40S initiates translation at a downstream CDS after

completing the translation of an upstream ORF (uORF). uORF translation forms a short peptide encoded by a short ORF (sPEP). C- In leaky scanning, the 40S ribosomal subunit can either recognize the start codon of an uORF and further translate an sPEP, or it can scan past the upstream start codon and initiate translation at a downstream start codon. (Adapted from Andrews & Rothnagel, 2014)

The translation efficiency of lncRNA compared to that of mRNA is almost similar implying that ncRNA that are translated have immediate access to the translation machinery (Ji et al., 2015). This could also account for the pervasive translation of the transcriptome (Ruiz-Orera & Albà, 2019). However ribosomal translation does not imply that an sORF encodes a functional peptide (Bazzini et al., 2014). It is possible that the peptide may be unstable or could function to alter transcript stability (Bazzini et al., 2014). Biochemical assays such as in-vitro translation can confirm micropeptide stability (Makarewich & Olson, 2017). Similarly, functional micropeptides are identified by epitope tagging using CRISPR-Cas gene editing techniques which allows for downstream applications such as immunoprecipitation, immunocytochemistry, and western blot (Makarewich & Olson, 2017)

Nevertheless, a few well characterized examples of functional micropeptides are reported by several research groups (Makarewich & Olson, 2017). The following sections will present some examples of functional micropeptides based on the class of ncRNA of their origin along with the mechanisms by which they are synthesized.

Micropeptides Encoded by Long Non-Coding RNAs

A small subset of lncRNA described to date, function mainly as modulators of gene expression and are involved in various cellular processes such as chromatin remodelling, splicing, mRNA stability and translation inhibition (Hartford & Lal, 2020). One of the earliest examples of a coding lncRNA, Steroid Receptor RNA (SRA) was initially identified as a regulator of steroid

receptor-dependent gene expression (Li & Liu, 2019). The protein product of the SRA lncRNA, SRAP, was found to regulate the transcription of *SRA1* gene. The switch between coding and non-coding functions of SRA, was a result of an alternative splicing event, suggesting that alternative splicing is an important determinant of ncRNA bifunctionality (Li & Liu, 2019). Further, studies have shown lncRNA with regulatory roles are expressed in a tissue specific manner (Yao et al., 2019). Matsumoto et al hypothesized that micropeptides expressed within tissues may fine-tune biological processes specific to the cell type (Matsumoto et al., 2017). An example is mitoregulin (MtlN) encoded by *LINC00116* and is expressed in human adipocytes. MtlN regulates lipolysis and mitochondrial β -oxidation in adipocytes and is also found to be conserved in mice (Stein et al., 2018). Similarly muscle specific lncRNA *LINC00948* in humans and mice (*2310015B20Rik*, the lncRNA in mice.) encodes a 46aa micropeptide, Myoregulin and interacts with sarco/endoplasmic reticulum Ca^{2+} -ATPase to affect muscle relaxation rates (Li & Liu, 2019). Cellular compartmentalization is another factor that predicts lncRNA translation. lncRNA are localized to both the nucleus and the cytoplasm (Li & Liu, 2019). Ji et al found that translated lncRNA, more likely localized to the cytoplasm than to the nucleus and showed translation efficiency comparable to that of mRNA (Ji et al., 2015). Although the number of lncRNA encoded micropeptides remains small, the few that are reported show cross species conservation and have distinct functional roles within the cellular landscape (Wu et al., 2020). implying the need to re-classify transcripts previously annotated as non-coding.

Micropeptides Encoded by Circular RNAs

Circular RNAs (circRNAs) are a type of ncRNA transcribed as covalently closed loops that lack 5'-3' polarity or polyA tails (Wu et al., 2020). They are transcribed by the process of back-splicing whereby a down-stream donor 5' splice site is fused to an upstream acceptor 3' splice site (Fig 4). This process re-orders the exons and is an alternative to canonical splicing (Di Timoteo et al., 2020). circRNA undergo translation by recruiting ribosomes at internal ribosome

entry sites (IRES) (Hartford & Lal, 2020) . An IRES is a regulatory element found in the 5' UTR of a subset of genes and is responsible for cap independent translation initiation (Wu et al., 2020) .

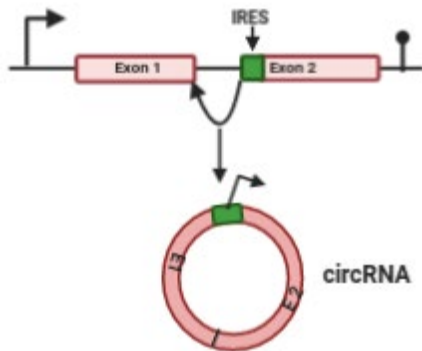


Figure 4: Backsplicing: circRNA are transcribed as products of backsplicing by the reordering of exons. (Adapted from Wang & Wang, 2015).

Timoteo et. al, reported Circ-ZNF609, a circRNA involved in myoblast differentiation, to undergo translation via an IRES. Circ-ZNF609 contains an ORF that is translated into small peptides either by canonical or cap independent splicing pathway (Di Timoteo et al., 2020). Additionally, it was an N⁶-methyladenosine (m⁶A) post-transcriptional modification at the ORF that resulted in the back-splicing method being selected and subsequently encoding a peptide (Di Timoteo et al., 2020). Similarly, ribosomal footprinting data has shown that circRNAs bind to ribosomes in the brain tissue of drosophila and undergo translation to form peptides detected by mass spectrometry (Kong et al., 2020) . Further study found that starvation likely regulates the cap independent translation of Ribo-circRNA indicating that the cell uses alternative splicing and further translates circRNA transcripts in response to physiological and environmental stress conditions (Kong et al., 2020) . Cap-independent proteins are translated during hypoxia, heat shock, or viral infection and help the cell survive (Kong et al., 2020). It is possible that

micropeptides originating from circRNA are synthesized by a similar mechanism and may help the cell survive under stress conditions. Therefore, regulatory elements such as m6A modification and the presence of IRES mediated ribosome binding are considered important drivers of translation and are incorporated in several translation prediction methods such as DeepM6ASeq and IRESite (Wu et al., 2020). Finally, biochemical studies have revealed circRNA encoded peptides are implicated in several human cancers (Wu et al., 2020) and show a direct functional role for micropeptides. Examples include the SNF2 histone linker PHD RING helicase gene encoded, SHPRH-146aa which is downregulated in glioblastoma and the *circ-FBXW7* encoded circ-FBXW7-185aa which induces cell cycle arrest and hinders glioma cell proliferation (Wu et al., 2020). Both these examples demonstrate that circRNA translated peptides affect cancer behaviour and present context specific mechanisms for ncRNA encoded peptide synthesis.

Micropeptides Encoded by MicroRNA

miRNAs are single stranded RNA molecules synthesized as enzymatically cleaved products of precursor RNA and suppress or inhibit translation by base pairing to complementary regions in the 3' UTR of protein coding mRNA transcripts (Santulli, 2015). miRNA are transcribed as precursor primary miRNA or pri-miRNA by RNA polymerase II and then subjected to splicing, capping and polyadenylation just like protein coding genes (Chen et al., 2020). Pri-miRNA possibly code for functional peptides since they are large precursor molecules that undergo maturation to form miRNA (Chugunova et al., 2018). Examples of coding pri-miRNA include, pri-miR165a from alfalfa and pri-miR171b from *A. thaliana*, and are reported to produce peptides that regulate root development (Wu et al., 2020). Similarly primary miRNA *miR-200a* and *miR 200b* encoded micropeptides, designated as miPEP-200a and miPEP-200b respectively, regulate epithelial to mesenchymal transition in prostate cancer cells (Wu et al.,

2020). These examples denote pri-miRNA as important classes of gene regulators, with additional coding functions in both plants and animals.

