Functions of long non-coding RNAs in human disease and their conservation in Drosophila development

Article  (Accepted Version)

Rogoyski, Oliver M, Pueyo-Marques, Jose, Couso, Juan Pablo and Newbury, Sarah F (2017)
Functions of long non-coding RNAs in human disease and their conservation in Drosophila development. Biochemical Society Transactions, 45 (4). pp. 895-904. ISSN 0300-5127

This version is available from Sussex Research Online: http://sro.sussex.ac.uk/id/eprint/68277/

This document is made available in accordance with publisher policies and may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher’s version. Please see the URL above for details on accessing the published version.

Copyright and reuse:
Sussex Research Online is a digital repository of the research output of the University.

Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable, the material made available in SRO has been checked for eligibility before being made available.

Copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.
Functions of long non-coding RNAs in human disease and their conservation in
*Drosophila development.*

Oliver M. Rogoyski¹, Jose Ignacio Pueyo¹, Juan Pablo Couso¹2 and Sarah F. Newbury¹*  

¹ Brighton and Sussex Medical School, Medical Research building, University of Sussex, Falmer, Brighton, BN1 9PS, U.K.

² Centro Andaluz de Biología de Desarrollo, CSIC-Universidad Pablo de Olavide, Ctra. de Utrera, Km.1, Sevilla, 41013, Spain

*Corresponding author: s.newbury@bsms.ac.uk +44 (0)1273 877874

Abstract

Genomic analysis has found that the transcriptome in both humans and *Drosophila melanogaster* features large numbers of long non-coding RNA transcripts (lncRNAs). This recently discovered class of RNAs regulates gene expression in diverse ways, and has been involved in a large variety of important biological functions. Importantly, an increasing number of lncRNAs have also been associated with a range of human diseases, including cancer. Comparative analyses of their functions among these organisms suggest that some of their modes of action appear to be conserved. This highlights the importance of model organisms such as *Drosophila*, which shares many gene regulatory networks with humans, in understanding lncRNA function and its possible impact in human health. This review discusses some known functions and mechanisms of action of lncRNAs and their implication in human diseases, together with their functional conservation and relevance in *Drosophila* development.

Introduction

The central dogma of molecular biology as proposed by Crick in 1958, often paraphrased as “DNA encodes RNA, RNA encodes protein”, implicates RNA as a molecular intermediate in the process of protein synthesis from the relevant encoding gene. As early as the 1950s however, other roles for non-coding RNAs, such as transfer RNAs and ribosomal RNAs, have been known to be vital to biological function. This showed the central dogma to be an over-simplified, if eloquent, summary of the flow of genetic information. Since then, many other types of non-coding RNA have been shown to exist, and furthermore, to be biologically relevant. In the 1990s, several studies began investigating the biological purpose of longer non-protein-coding RNAs, such as Xist [1], which did not fit well into the RNA classifications existing at the time. With further advances in molecular techniques suggesting that only 2% of the human genome is comprised of protein-coding genes [2], and rapidly revealing lncRNAs with biological functions (including in human diseases), the topic has become an extremely promising and popular avenue of investigation.

In this review, we have used the definition of lncRNAs as being RNA transcripts longer than 200 nucleotides, which lack a significant open reading frame (greater than 100 amino acids in length) [3]. This definition is routinely used in the annotation of the *Drosophila* and other genomes. LncRNAs are highly abundant, and are found in many organisms across different taxa, including humans, mouse, *Xenopus tropicalis*, *Drosophila melanogaster*, *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Medicago truncatula*, and *Zea mays* [4]. LncRNAs have been shown to regulate gene expression transcriptionally [5-8] and post-transcriptionally [9-13], and have a wide
range of cellular and molecular functions. Despite these proven non-coding functions, there exist a handful of IncRNAs that have been shown to encode small open reading frame (smORF) peptides with proven cellular functions [14-19]. Recent work has shown that IncRNAs can simultaneously display biological function as both a coding, and a non-coding RNA, for example where primary transcripts of microRNAs encode regulatory peptides [20, 21]. Additionally, ribosome profiling and bioinformatics analyses have identified the existence of thousands of IncRNAs containing putatively functional translated smORFs [19, 22-25], the extent of which may depend on developmental or tissue specific context. We have therefore used the accepted definition above, which coincides with genome annotations.

*Drosophila melanogaster*, the common fruit fly, is a well-established model organism for geneticists, and one in which IncRNAs are known to be abundant. With an estimated 75% of human disease-linked genes having a functional orthologue in *Drosophila*, and many basic molecular and biological functions conserved between species [26, 27], *Drosophila* are an appealing whole animal model for understanding human disease. In addition to their genetic similarities, the fly genome has been incredibly well studied and fully sequenced, with a wide range of genetic tools and gene-specific knockdown and mutant lines readily available. Combined with their low maintenance cost, short generation time, high fecundity, and compound factors lending themselves to ease of establishing genetic crosses, it is easy to see why *Drosophila* have emerged as one of the foremost systems for studying the genetic components of human disease, and have already been successfully used to dissect the roles and mechanisms of certain IncRNAs [28].

As well as the general excellence of *Drosophila* as a model organism, they stand out as particularly apt for the study of IncRNA. LncRNAs evolve rapidly, and can act as flexible scaffolds tethering together one or more functional elements [29]. *Drosophila* IncRNAs also appear to accumulate relatively few deleterious changes, due to genetic drift, compared to mammalian IncRNAs [30], and therefore can be useful in developing strategies to identify IncRNA orthologues, as shown for *roX* IncRNA orthologues in Drosophilid species [31]. Additionally, *Drosophila* is an excellent model system to functionally characterise IncRNA-protein complexes, for example by using the GAL4-UAS system to express IncRNAs in specific tissues or by characterising the localisation of RNA-proteins within cells (e.g. 7SK snRNA [32]).

Molecular functions and mechanisms of IncRNAs, such as their binding to protein complexes, definitively need to be tested in vivo in order to be well characterized. For example, in vivo experiments have shown that only the IncRNA transcribed in the reverse direction from the Polycomb/Trithorax response elements can bind the the Polycomb Repressive Complex 2 component Enhancer of Zeste, which provides the critical Histone Methyl Transferase activity required for transcriptional silencing. This level of understanding of such complex mechanisms and interactions would be extremely difficult to achieve without the use of a tractable in vivo system such as that provided by *Drosophila*.

In this review, we will be examining the emerging roles and relevance of IncRNAs using recent work illustrating their biological and molecular functions in *Drosophila*. We aim to examine these recent advances in our understanding of IncRNAs through the lens of their potential relevance to humans, and particularly human disease. By doing so, we hope to provide a concise synopsis of the topic, and demonstrate the value of using *Drosophila* as a model organism for understanding the roles of IncRNAs at molecular and cellular levels, and their implications in human disease.

**Abundance and localisation of IncRNAs in the human and Drosophila genomes**
According to the Ensembl database, lncRNAs comprise 7841 of the 63898 annotated genes in the human genome, and 2366 of the 17559 in the *Drosophila* genome. In both species, they account for a similar and substantial proportion of the entire genome (12.4% and 13.5% respectively). Although only a fraction of these have been investigated experimentally, information on their sequences and loci are readily available through various genomic databases, both non-specific (such as Ensembl), and dedicated non-coding RNA databases (such as LNCipedia, IncRNome, and IncRNAdb). Additionally, significant bioinformatic work has been carried out on them in terms of their expression and conservation within and across species [33].

With so much information on lncRNA now available, exploring this class of genes with a thorough experimental approach has become more feasible in recent years.

LncRNAs vary significantly in their distribution throughout cellular compartments, with the majority of transcripts residing predominantly in the nucleus, others in the cytoplasm, and some distributed more evenly between the two [34, 35]. For example, the *roX* transcripts in *Drosophila* are found in the nucleus, while *yar* is cytoplasmic [35]. The localisation of lncRNAs can give clues about their function; in the case of a chromatin restructuring lncRNA such as *roX1* or *roX2* it must be nuclear in order to access the chromatin. Localisation of particular lncRNAs can also affect their susceptibility to suppression by RNA interference and antisense oligonucleotides. An example of this is the suppression of nuclear lncRNAs *MALAT1* and *NEAT1* which in humans is more efficient using antisense methods, whereas cytoplasmic lncRNAs *DANCR* and *OIP5-AS1* are better suppressed with RNAi methods [35].