Couzigou et. al, reported plant miRNA loci to contain coding ORFs (Couzigou et al., 2015). The peptides produced from these loci increased the transcription of associated miRNA leading to enhanced silencing of target genes. Further, *A. thaliana* plants treated with synthetic micropeptides showed specific phenotypes seen with the corresponding miRNA overexpression, suggesting that endogenous application of synthetic micropeptides can assess the coding function of miRNA encoded micropeptides. Additionally, synthetic micropeptides can be used as a powerful tool to directly target genes in a non-model organism or commercially useful plants without the need for genetic manipulation (Couzigou et al., 2015).

Most pr-miRNA encoded peptides are reported in plants, with only a few examples seen in mammals (Chugunova et al., 2018) . One of the best characterized miRNA in mammals, is miR155 and is produced by lncRNA *MIR155HG* (Niu et al., 2020). miR155 and its parent lncRNA *MIR155HG* regulate inflammation and immune responses in inflamed dendritic cells (Niu et al., 2020). As a precursor to miR155, *MIR155HG* is localized to the nucleus to be further processed into mature miRNA. Niu et. al found measurable levels of *MIR155HG* in the cytoplasm and was found to encode a 17 aa micropeptide, P155. P155 selectively bound to the ATP binding site of HSP70, a peptide chaperone and modulated antigen presentation to MHC class II and CD 4+ cells (Niu et al., 2020).

From the examples given above representing various classes of ncRNA encoded micropeptides, it is clear that coding and non-coding functions exist within the same RNA molecule and the cell seems to have conserved mechanisms in place that utilize either of these functions in a context specific manner. But not all ncRNA with ability to be translated form micropeptides. In order to study the mechanistic details for making the switch between coding

and non-coding, it is important to devise methods to identify and classify ncRNA encoding functional peptides.

Challenges of Non-Coding RNA classification

Advances in RNAseq technology have allowed researchers to map transcriptomes with greater coverage and improved quality. (Tripathi et al., 2017) RNASeq data has revealed new classes of lncRNA ((St.Laurent et al., 2015) adding to an ever increasing repository of ncRNA.

Classification is based on transcript length, subcellular localization, genomic location and transcript properties and function (St.Laurent et al., 2015). Novel transcripts with coding potential are annotated based on their resemblance with protein coding genes. There are two criteria for defining coding genes. They must be an independent transcription unit and have some evidence of translation. (Hanada et al., 2013) lncRNA have most of the attributes of mRNA including a cap, polyA tails, UTRs, and their own promoters and terminators. (St.Laurent et al., 2015). Moreover, sufficient examples of sORFs originating from lncRNA, are translated into functional peptides, suggesting that there are no definitive signatures that demarcate a strict boundary between coding and non-coding transcripts. Similarly, most circRNAs are derived from overlapping exons, which suggests that coding circRNA may be re-classified as other types of mRNA (Wang et al., 2019). Further, existing classes represent only a small fraction of the lncRNA present within the cell, of which only 0.05–1.12% are similar to protein coding genes (St.Laurent et al., 2015). Therefore, current classification strategies may be underrepresenting a vast majority of lncRNA, to which sORFs have been mapped (Couso & Patraquim, 2017)

Since lncRNA are expressed in a cell specific or developmental stage specific manner, initial studies were done to obtain transcriptome data sets from different cell-types (Mackowiak et al., 2015). These studies provided sequence and locus information. Hence, initial lncRNA classification was based on sequence derived features such as predicted ORF length,

sequence similarity with known protein coding genes and sequence conservation. (Choi et al., 2019) The following sections describe some of the classification strategies as well as the challenges associated with existing classification paradigms.

OMICs resolution

Read length is one of the most used metrics to classify lncRNA. Next generation sequencing protocols typically filter transcripts greater than 300 nt or 100aa and classify them as non-coding (Makarewich & Olson, 2017) based on the premise that protein coding genes would contain longer ORFs (Choi et al., 2019). Computational gene annotation methods are therefore biased against sORFs, some of which encode peptides, as little as 9aa (Wang et al., 2019) .

Ribosome profiling is a deep sequencing based tool that measures translation dynamics at the genome wide level with single nucleotide resolution, by sequencing ribosome protected fragments (RPF) (Mackowiak et al., 2015) Although ribosome profiling is extensively used as a method to classify coding and non-coding transcripts, it does not distinguish between actively translating and non-specific ribosome interacting transcripts (Choi et al., 2019) RibSeq analysis methods were later modified to contain features that compared RPF coverage in ORFs versus UTRs (Ingolia et al., 2011). Further, a subset of the translated ORFs in lncRNA conserved regions are likely to encode proteins that are under selective constraints (Ruiz-Orera & Albà, 2019) . Ruiz-Orera et. al, used RibORF, a program to score read periodicity and reported that conserved regions in lncRNA were thrice as likely to contain translated ORFs than non-conserved regions (Ruiz-Orera & Albà, 2018) . Similarly, Seo. et al recently reviewed several such additional machine learning algorithms that were used in combination with ribosome profiling analysis to determine translation in identified sORFs (Choi et al., 2019) (Table1)

Bioinformatic tools used for classifying coding potential include properties that define ribosome dynamics such as 3 nucleotide periodicity, Ribosome A site alignment and frequency of codon

usage (Ji et al., 2015) and thereby provide direct evidence of translation. Recent studies have incorporated machine learning tools with intrinsic features such as Ribosome release scores (RRS) and ORFscores that measure biased distributions of RPFs towards the first frame of the CDS (Bazzini et al., 2014). These features rely on triplet periodicity and sub-codon phasing similarly imply active translation. However, what they gain in translation prediction they lack in secondary structure conservation prediction, specific to lncRNA (Tripathi et al., 2017) implying that these computational methods are limited in their robustness to identify lncRNA with coding functions (Tripathi et al., 2017). In conclusion commonly used classifiers must be free of length restrictions and should specifically detect lncRNA with coding potential based on structural and functional homologies. (Mackowiak et al., 2015)







Species	Method	Experimental Data	Number of transcripts or sORFs analysed	Translated sORFs detected in lncRNAs	MS evidence	Reference
Human 	ORFscore	RiboSeq	261 from lncRNAs	261	-	Choi et al., 2019
	RibORF	RiboSeq	925 from 233 lncRNAs		18 lncRNAs	
	PhyloCSF	RiboSeq, MS	354 from lncRNAs	354	22 peptides	
Mouse 	sORF finder	Ribo-seq	514 from lncRNAs	514	-	Choi et al., 2019
	Hexamer-based coding score	Ribo-seq	137 from 403 lncRNAs	107	-	
	PhyloCSF	MS	98 from lncRNAs	98	11 peptides	
<i>Drosophila melanogaster</i> 	PhyloCSF	MS	53 from lncRNAs	53	2 peptides	Choi et al., 2019
	Hexamer-based coding score	Ribo-seq	7 from 22 lncRNAs	7	-	
Yeast 	-	Ribo-seq, PolySeq	47 from 331 lncRNAs	47	-	Choi et al., 2019
	Hexamer-based coding score	Ribo-seq	5 from 6 lncRNAs	5	-	
<i>Arabidopsis thaliana</i> 	Hexamer-based coding score	Ribo-seq	43 from 93 lncRNAs	43	-	Choi et al., 2019
Zebrafish 	ORFscore	Ribo-seq, MS	535 from lncRNAs	535	6 peptides	Choi et al., 2019
	PhyloCSF	MS	99 from lncRNAs	99	-	
	Hexamer-based coding score	Ribo-seq	379 from 726 lncRNAs	155	-	