However, the sub-cellular localisation of the majority of lncRNAs has not been well characterised, with the localisation of relatively few being experimentally visualised. Single molecule RNA fluorescence *in situ* hybridisation has now been used to give high resolution data for the distribution of lncRNAs in human cells [34], and a systematic investigation of lncRNA localisation has been suggested as an important next step in expanding our understanding of their function; as well as a useful way to shed light on the potential relevance of lncRNAs to a particular mechanism.

**LncRNA in human disease**

LncRNAs have now been implicated as important factors linked to a range of human diseases. The broad range of biological functions of lncRNAs is reflected in the variety of different pathologies in which their aberrant expression is thought to be a contributing factor. Many lncRNAs have been shown to either be expressed at aberrant levels in cancerous cells [36-67], or their levels shown to affect the growth and behaviour of cancerous cells [46, 47, 49, 50, 52-56] (Table 1). This has prompted speculation that if better characterised, this class of genes may present many promising biomarkers, and even novel potential therapeutic targets. We cannot comprehensively cover this topic within the scope of this review, and point the reader to a comprehensive review of the topic for more information [57], but instead demonstrate this point with two well documented examples, below.

*MALAT1*, a highly conserved mammalian lncRNA, has been found to be overexpressed in human osteosarcoma cells and cell lines [46, 47]. It is hypothesised to function as a molecular scaffold for ribonucleoprotein complexes, acting as a transcriptional regulator for certain genes. Higher levels of *MALAT1* have been shown to be associated with “aggressive” cancer traits such as increased migration, metastasis, and clonogenic growth in non-small cell lung cancer [36-38] pancreatic [58], and prostate cancer cells [39]. Indeed, inducing a knockdown of *MALAT1* in osteosarcoma cell lines inhibited cell proliferation and invasion [46, 47].
The HOTAIR IncRNA, transcribed from an antisense Hox gene, plays an important role in the epigenetic regulation of genes thought to be due to its interactions with the Polycomb Repressive Complex 2 (PRC2) [43, 59], although recent work has indicated that PRC2 recruitment may be a downstream consequence of gene silencing, rather than initiating it [68]. HOTAIR is thought to act as a molecular scaffold, and is required for histone modification of particular genes across different chromosomes. Higher levels of HOTAIR have been found in colorectal cancer tissues, and are associated with increased tumour invasion, metastasis, vascular invasion, advanced tumour stage, and a worse prognosis in patients [43, 44]. HOTAIR has since been suggested for use as a biomarker for the progression and prognosis of certain cancers [44].

A Drosophila homologue for HOTAIR has not been identified, but given the similarities in polycomb regulation between species, it is likely that a targeted search might reveal such an equivalent.

Aside from cancer, strong evidence now exists linking certain IncRNAs to certain neurological pathologies [60]. LncRNAs have been shown to be relevant factors in amyotrophic lateral sclerosis, multiple sclerosis [61, 62], Alzheimer’s disease [10, 63], Huntington’s disease [64, 65], and Parkinson’s disease, among others. For example, the BACE1 antisense transcript (BACE1-AS) regulates mRNA stability of BACE1, a key enzyme in Alzheimer’s disease pathology [10]. This subsequently affects amyloid-β 1-42 abundance, the increased expression of which is a hallmark of Alzheimer’s disease. One mechanism by which IncRNAs have been hypothesized to impact neurodegenerative disease is through their induction of R-loop formation (which may be triggered by trinucleotide repeat expansion). R-loops have been shown to be capable of controlling the fate of neuroprotective genes [69], and are thought to contribute to the pathogenesis of fragile X syndrome and Friedrich’s Ataxia [70, 71] by their silencing of certain genes. Additionally, work in S. pombe and Arabidopsis has suggested that R-loops may regulate IncRNA expression [72, 73], although whether this is true of IncRNAs linked to neurodegenerative diseases remains unclear. Trinucleotide repeats in IncRNAs are also known to be important in the pathogenesis of SCA8, by production of toxic noncoding CUG expansion RNAs from the ataxin 8 opposite strand (ATXN8OS), thought to cause a toxic gain of function at both the RNA and protein level [74, 75].

Another area of disease in which IncRNAs have been proven relevant is cardiovascular disease [66, 67]. Evidence now shows that IncRNAs are an important factor in susceptibility to coronary artery disease and myocardial infarction, prognosis in recovery from myocardial infarction, cardiovascular disease mortality, and heart failure [67]. Once again their correlations with prognosis and susceptibility have placed IncRNAs in the spotlight as a promising avenue of investigation in finding novel biomarkers.

Interestingly, Drosophila IncRNAs have been shown hold functional roles very relevant to these pathologies. Hsromea [76-80] and bft [81] are required for proper apoptosis process and cell differentiation, yar [82] and CRG [83] serve regulatory roles in the nervous system, and sclA and sclB are required for normal calcium transients and cardiac muscle contractility [19]. This is particularly promising given that these links can be made from the limited pool of Drosophila IncRNAs that have been experimentally characterised.

Molecular functions of IncRNA conserved in Drosophila

LncRNAs have been shown to function via a wide range of molecular mechanisms, falling under the broad categories of signals, molecular decoys, guide RNAs, or scaffolds [84]. Some IncRNAs have convincingly been shown to be translated, with the small peptide products (smORFs) having important biological functions [14-19, 22-25]. Through these various mechanisms (Figure 1), they have been implicated in
regulation of a diverse array of processes, such as differentiation, development, cell proliferation, nervous system function, and cardiovascular function in both *Drosophila* and humans, despite the lack of sequence conservation in lncRNAs across species. Importantly, similarities in the modes of action of lncRNAs have been found at the molecular level between organisms, discussed below.

**LncRNA in the regulation of chromatin structure and gene expression**

One of the most extensively studied molecular mechanisms of lncRNA modes of action is their role in sex chromosome dosage compensation pathways. Due to the difference in the number of X chromosome copies between males and females, there exists a compensation pathway required to maintain a similar level of expression for genes located on the X chromosome. In *Drosophila*, this is achieved by transcriptional hyperactivation of the single copy of the genes in males, allowing their expression at comparable levels to that given by the two copies of the gene found in females [85]. In humans, by contrast, the genes located on the X-chromosome in human females are partially transcriptionally repressed, giving a similar level of expression to that seen in males [86].

In *Drosophila*, the RNA on the X genes, *roX1* and *roX2*, are expressed in males, and regulate the assembly of the Male Specific Lethal (MSL) complex in *Drosophila*; a chromatin modifier that functions in histone modification [87-90]. The recruitment and binding of MSL proteins by high affinity sequences on the nascent *roX* transcripts covering the X chromosome allows the assembly of the active MSL complex, which can then spread in cis, allowing chromatin restructuring and hyperactivation of specific regions of the chromosome.

An immediate comparison can be made between the *roX* genes in *Drosophila*, and lncRNAs involved in the sex chromosome dosage compensation pathway in humans and other mammals; X-inactive specific transcript (*Xist*) and its antisense transcript, *Tsix*. Like the *roX* genes, *Xist* coats the X chromosome, where it regulates chromatin modifications, with consequent effects on the expression of particular target genes [91, 92]. Unlike *roX*, *Xist* is expressed in females, and regulates the inactivation of the X chromosome by facilitating the initiation and stabilising of the X chromosome inactivation process [86].

Although these lncRNA genes differ in their sequence, there are striking similarities between their role in specific regulation of the X-chromosome and the molecular mechanisms by which they are thought to achieve this. Interestingly, a subset of lncRNAs involved in chromatin looping, called topological anchor point RNAs (tapRNAs), have been identified in the human and mouse genomes, with conserved zinc-finger motifs capable of binding DNA and RNA [93]. Whether these are conserved in *Drosophila* has not yet been studied, but given the involvement of lncRNAs in *Drosophila* chromatin regulation so far, this may be a promising avenue to explore, and may reveal a wider conservation of this class of lncRNA chromatin regulators.