Table 1: Studies done in model organisms have identified sORFs encoded peptides in lncRNA. (Adapted from Choi et al., 2019)

Another observed attribute of lncRNA is their close associations with protein coding genes. The FANTOM consortium described transcription forests as overlapping regions between non-coding and exons of coding regions at a given loci. (St.Laurent et al., 2015) It is from these regions that several subclasses of lncRNA can be obtained based on whether they arise from sense or antisense strand, whether they contain spliced or unspliced sections of their associated genes or whether they exclusively arise from intronic regions of the gene. (St.Laurent et al., 2015) Each category has its own nomenclature and may be expressed as a percentage of how frequently they occur in the genome (St.Laurent et al., 2015). Similarly data is derived using a wide range of experimental methods (RNAseq, Random amplification of cDNA ends, tiling microarrays) or by performing in-silico analysis of existing databases (EST, RefSeq) (St.Laurent et al., 2015). Here the challenge lies in delineating various classes for two reasons: Firstly, there is very little overlap in the identified lncRNA amongst different groups of researchers due to differing underlying hypothesis and techniques applied (Mackowiak et al., 2015). Secondly, many of these classes are not mutually exclusive. For example, CDKN2B-AS1 also known as ANRIL is both a Natural Antisense RNA (NAT), and a circRNA. (St.Laurent et al., 2015) Therefore a classification strategy that accounts for a continuum of functions from non-coding to coding rather than genomic location may be better suited for experimental design and predicting ncRNA classes with coding potential.

Functional Prediction

As is the case with protein coding genes, functional annotation of sORF encoded micropeptides are based on computational and experimental evidence (Makarewich & Olson, 2017).

Computational tools rely on sequence similarity, as cross species conservation between

putative coding sequences denotes selective value and hence function (Couso & Patraquim, 2017) . For example, PhyloCSF is a tool that compares amino acid conservation between canonical protein sequences and micropeptides and is used to predict conservation in start/stop codons in related species (Mackowiak et al., 2015). Additionally, comparative genomics data is mined for the depletion of non-synonymous mutations over synonymous mutations, and indicates conservation within lncRNA sequences (Mackowiak et al., 2015). However, since sORFs code for fewer amino acids, they often get lower quantitative conservation scores compared to longer proteins and may be incorrectly classified as non-functional (Couso & Patraquim, 2017) . Further, several computational approaches use support vector machines to train models that can predict coding potential of non-coding transcripts (Table 2) Training data sets consist of BLAST related features and look for sequence homology to classify coding and non-coding transcripts (Choi et al., 2019). But the training sets are based on protein coding properties of longer canonical proteins and may not accurately represent micropeptides (Peeters & Menschaert, 2020). Moreover, the training process and the sequence alignment steps are time consuming and do not take into consideration the biological characteristics of lncRNA (Choi et al., 2019) . As a result, it is difficult to identify functional micropeptides based

on computational methods alone.

Computational coding RNA Classifiers	Experimental Data	Method	Objective	Reference
PhyloCSF	RiboSeq, MS	Expectation Maximization	Evaluate evolutionary signatures in the form synonymous or nonsynonymous codon substitution rates	Choi et al., 2019
CPC2	RiboSeq	Support Vector Machine	Detect and predict sORF coding potential based on ORF length and nucleotide composition	Peeters & Menschaert, 2020
RibORF	RiboSeq	Maximum Entropy Value/ SVM	Identify translated ORFs based on 3-nucleotide periodicity, A site alignment and RPF coverage	Choi et al., 2019
sORF finder	RiboSeq	Hexamer Frequency	Detects sORFs with high coding potential based on nucleotide composition and substitution ratio (dN/dS)	Choi et al., 2019
ORFscore	RiboSeq	Chi-square test	Quantify biased distribution of RPF in the first reading frame	Choi et al., 2019
RRS	RiboSeq	Read Counts of Ribosome Occupancy	Measure discrepancies in ribosome occupancy in oRF versus non-ORF regions	Choi et al., 2019
TOC	RiboSeq	Random Forrest	Detect sORFs based on RPF coverage and translation efficiency	Choi et al., 2019
FLOSS	RiboSeq	RPF length distribution	Predicts coding versus non-coding nature of transcript based on RPF coverage and RPF length distribution	Choi et al., 2019
PROTEOFORMER	RiboSeq, MS	Proteogenomics Workflow	Custom database based on RiboSeq to predict novel isoforms	Peeters & Menschaert, 2020

Table 2: Overview of bioinformatics tools used in ribosome profiling to classify and predict coding potential of sORFs. (Adapted from Choi et al., 2019 & Peeters & Menschaert, 2020).

Bioinformatic analyses must be supported with experimental evidence to demonstrate true coding potential. A study done in *Drosophila melanogaster* identified translated sORFs that were enriched in peptides allocated to cell membranes and organelles (Couso & Patraquim, 2017). Therefore, subcellular localization of ncRNA transcripts is important for classification as translatability is directly correlated with location within the cell (Yao et al., 2019). Classification methods should link sequence information with molecular functions and biochemical properties (Couso & Patraquim, 2017) and be experimentally validated through gene-knockdown, overexpression and CRISPR editing techniques (Makarewich & Olson, 2017). But it is difficult

to individually assign a function to every lncRNA identified, using the said techniques as these are time consuming and cannot practically keep up with the large number of existing and newly discovered lncRNA.

The challenge of case-by-case functional prediction of coding lncRNA may be overcome by using the guilt-by-association principle (Lefever et al., 2017). lncRNA are closely associated with DNA, mRNA and proteins and the guilt-by-association principle utilizes the power of clustering algorithms to show co-expression of lncRNA with protein coding genes (Lefever et al., 2017). The principle assumes that if a pair of lncRNA and mRNA show high correlation in their expression, then their functions may be related and thus can be deduced. RNAseq or microarray data sets under varying experimental conditions are used as sources for paired expression profiles, (Tripathi et al., 2017) based on which common clusters containing pairs of lncRNA and associated mRNA are identified. The function of an unknown lncRNA can be deduced since a common cluster is indicative of a common biological process (Lefever et al., 2017). Similarly, web-based computational tools such as Co-LncRNA, are used to perform enrichment analyses of expression-related genes with individual or multiple lncRNAs in all known GO annotations and KEGG pathways (Zhao et al., 2015) which may be useful for further assigning function to newly identified sORFs present within lncRNA.

Therefore, the integration of computational and experimental data is crucial to characterize newly discovered sORF with coding potential. But to increase the confidence that sORF are indeed translated, methods that identify the products of translation add additional value (Peeters & Menschaert, 2020). The application of ribosome profiling has provided novel insights into protein-coding sORFs including their identification in polycistronic genes, upstream ORFs, overlapping ORFs, and the non-canonical use of ATG codons (Yeasmin et al., 2018). However, association of ribosomes does not mean that the ORFs will be translated and additional evidence of translation is needed to prove that sORFs indeed get translated to functional

peptides. The following section will highlight some of the techniques that can detect the peptides products of sORF translation, and hence provide a better resolution of the sORF translation landscape.

Advancements in Identification of Micropeptides

Small Open Reading Frames

Recent classification strategies have revealed that sORFs exist randomly throughout the genome as inert DNA sequences, while others may be transcribed and translated to act as cis-regulators of associated protein coding genes. Finally, only a small subset of sORFs may be actively translated to code for functional peptides with the propensity to modulate canonical proteins (Couso & Patraquim, 2017) Therefore it is clear that not all sORFs are translated to produce functional peptides. Although Ribo-seq data predicts novel translation events at sub-codon resolution, reproducing RibSeq data is challenging since the tools and metrics applied cannot handle biological replicates (Peeters & Menschaert, 2020). RibSeq experiments that are validated over multiple replicates can be used to increase the likelihood of active translation. Similarly, targeted RiboSeq experiments that utilize the expression-specific property of micropeptides such as during stress or antigen presentation, can add value to sORF discovery (Peeters & Menschaert, 2020). However experimental data to detect translated sORF is not restricted to RibSeq. The next section will highlight some of these advancements in functional proteomics that are used to understand the biological nature of the translated products of sORFs.

Pathways for Identification

The Global Translation Initiation or GTI technique includes a pretreatment with translation initiation inhibitors like harringtonin or lactimidomycin and provides mapping of translation start sites (Chugunova et al., 2018) . When combined with samples treated with a translation elongation inhibitor like cycloheximide now generates two Rib-seq data sets and gives a better accuracy of predicting translating regions (Choi et al., 2019) . Similarly, a modified version of GTI is Quantitative Translation Initiation Sequencing captures real time translation initiation events with greater qualitative and quantitative precision (Chugunova et al., 2018) .

Experimental data generated from these techniques are often combined with computational tools such as PROTEOFORMER which additionally maps RPFs, identifies translated transcripts and translation initiation sites and creates a protein sequence database that can be used for MS-based proteomics analysis (Choi et al., 2019). Another modification of ribosome profiling, Poly-Ribo-Seq accounts for multiple ribosomes or polysome binding to the coding transcripts and enriches small polysomes more likely to form during sORF translation (Hartford & Lal, 2020)

Mass Spectrometry is an analytical tool that ionizes peptides and identifies their aa sequences by measuring the mass-to-charge ratio of the ionized peptide (Choi et al., 2019) . MS is often used in tandem with transcriptomics to detect lncRNA encoded micropeptides (Hartford & Lal, 2020). For example, Slavoff et. al, used proteogenomics to generate a custom reference database of potential micropeptides greater than 8 amino acids from the human genome (RefSeq) and a database with peptides from MS/MS spectra (Sequest database) (Slavoff et al., 2013) . The custom database (DB) allowed for identification of 86 novel micropeptides in K562 cells. Despite being a powerful technique MS identification of microproteins is impeded by their short length, low abundance, and the possible removal of small translation products during MS sample preparation. Similarly, customized DBs designed using 6 frame translation (6FT) contain hypothetical protein sequences and are difficult to use due their large size (Chugunova et al.,

2018) . Recent proteogenomic workflows reduce the search space 6FT databases by applying predicted isoelectric point filters with the help of algorithms such as Predpl, while simultaneously increasing the identification rate (Peeters & Menschaert, 2020) Finally MS sensitivity to detect small, low abundant proteins can be enhanced by the application of newly developed techniques such as data independent acquired (DIA) MS and trapped ion mobility spectrometry (TIMS) using a time of flight mass analyzer (TOF). Data from DIA and TIMS-TOF are analyzed by machine learning algorithms that learn and train a method for MS2 fragment intensity predictions and can produce a full coverage of the cleaved peptides (Peeters & Menschaert, 2020) .

Finally, methods such as *in-vitro* translation use fusion-proteins, expressed in the cell to probe for micropeptide interacting proteins. An example is the APEX method, in which Ascorbate peroxidase fusion protein reacts with peroxide in the presence of phenol-biotin and adds a biotin tag to neighboring proteins in the process. Biotinylated proteins are further enriched by MS and hence provides information about the protein environment of the fusion protein (Yeasmin et al., 2018) .

Industrial Applications of Non-Coding RNA data

The main motive of functional characterization of micropeptides is to utilize them as pharmaceutical and agriculture targets. Their small sizes may facilitate easy purification and downstream processing. As the field of micropeptides is relatively nascent, less data is available thus far of scalable technologies utilizing micropeptides. Moreover, much of their biology still needs to be understood. There are however examples reported in the literature that illustrate micropeptide roles in various cellular biological processes in animals and plants (Wang et al., 2019)(Hartford & Lal, 2020)(Crappé et al., 2014)This information can form the basis of further

applying them on a commercial scale. The following section will provide details of micropeptides with respect to their use in industrial applications.

Diagnostic & Pharmaceutical Applications

CircRNA encoded micropeptides are translated as alternative spliced products of the maternal gene. Therefore it is possible to imagine peptides encoded by circRNA to act as competitive inhibitors to the proteins with which they are homologous (Wang et al., 2019) Since several circRNA peptides are implicated in cancer (Wu et al., 2020), it can be hypothesized that as competitive inhibitors, they may be used to block key signalling pathways disrupted in cancer and therefore have therapeutic value. For example SHPRH-146aa and FBXW7-185aa are circRNA encoded micropeptides; both of which act as competitive inhibitors of their parent genes SHPRH and c-myc respectively by a similar mechanism and inhibit cell proliferation (Wang et al., 2019). While some classes of micropeptides abrogate their homologous proteins, others act as domain specific modulators and cause a conformational change within binding partners via micropeptide secondary structures. For instance, the *D. Melanogaster* lncRNA tarsal-less (tal) encodes a 11-32aa pri peptide and mediates the binding of an E3 ligase to Shavenbaby, marking it for proteasomal degradation (Ruiz-Orera & Albà, 2019) . The tal gene has a vital role in tarsal morphogenesis in the fly leg and is conserved in metazoans (Li & Liu, 2019) Therefore it is clear that micropeptides can interact with larger proteins and have important roles in developmental processes, cell cycle and cancer. The ability of microproteins to modulate, interfere or change the conformation of larger proteins can be exploited while designing drugs that can specifically target these proteins in perturbed diseased states.

For debilitating diseases such as cancer, the discovery of novel biomarkers that facilitate early diagnosis and better prognosis is equally important. Using micropeptides as biomarkers is advantageous since they are more stable compared to their host lncRNA (Chakraborty et al.,

2019). Studies that probe the tissue specific versus ubiquitous expression of lncRNA encoded peptides can expedite the utility of micropeptides as biomarkers. Notably, Chakraborty et al established a computational proteogenomic workflow that was used to quantify abundance of lncRNA encoded micropeptides, in human tissues and cancer cell lines (Chakraborty et al., 2019). This study proved the benefits of reprocessing publicly available mass spectrometry data to probe for differentially expressed lncRNA peptides between cancer tissue, plasma samples and their normal counterparts (Chakraborty et al., 2019).