**LncRNAs in the production of small peptides**

The *Drosophila* sarcocolbamban (*scl*) gene, originally classified as a lncRNA *pncr003* [94], is transcribed into a 992 base-pair mRNA, which is translated to produce two related peptides of less than 30 amino acids [19]. The *scl* gene is expressed in muscle cells, and *scl* null mutants show arrhythmic cardiac contractions, a phenotype produced by abnormal intracellular calcium levels in contracting muscle cells [19].
Interestingly, the scl genes were found to have homologues in humans, namely *sarcolipin* (sln) and its longer parologue, *phospholamban* (pln), encoding peptides of 31 and 52 amino acids respectively [19]. Phylogenetic analysis suggests that these genes belong to the same gene family, derived from a single ancestral gene, conserved for more than 550 million years. Furthermore, their function also seems to be conserved, with Sln and Pln regulating calcium transport in mammalian muscle cells, via dampening of Sarco-endoplasmic Reticulum Ca\(^{2+}\) adenosine triphosphate (SERCA) pump function. Scl peptides were able to colocalise and interact with *Drosophila* SERCA. Exogenous expression of the human Pln and Sln peptides in *Drosophila* scl mutant muscle cells were sufficient to rescue muscle function. Importantly, aberrant levels of Sln in humans have been linked to heart arrhythmias [95]. Regulation of SERCA by micropeptides (encoded by IncRNAs) has been extensively exploited in mammals; with tissue specific positive and negative regulators being found [22, 96, 97]. In addition, the number of characterized IncRNA genes encoding micropeptides is rapidly increasing, with roles found in a myriad of essential, conserved cellular functions, from phagocytosis [17] and cellular motility [98] to RNA degradation [18]. Thus, these examples show that IncRNAs that produce biologically relevant peptides may be conserved in structure, function, and relevance to pathologies between humans and *Drosophila* [19, 22].

**Future directions**

As previously shown in *sarcolamban*, proving the protein-coding potential of IncRNAs is a painstaking process, and an extremely difficult topic to broach; with genes having previously been catalogued as “non-coding” by arbitrary rules. Definitively showing the translation, or lack thereof, of an RNA using experimental techniques can be an arduous process, making this approach impractical to apply to the entire catalogue of identified IncRNAs. Ribosome profiling (in which a protease digestion is used to degrade RNA not protected by a bound ribosome,) and polysome profiling (where RNAs are separated by the number of ribosomes that are attached to different transcripts) have been used to provide a translational snapshot for several IncRNAs so far. This data has given a profile for IncRNA translation, but the threshold for significant translation is difficult to define in a non-arbitrary fashion. Therefore, use of model organisms to determine the biological function of any particular IncRNA remains crucial to gaining a meaningful understanding of the function of these molecules. A thorough and processive approach to clarifying this aspect of the gene class, as well as standardising measures and cut-offs for translational activity is an important priority for those in the field.

Bioinformatic approaches to elucidating the possible biological functions of IncRNAs are also being developed, although this method is not without its difficulties. Due to the poor sequence conservation characteristic of IncRNAs, standard approaches used to identify biologically relevant transcripts by their conservation within and across species are significantly less effective within this gene class. However, recent work has noted distinctive selection patterns in IncRNAs based on secondary structure [99], which may be of help in future analyses.

To conclude, we suggest that the studies currently being carried out on IncRNA in *Drosophila* should be of interest to a far wider audience than just fly geneticists, having shown that as a model organism, *Drosophila* is a logical choice both for better characterising this gene class, and for precursor studies to highlight genes and mechanisms that can be carried forward into more expensive and laborious large animal and human work. The superb annotation of the *Drosophila* genome and transcriptome, coupled with further increases in RNA-sequencing data available, will provide a candidate pool of IncRNAs for a rapid functional characterization (using the sophisticated genetic tools available in *Drosophila*). Therefore, further IncRNA
studies in *Drosophila*, of a suitably high calibre, are likely to provide us not only with a better understanding of the basic science behind this gene class, but promise to highlight potential biomarkers, elucidate genetic mechanisms behind a range of diseases, and perhaps provide novel targets for next generation therapeutics.