Cancer	coding RNA name	circRNA/lncRNA ID	Protein/Polypeptide	Translation driver	Function	Ref.
Glioblastoma	circFBXW7	novel_circ_022705	FBXW7-185aa	ORF, IRES	Induces cell cycle arrest, reduces proliferation in glioma cell, independent prognostic marker	Kong et al., 2020
Glioblastoma	circSHPRH	hsa_circ_0001649	SHPRH-146aa	ORF, IRES	Reduces malignant behaviour, prognostic marker	
Colon Cancer	circPPP1R12A	has_circ_0000423	circPPP1R12A_73aa	ORF, IRES	Promotes proliferation, migration and invasion of CC, potential therapeutic agent	
Colorectal Cancer	LncRNA HOXB-AS3	LncHOXB-AS3	HOXB-AS3 peptide	ORF	Suppresses CRC growth, diagnostic marker	Wu et al., 2020
Nasopharyngeal Carcinoma	Nobody lncRNA	LINC01420	nobody	ORF	Promotes carcinoma cell invasion, prognostic marker	
Non-small cell lung cancer	LINC00961 lncRNA	LINC00961	SPAR	ORF	Promotes lymph node metastasis, prognosis marker	Matsumoto et al., 2017

Table 3: Example of ncRNA encoded micropeptides implicated in different cancers. (Adapted from Kong et al., Wu et al., & Matsumoto et al.)

sORFs translated via non-AUG start codons in mammalian cells have recently been demonstrated to be functional in disease and stress states (Cao & Slavoff, 2020). For example, a study revealed 55% of sORFs were translated as revealed by ribosome profiling of LPS treated mouse macrophages. In most cases a non-canonical start codon was used for translation initiation. Notably, lncRNA Aw112010, translated sORF, utilized a CUG start codon

and conferred resistance in mice against *S. Typhimurium* infection (Cao & Slavoff, 2020). This study demonstrated

Agricultural Applications

Pri-miRNA encoded functional peptides have a regulatory function and promote the accumulation of their associated pri-miRNAs to down-regulate target genes (Couzigou et al., 2015) . Analysis of fifty pri-miRNA in *Arabidopsis thaliana* revealed the presence of at least one putative sORF encoded peptide (Lauressergues et al., 2015). Moreover, the peptides contained unique signatures suggesting that each of these micropeptides specifically regulate their associated miRNA (Lauressergues et al., 2015). The exogenous application of micropeptides can therefore be exploited to modulate the expression of target genes, critical to plant development (Couzigou et al., 2015) . Furthermore, genetic manipulation of non-model, agronomically important plants is not always feasible (Couzigou et al., 2015) . Hence targeting specific genes in such cases presents a unique opportunity to develop personalized treatments for such plants. Similarly, synthetic micropeptides have the potential to impart disease resistance in plants, stimulate plant-microbe symbiotic relationships, alter germination and flowering rates, and induce nutrient uptake (Zhang et al., 2013)

Finally, Cai et. al used differential gene expression analysis in chicken breast tissue and revealed lncRNA-Six1 to promote cell division and proliferation in muscle growth related genes (Cai et al., 2017). This example illustrates lncRNA encoded peptide utility in meat production thereby adding economic value to the poultry industry. (Cai et al., 2017)

Bifunctional mechanism

The central theme underlying bifunctionality of ncRNA assumes that the same RNA molecule can encode protein as well as retain its regulatory power as a ncRNA and in part may be true for a subset of molecules. (Hubé & Francastel, 2018) A case in point is p53, a key mediator of cell cycle control (Hubé & Francastel, 2018). p53 interaction with Mdm2, an E3 Ubiquitin ligase, marks it for degradation via the proteasomal pathway (Hubé & Francastel, 2018). However, under cellular stress, p53 mRNA interacts with Mdm2 and prevents its own degradation (Hubé & Francastel, 2018) suggesting that the same ORF has closely intertwined RNA and protein functions (Li & Liu, 2019) . Similarly, the presence of cis-acting factors in the UTRs of some protein coding genes suggests that bifunctionality could be an intrinsic property embedded within the mRNA (Li & Liu, 2019) . It is possible that competitive regulation exists between translation and structural regulation but the factors that promote either of these competitive states remain unclear (Hubé & Francastel, 2018). It may indicate that the bifunctional nature of RNA is an adaptation to physiological changes like cellular stress (Ulveling et al., 2011). Whether the regulatory RNA and the translated protein are both involved in this adaptation or whether their functions are independent can better explain the nature of bifunctionality. For example, SgrS mRNA encodes SgRT micropeptide in bacteria; both of which independently alleviate glucose-phosphate stress in a physiologically redundant but mechanistically distinct manner (Ulveling et al., 2011). This suggests that other bifunctional RNA and their expressed products may also have redundant functions and raises the question of how bifunctionality confers an evolutionary benefit to the cell.

Secondly, for several ncRNA, it is not the same molecule with dual functionality. Often different isoforms coming from the same locus are transcribed as products of alternative splicing or by the use of alternative promoters. If this is the case it is not the RNA that is bifunctional but in fact

the locus (Hubé & Francastel, 2018). Williamson et. al described a dual role for ASCC3 mRNA which is formed as a short isoform upon UV treatment, attenuating the transcription of the long form in response to UV damage (Williamson et al., 2017) This implies that there is crosstalk between the coding and non-coding isoforms. Although the short isoform encodes a 13aa peptide, it is the RNA that has the effector functions and affects transcription recovery after DNA damage (Williamson et al., 2017). Therefore, it is necessary to first identify the protein product of the bifunctional RNA, as well as confirm whether the protein or the RNA has effector functions. As such micropeptides may not be expressed in detectable amounts and it may be difficult to conclusively predict their functions, leaving the problem of assigning effector functions, to either the protein or RNA, unresolved.

Finally, the spatio-temporal expression of the bifunctional transcripts can explain their functional relevance. Subcellular localization is an obvious parameter that may clearly demarcate bifunctional modality. Similarly, stage of development, environmental cues and pathogen insults may also provide additional information about the same (Hubé & Francastel, 2018). Therefore, a clear distinction between whether the acting molecule is a lncRNA or its encoded peptide or both will benefit their application in future therapies (Choi et al., 2019) .

Identification of targets

One of the first steps in utilizing micropeptides for commercial purposes lies in the ability to probe for binding partners. Target identification is achieved by generating antibodies against the peptides, to probe for interacting partners using in-vitro pull down assays. (Hartford & Lal, 2020) This is one of the gold standard methods to elucidate mechanisms of action in the cell and is particularly important in identifying signaling pathways that the micropeptide may be involved in (Peeters & Menschaert, 2020). The small size of the micropeptides makes generating antibodies against specific epitopes difficult. Similarly, micropeptides localize to cellular membranes

(Makarewich, 2020) which hinder epitope design. The matter may be further complicated as micropeptides are as such expressed in low quantities in the cell compared to mRNA encoded proteins (Peeters & Menschaert, 2020). Using CRISPR/CAS9 technology to insert an epitope to determine localization and endogenous expression in the cell may overcome this problem (Makarewich & Olson, 2017). But the addition of a tag to an already small protein may have undesired consequences when it comes to folding, subcellular localization, and as such influence the interacting partners. As such it is not clear whether experiments to identify micropeptides should be case specific or may be applied to a broad spectrum of candidates and is an area of future research.