**Abbreviations**

IncRNA, long non-coding RNA; MSL, Male Specific Lethal; pln, phospholamban; PRC2, Polycomb Repressive Complex 2; roX, RNA on the X; scl, sarcolamban; sln, sarcolipin; small open reading frame, smORF; tapRNA, topological anchor point RNA; Xist, X-inactive specific transcript.

**Funding**

This work was funded by a Brighton and Sussex Medical School studentship [grant number WB002-61] and the Biotechnology and Biological Sciences Research Council [grant number BB/N001753/1].

**Competing Interests**

The Authors declare that there are no competing interests associated with the manuscript.

**Acknowledgements**

The authors thank Benjamin Towler and Pedro Patraquim for their input and advice.
References


resistant and migration. The evolutionarily novel disease source of new peptides. Elife. 3, e03523


524 93  Nature. 521, 232-236
Table 1) A table summarising the lncRNAs linked to various kinds of cancer, as covered in this review.

<table>
<thead>
<tr>
<th>IncRNA</th>
<th>Associated disease</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>MALAT1</td>
<td>Osteosarcoma</td>
<td>[46, 47]</td>
</tr>
<tr>
<td></td>
<td>Non-small cell lung cancer</td>
<td>[36-38]</td>
</tr>
<tr>
<td></td>
<td>Prostate cancer</td>
<td>[39]</td>
</tr>
<tr>
<td></td>
<td>Pancreatic cancer</td>
<td>[58]</td>
</tr>
<tr>
<td>HOTAIR</td>
<td>Colorectal cancer</td>
<td>[43, 44]</td>
</tr>
<tr>
<td>EWSAT1</td>
<td>Ewing sarcoma</td>
<td>[53]</td>
</tr>
<tr>
<td>HOTTIP</td>
<td>Osteosarcoma</td>
<td>[52]</td>
</tr>
<tr>
<td>HIF2PUT</td>
<td>Osteosarcoma</td>
<td>[54]</td>
</tr>
<tr>
<td>ANCR</td>
<td>Osteosarcoma</td>
<td>[55]</td>
</tr>
<tr>
<td>TUSC7</td>
<td>Osteosarcoma</td>
<td>[50]</td>
</tr>
<tr>
<td>FGFR3-AS1</td>
<td>Osteosarcoma</td>
<td>[49]</td>
</tr>
<tr>
<td>SNHG12</td>
<td>Osteosarcoma</td>
<td>[56]</td>
</tr>
<tr>
<td>TUG1</td>
<td>Osteosarcoma</td>
<td>[51]</td>
</tr>
<tr>
<td>H19</td>
<td>Wilms tumour</td>
<td>[40]</td>
</tr>
<tr>
<td></td>
<td>Gastric cancer</td>
<td>[41, 42]</td>
</tr>
<tr>
<td>LINC00152</td>
<td>Gastric cancer</td>
<td>[45]</td>
</tr>
</tbody>
</table>
Figure 1) A cartoon depicting the molecular mechanisms by which IncRNAs can function.

a) Some IncRNAs (red), such as Xist and RoX1, can act to modulate expression of a certain gene by binding to a transcription modifier or chromatin modifier (purple). b) LncRNAs (red) such as HOTAIR can act as molecular scaffolds, allowing the assembly of protein complexes (teal, green, dark purple) with genetic regulatory roles e.g. polycomb complex PRC2. c) LncRNAs (red) can act as molecular decoys, to sequester miRNAs (orange) or proteins (purple). d) Alternatively, LncRNAs (red) can act as molecular decoys, occluding or removing transcription factors, proteins, or RNAs (purple) from their functional location. e) LncRNAs (red) can act as a molecular guide, allowing formation of ribonucleoprotein complexes (yellow) to specific target sites. f) It has also been shown that IncRNAs (blue as DNA, red as RNA) can be actively translated into functional smORF peptides (orange) such as the SclA and SclB peptides, which function in regulating calcium transport in cardiac muscle.