Functional prediction

A large proportion of micropeptides are predicted to contain transmembrane α -helix motifs, indicating that some of these microproteins are targeted to biological membranes (Makarewich, 2020). Moreover, GO data for lncRNA containing sORFs, display an enrichment of membrane-related terms (Aspden et al., 2014) further support membrane localization. Due to their small sizes, it was hypothesized that can fit into larger proteins and regulate membrane complexes (Makarewich, 2020). But functional micropeptides that have not yet been characterized may localize to membranes which may make biochemical detection difficult (Couso & Patraquim, 2017). Matsumoto et. al, reported SPAR, a lncRNA LINC00961 encoded polypeptide to localize to the late endosome and negatively interact with mTORC (Matsumoto et al., 2017). But the interaction was specifically induced upon amino-acid starvation suggesting that micropeptide respond to external stimuli and that micropeptide requirement may not be immediately obvious. Detecting the expression of micropeptides in cells under resting or in an unstressed state may result in many of them being undetected (Wang et al., 2019). Therefore, predicting special cases that will enrich micropeptide expression may preclude further characterization. Finally,

Jorge Ruiz-Orera et. al proposed a model to show sORF within lncRNA as “*de novo* genes”, in the process of acquiring new functions (Ruiz-Orera et al., 2020). The model suggested sORF translation events may be dispensable, and the expression of transcripts may not persist over time (Ruiz-Orera et al., 2020). Over time selective advantage and subsequent maturation of these evolving populations may result in fully functional stable peptides. This implies that sORFs are continuously subjected to evolutionary forces (Couso & Patraquim, 2017). Hence functional peptides that have arisen *de novo* may require new tools and new technologies for their identification and characterization (Ruiz-Orera et al., 2020).

Perspective

Micropeptides represent a novel class of promising molecules for two reasons: First, they are products of previously annotated non-coding RNA classes which suggests that we need to reevaluate classical concepts of RNA biology. Secondly as we have seen in this review, identifying and functionally characterizing micropeptides is not as simple as canonical proteins. Moreover, micropeptides and the sORFs that encode them add an additional layer of complexity to the proteome. But as a viable source of small lab grown molecules with practical applications it is imperative that we develop sensitive and specific tools and technologies that will aid in their identification. Further as a future research perspective the evolutionary roles of micropeptides may explain genome complexity in higher organisms and is worth investigating. Finally, the discovery of sORF encoded micropeptides provides insights into important cellular processes and broadens our understanding of gene expression.

References

- Anderson, D. M., Anderson, K. M., Chang, C., Makarewich, C. A., Nelson, B. R., Mcanally, J. R., Kasaragod, P., Shelton, J. M., Liou, J., Bassel-duby, R., & Olson, E. N. (2016). *Regulates Muscle Performance*. *160*(4), 595–606. <https://doi.org/10.1016/j.cell.2015.01.009.A>
- Andrews, S. J., & Rothnagel, J. A. (2014a). Emerging evidence for functional peptides encoded by short open reading frames. *Nature Reviews Genetics*, *15*(3), 193–204. <https://doi.org/10.1038/nrg3520>
- Aspden, J. L., Eyre-Walker, Y. C., Phillips, R. J., Amin, U., Mumtaz, M. A. S., Brocard, M., & Couso, J. P. (2014). Extensive translation of small open reading frames revealed by poly-ribo-seq. *ELife*, *3*(August2014), 1–19. <https://doi.org/10.7554/eLife.03528>
- Bazzini, A. A., Johnstone, T. G., Christiano, R., Mackowiak, S. D., Obermayer, B., Fleming, E. S., Vejnar, C. E., Lee, M. T., Rajewsky, N., Walther, T. C., & Giraldez, A. J. (2014). Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO Journal*, *33*(9), 981–993. <https://doi.org/10.1002/emboj.201488411>
- Bi, P., Ramirez-Martinez, A., Li, H., Cannavino, J., McAnally, J. R., Shelton, J. M., Sánchez-Ortiz, E., Bassel-Duby, R., & Olson, E. N. (2017). Control of muscle formation by the fusogenic micropeptide myomixer. *Science*, *356*(6335), 323–327. <https://doi.org/10.1126/science.aam9361>
- Cai, B., Li, Z., Ma, M., Wang, Z., Han, P., Abdalla, B. A., Nie, Q., & Zhang, X. (2017). LncRNA-Six1 encodes a micropeptide to activate Six1 in Cis and is involved in cell proliferation and muscle growth. *Frontiers in Physiology*, *8*(APR), 1–13. <https://doi.org/10.3389/fphys.2017.00230>
- Cao, X., & Slavoff, S. A. (2020). Non-AUG start codons: Expanding and regulating the small and alternative ORFeome. *Experimental Cell Research*, *391*(1), 111973. <https://doi.org/10.1016/j.yexcr.2020.111973>
- Chakraborty, S., Andrieux, G., Hasan, A. M. M., Ahmed, M., Hosen, M. I., Rahman, T., Hossain, M. A., & Boerries, M. (2019). Harnessing the tissue and plasma lncRNA-peptidome to discover peptide-based cancer biomarkers. *Scientific Reports*, *9*(1), 1–17. <https://doi.org/10.1038/s41598-019-48774-1>
- Chang, K. J., & Wang, C. C. (2004). Translation Initiation from A Naturally Occurring Non-AUG Codon in *Saccharomyces cerevisiae*. *Journal of Biological Chemistry*, *279*(14), 13778–13785. <https://doi.org/10.1074/jbc.M311269200>

Chen, Q., Deng, B., Gao, J., Zhao, Z., Chen, Z., Song, S., Wang, L., Zhao, L., Xu, W., Zhang, C., Ma, C., & Wang, S. (2020). A miRNA-Encoded Small Peptide, vvi-miPEP171d1, Regulates Adventitious Root Formation. *Plant Physiology*, *183*(2), 656–670. <https://doi.org/10.1104/pp.20.00197>

Choi, S. W., Kim, H. W., & Nam, J. W. (2019). The small peptide world in long noncoding RNAs. *Briefings in Bioinformatics*, *20*(5), 1853–1864. <https://doi.org/10.1093/bib/bby055>

Chooniedass-Kothari, S., Emberley, E., Hamedani, M. K., Troup, S., Wang, X., Czosnek, A., Hube, F., Mutawe, M., Watson, P. H., & Leygue, E. (2004). The steroid receptor RNA activator is the first functional RNA encoding a protein. *FEBS Letters*, *566*(1–3), 43–47. <https://doi.org/10.1016/j.febslet.2004.03.104>

Chugunova, A., Navalayeu, T., Dontsova, O., & Sergiev, P. (2018). Mining for Small Translated ORFs. *Journal of Proteome Research*, *17*(1), 1–11. <https://doi.org/10.1021/acs.jproteome.7b00707>

Couso, J. P., & Patraquim, P. (2017). Classification and function of small open reading frames. *Nature Reviews Molecular Cell Biology*, *18*(9), 575–589. <https://doi.org/10.1038/nrm.2017.58>

Couzigou, J. M., Laressergues, D., Bécard, G., & Combie, J. P. (2015). miRNA-encoded peptides (miPEPs): A new tool to analyze the roles of miRNAs in plant biology. *RNA Biology*, *12*(11), 1178–1180. <https://doi.org/10.1080/15476286.2015.1094601>

Crappé, J., van Crielinge, W., & Menschaert, G. (2014b). Little things make big things happen: A summary of micropeptide encoding genes. *EuPA Open Proteomics*, *3*, 128–137. <https://doi.org/10.1016/j.euprot.2014.02.006>

Derrien, Thomas, Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J. B., Lipovich, L., Gonzalez, J. M., ... Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research*, *22*(9), 1775–1789. <https://doi.org/10.1101/gr.132159.111>

di Timoteo, G., Dattilo, D., Centrón-Broco, A., Colantoni, A., Guarnacci, M., Rossi, F., Incarnato, D., Oliviero, S., Fatica, A., Morlando, M., & Bozzoni, I. (2020). Modulation of circRNA Metabolism by m6A Modification. *Cell Reports*, *31*(6), 107641. <https://doi.org/10.1016/j.celrep.2020.107641>

Dinger, M. E., Pang, K. C., Mercer, T. R., & Mattick, J. S. (2008). Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities. *PLoS Computational Biology*, *4*(11), e1000176. <https://doi.org/10.1371/journal.pcbi.1000176>

D'Lima, N. G., Ma, J., Winkler, L., Chu, Q., Loh, K. H., Corpuz, E. O., Budnik, B. A., Lykke-Andersen, J., Saghatelian, A., & Slavoff, S. A. (2017). A human microprotein that interacts with the mRNA decapping complex. *Nature Chemical Biology*, *13*(2), 174–180. <https://doi.org/10.1038/nchembio.2249>

Friesen, M., Warren, C. R., Yu, H., Toyohara, T., Ding, Q., Florido, M. H. C., Sayre, C., Pope, B. D., Goff, L. A., Rinn, J. L., & Cowan, C. A. (2020). Mitoregulin Controls β -Oxidation in Human and Mouse Adipocytes. *Stem Cell Reports*, *14*(4), 590–602. <https://doi.org/10.1016/j.stemcr.2020.03.002>

Hanada, Kousuke, Higuchi-Takeuchi, M., Okamoto, M., Yoshizumi, T., Shimizu, M., Nakaminami, K., Nishi, R., Ohashi, C., Iida, K., Tanaka, M., Horii, Y., Kawashima, M., Matsui, K., Toyoda, T., Shinozaki, K., Seki, M., & Matsui, M. (2013). Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(6), 2395–2400. <https://doi.org/10.1073/pnas.1213958110>

Hartford, C. C. R., & Lal, A. (2020). When Long Noncoding Becomes Protein Coding. *Molecular and Cellular Biology*, *40*(6). <https://doi.org/10.1128/mcb.00528-19>

Hon, C., Ramilowski, J., Harshbarger, J., Bertin, N., Rackham, O., & Gough, J. et al. (2017). An atlas of human long non-coding RNAs with accurate 5' ends. *Nature*, *543*(7644), 199–204. <https://doi.org/10.1038/nature21374>

Huarte, M. (2013). LncRNAs have a say in protein translation. *Cell Research*, *23*(4), 449–451. <https://doi.org/10.1038/cr.2012.169>

Hubé, F., & Francastel, C. (2018). Coding and non-coding RNAs, the frontier has never been so blurred. In *Frontiers in Genetics* (Vol. 9, Issue APR). Frontiers Media S.A. <https://doi.org/10.3389/fgene.2018.00140>

Ingolia, N. T., Lareau, L. F., & Weissman, J. S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, *147*(4), 789–802. <https://doi.org/10.1016/j.cell.2011.10.002>

Ji, Z., Song, R., Regev, A., & Struhl, K. (2015). Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *ELife*, *4*(DECEMBER2015). <https://doi.org/10.7554/eLife.08890>

Kong, S., Tao, M., Shen, X., & Ju, S. (2020). Translatable circRNAs and lncRNAs: Driving mechanisms and functions of their translation products. *Cancer Letters*, *483*(20), 59–65. <https://doi.org/10.1016/j.canlet.2020.04.006>

Lefever, S., Anckaert, J., Volders, P. J., Luybaert, M., Vandesompele, J., & Mestdagh, P. (2017). decodeRNA- predicting non-coding RNA functions using guilt-by-association.

Database : *The Journal of Biological Databases and Curation*, 2017, 1–8.

<https://doi.org/10.1093/database/bax042>

Li, J., & Liu, C. (2019b). Coding or noncoding, the converging concepts of RNAs. *Frontiers in Genetics*, 10(MAY), 1–10. <https://doi.org/10.3389/fgene.2019.00496>

Li, Y., Xu, J., Shao, T., Zhang, Y., Chen, H., & Li, X. (2017). RNA function prediction. In *Methods in Molecular Biology* (Vol. 1654, pp. 17–28). Humana Press Inc.

https://doi.org/10.1007/978-1-4939-7231-9_2

Losko, M., Kotlinowski, J., & Jura, J. (2016). Long noncoding RNAs in metabolic syndrome related disorders. *Mediators of Inflammation*, 2016. <https://doi.org/10.1155/2016/5365209>

Mackowiak, S. D., Zauber, H., Bielow, C., Thiel, D., Kutz, K., Calviello, L., Mastrobuoni, G., Rajewsky, N., Kempa, S., Selbach, M., & Obermayer, B. (2015). Extensive identification and analysis of conserved small ORFs in animals. *Genome Biology*, 16(1).

<https://doi.org/10.1186/s13059-015-0742-x>

Makarewich, C. A. (2020). The hidden world of membrane microproteins. *Experimental Cell Research*, 388(2). <https://doi.org/10.1016/j.yexcr.2020.111853>

Makarewich, C. A., & Olson, E. N. (2017b). Mining for Micropeptides. In *Trends in Cell Biology* (Vol. 27, Issue 9, pp. 685–696). Elsevier Ltd.

<https://doi.org/10.1016/j.tcb.2017.04.006>

Markus Wolfien, David Leon Brauer, A. B., & Wolkenhauer, and O. (2019). *Computational Biology of Non-Coding RNA*. 1912(January). <https://doi.org/10.1007/978-1-4939-8982-9>

Matsumoto, A., Pasut, A., Matsumoto, M., Yamashita, R., Fung, J., Monteleone, E., Saghatelian, A., Nakayama, K. I., Clohessy, J. G., & Pandolfi, P. P. (2017). MTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature*, 541(7636), 228–232. <https://doi.org/10.1038/nature21034>

Nam, J. W., Choi, S. W., & You, B. H. (2016). Incredible RNA: Dual functions of coding and noncoding. In *Molecules and Cells* (Vol. 39, Issue 5, pp. 367–374). Korean Society for Molecular and Cellular Biology. <https://doi.org/10.14348/molcells.2016.0039>

Niu, L., Lou, F., Sun, Y., Sun, L., Cai, X., Liu, Z., Zhou, H., Wang, H., Wang, Z., Bai, J., Yin, Q., Zhang, J., Chen, L., Peng, D., Xu, Z., Gao, Y., Tang, S., Fan, L., & Wang, H. (2020). A micropeptide encoded by lncRNA MIR155HG suppresses autoimmune inflammation via modulating antigen presentation. *Science Advances*, 6(21).

<https://doi.org/10.1126/sciadv.aaz2059>

Peeters, M. K. R., & Menschaert, G. (2020). The hunt for sORFs: A multidisciplinary strategy. *Experimental Cell Research*, 391(1). <https://doi.org/10.1016/j.yexcr.2020.111923>

- Ruiz-Orera, J., & Albà, M. M. (2019). Conserved regions in long non-coding RNAs contain abundant translation and protein–RNA interaction signatures. *NAR Genomics and Bioinformatics*, 1(1), e2–e2. <https://doi.org/10.1093/nargab/lqz002>
- Ruiz-Orera, J., & Albà, M. M. (2019). Translation of Small Open Reading Frames: Roles in Regulation and Evolutionary Innovation. In *Trends in Genetics* (Vol. 35, Issue 3, pp. 186–198). Elsevier Ltd. <https://doi.org/10.1016/j.tig.2018.12.003>
- Ruiz-Orera, J., Villanueva-Cañas, J. L., & Albà, M. M. (2020). Evolution of new proteins from translated sORFs in long non-coding RNAs. *Experimental Cell Research*, 391(1), 111940. <https://doi.org/10.1016/j.yexcr.2020.111940>
- Santulli, G. (2015). *microRNA: Basic Science*. 887, 79–100. <https://doi.org/10.1007/978-3-319-22380-3>
- Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., Karger, A. D., Budnik, B. A., Rinn, J. L., & Saghatelian, A. (2013). Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nature Chemical Biology*, 9(1), 59–64. <https://doi.org/10.1038/nchembio.1120>
- Stein, C. S., Jadiya, P., Zhang, X., McLendon, J. M., Abouassaly, G. M., Witmer, N. H., Anderson, E. J., Elrod, J. W., & Boudreau, R. L. (2018). Mitoregulin: A lncRNA-Encoded Microprotein that Supports Mitochondrial Supercomplexes and Respiratory Efficiency. *Cell Reports*, 23(13), 3710–3720.e8. <https://doi.org/10.1016/j.celrep.2018.06.002>
- St.Laurent, G., Wahlestedt, C., & Kapranov, P. (2015). The Landscape of long noncoding RNA classification. In *Trends in Genetics* (Vol. 31, Issue 5, pp. 239–251). Elsevier Ltd. <https://doi.org/10.1016/j.tig.2015.03.007>
- Tang, R., Long, T., Lui, K. O., Chen, Y., & Huang, Z. P. (2020). A Roadmap for Fixing the Heart: RNA Regulatory Networks in Cardiac Disease. *Molecular Therapy - Nucleic Acids*, 20(June), 673–686. <https://doi.org/10.1016/j.omtn.2020.04.007>
- Tripathi, R., Chakraborty, P., & Varadwaj, P. K. (2017). Unraveling long non-coding RNAs through analysis of high-throughput RNA-sequencing data. In *Non-coding RNA Research* (Vol. 2, Issue 2, pp. 111–118). KeAi Communications Co. <https://doi.org/10.1016/j.ncrna.2017.06.003>
- Ulveling, D., Francastel, C., & Hubé, F. (2011). When one is better than two: RNA with dual functions. *Biochimie*, 93(4), 633–644. <https://doi.org/10.1016/j.biochi.2010.11.004>
- Wang, L., Fan, J., Han, L., Qi, H., Wang, Y., Wang, H., Chen, S., Du, L., Li, S., Zhang, Y., Tang, W., Ge, G., Pan, W., Hu, P., & Cheng, H. (2020). The micropeptide LEMP plays an evolutionarily conserved role in myogenesis. *Cell Death and Disease*, 11(5). <https://doi.org/10.1038/s41419-020-2570-5>

- Wang, S., Mao, C., & Liu, S. (2019). Peptides encoded by noncoding genes: Challenges and perspectives. *Signal Transduction and Targeted Therapy*, 4(1), 1–12. <https://doi.org/10.1038/s41392-019-0092-3>
- Wang, Y., & Wang, Z. (2015). Efficient backsplicing produces translatable circular mRNAs. *Rna*, 21(2), 172–179. <https://doi.org/10.1261/rna.048272.114>
- Williamson, L., Saponaro, M., Boeing, S., East, P., Mitter, R., Kantidakis, T., Kelly, G. P., Lobley, A., Walker, J., Spencer-Dene, B., Howell, M., Stewart, A., & Svejstrup, J. Q. (2017). UV Irradiation Induces a Non-coding RNA that Functionally Opposes the Protein Encoded by the Same Gene. *Cell*, 168(5), 843-855.e13. <https://doi.org/10.1016/j.cell.2017.01.019>
- Wolfien, M., Brauer, D. L., Bagnacani, A., & Wolkenhauer, O. (2019). Workflow development for the functional characterization of ncRNAs. In *Methods in Molecular Biology* (Vol. 1912, pp. 111–132). Humana Press Inc. https://doi.org/10.1007/978-1-4939-8982-9_5
- Wu, P., Mo, Y., Peng, M., Tang, T., Zhong, Y., Deng, X., Xiong, F., Guo, C., Wu, X., Li, Y., Li, X., Li, G., Zeng, Z., & Xiong, W. (2020a). Emerging role of tumor-related functional peptides encoded by lncRNA and circRNA. *Molecular Cancer*, 19(1), 1–14. <https://doi.org/10.1186/s12943-020-1147-3>
- Yao, R. W., Wang, Y., & Chen, L. L. (2019). Cellular functions of long noncoding RNAs. In *Nature Cell Biology* (Vol. 21, Issue 5, pp. 542–551). Nature Publishing Group. <https://doi.org/10.1038/s41556-019-0311-8>
- Yeasmin, F., Yada, T., & Akimitsu, N. (2018). Micropeptides encoded in transcripts previously identified as long noncoding RNAs: A new chapter in transcriptomics and proteomics. In *Frontiers in Genetics* (Vol. 9, Issue APR). Frontiers Media S.A. <https://doi.org/10.3389/fgene.2018.00144>
- Yin, X., Jing, Y., & Xu, H. (2019). Mining for missed sORF-encoded peptides. In *Expert Review of Proteomics* (Vol. 16, Issue 3, pp. 257–266). Taylor and Francis Ltd. <https://doi.org/10.1080/14789450.2019.1571919>
- Yin, X., Hu, J., & Xu, H. (2018). Distribution of micropeptide-coding sORFs in transcripts. *Chinese Chemical Letters*, 29(7), 1029-1032. <https://doi.org/10.1016/j.ccllet.2018.04.027>
- Yongsheng Li, Juan Xu, Tingting Shao, Yunpeng Zhang, Hong Chen, and X. L. (2017). RNA Function Prediction. *Functional Genomics: Methods and Protocols*, 1654, 151–164. <https://doi.org/10.1007/978-1-4939-7231-9>

Zhang, J., Mujahid, H., Hou, Y., Nallamilli, B. R., & Peng, Z. (2013). Plant Long ncRNAs: A New Frontier for Gene Regulatory Control. *American Journal of Plant Sciences*, *04*(05), 1038–1045. <https://doi.org/10.4236/ajps.2013.45128>

Zhao, Z., Bai, J., Wu, A., Wang, Y., Zhang, J., Wang, Z., Li, Y., Xu, J., & Li, X. (2015). Co-LncRNA: Investigating the lncRNA combinatorial effects in GO annotations and KEGG pathways based on human RNA-Seq data. *Database*, *2015*, 1–7. <https://doi.org/10.1093/database/bav082>

Zheng, G. zhen, Li, W., & Liu, Z. yong. (2019a). Alternative role of noncoding RNAs: coding and noncoding properties. In *Journal of Zhejiang University: Science B* (Vol. 20, Issue 11, pp. 920–927). Zhejiang University Press. <https://doi.org/10.1631/jzus.B1900336>