UNIVERSITAT DE BARCELONA

FACULTAT DE BIOLOGIA

DEPARTAMENT DE GENÈTICA

Programa de Doctorat en Genètica

Unravelling the Role of Long Noncoding RNAs in the Context of Cell-growth and Regeneration

Memòria presentada per

Raziel Amador Rios

per a optar al grau de Doctor per la

Universitat de Barcelona

Treball realitzat al Departament de Genètica, Microbiologia i Estadística de la Facultat de Biologia de la Universitat de Barcelona i al Centre de Regulació Genòmica (CRG)

Doctorand

Raziel Amador Rios

Co-Director

Montserrat Corominas

Universitat de Barcelona

Co-Director

Roderic Guigó

Centre de Regulació Genòmica

Barcelona, Novembre de 2021

Acknowledgments

I would like to express my profound gratitude to my thesis co-directors, Roderic Guigó and Montserrat Corominas, for their feedback, guidance, and for providing a research environment that was intellectually stimulating and in which I was able to improve my research skills. My work was also co-supervised by Assaf Bester, to whom I am also grateful for his guidance, and for the feedback in the second part of this study.

Besides Roderic, Montse and Assaf, I also thank Carlos Camilleri, for his remarkable scientific contributions to the improvement of this thesis. I am grateful to all members of Roderic's lab. The administrative support provided by Montse Ruano, and Romina Garrido have been exceptional. This work also benefited from the computational help of Emilio Palumbo.

Many thanks to the following people for useful feedback on various parts of the Thesis manuscript: Montserrat Corominas, Manuel Muñoz, David Brena, Reza Sodaei, Marc Elosua, Iman Sadeghi, and Roderic Guigó.

Additionally, the following non-exhaustive list of people, which positively impacted my PhD experience and this work:

Cecilia C. Klein, for your mentorship during the first year of PhD and share your insights in *Drosophila*, regeneration and bioinformatic. From you I learned how to combine biology and computational analysis to obtain novel results.

Manuel Muñoz, for making my adaptation process in the lab easier; help in statistics, plotting, programming, coding best practices; for introducing me into the Emacs world (the most efficient text editor/IDE) and, last but not least, the fun hackathon times we shared, we learned a lot. Your examples helped me realize the importance of hard working, curiosity-driven projects and never stop learning.

Reza Sodaei and Valentin Wucher, for giving the opportunity to collaborate with you in the circadian-seasonal manuscript. I learned why the science-core should be based on collaboration and use different expertise to solve a problem, and achieve a more complete conclusion.

Iman Sadeghi, for your friendship, all the adventures we had through our PhD years, personal advice, and time to time scientific counseling in this research. Julien Lagarde, it was very helpful in sharing his LaTeX code to everybody, using open-source software, providing thorough instructions of how his code works and helping when I needed it. Vasilis Ntasis, for the geek talks, sharing your linux/R tools and philosophy of mostly using the keyboard made me more productive and efficient.

Thanks to my love Vanessa Vega for her constant support, in graphic design, revisions, polishing plots, and otherwise.

Por último, agradezco a mis padres y hermanos por todo el apoyo, amor y proporcionarme las herramientas necesarias para ser quien soy. Sin ustedes esto no sería posible.

Abstract

Long noncoding RNAs (lncRNAs) have proven biological roles in plethora cellular contexts. Nonetheless, only a handful have been clearly characterized, leaving thousands of newly discovered lncRNAs without an associated function, and sometimes considered as transcriptional by-products. To this end, this thesis work had focused on exploring lncRNA functionality in two scenarios. First, in order to discern between lncRNAs affecting cell-growth rate (lncRNA-hits) and lncRNA-not-hits, we built a tree-based classifier based on high-throughput CRISPRi functional screen data in seven human cell lines, as well as, cell-specific ENCODE transcription factor ChIP-seq data; finding that the genomic features used in our study showed small effects and tend to be transcript-specific. Our classifier outperformed previous algorithms, displayed balanced sensitivity and specificity values, and uncovered a lncRNA (*LINC00879*) involved in cell-growth. Additionally, we unveiled a list of 40 lncRNAs as candidates for experimental validation. Second, we characterized the lncRNA profile during regeneration, using *Drosophila* wing imaginal disc as a regeneration-model. We selected a candidate lncRNA (*CR40469*) and evaluated its role in regeneration at the early stage of cell-damage. Subsequently, using RNA-seq data, we observed significant transcriptomic alterations in consequence of the *CR40469* genetic deletion, suggesting its role in regeneration. In this study we have generated a list of lncRNAs whose possible biological role in cell-growth and in regeneration can be further studied.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

bp - basepair
cDNA - complementary DNA
CRISPR - Clustered Regularly Interspaced Short Palindromic Repeats
CRISPRi - CRISPR interference
Ctrl - Control
DE - Differentially expressed
DEG - Differentially expressed genes
DNA - Deoxyribonucleic Acid
Down - Downregulated
ENCODE - ENCyclopedia Of DNA Elements
FC - Fold change
FPKM - Fragments per kilobase million
GEO - Gene expression omnibus
GTEx - Genotype-Tissue Expression
Gtf - Gene transfer format
H3K4me1 - monomethylation of histone H3 at lysine 4
H3K4me2 - dimethylation of histone H3 at lysine 4
H3K4me3 - trimethylation of histone H3 at lysine 4
H3K9ac - acetylation of histone H3 at lysine 9
H3K9me3 - trimethylation of histone H3 at lysine 9
H3K27ac - acetylation of histone H3 at lysine 27
H3K27me3 - trimethylation of histone H3 at lysine 27
H3K36me3 - trimethylation of histone H3 at lysine 36
H3K56ac - acetylation of histone H3 at lysine 56
KO - KnockOut
lincRNA - long intervening (sometimes intergenic) noncoding RNA
lncRNA - long noncoding RNA
ML - Machine Learning
modENCODE - model organisms ENCODE
mRNA - messenger RNA (protein-coding)
ncRNA - noncoding RNA
NDE - Not differentially expressed

nt - nucleotide
PCG - Protein coding gene
PCR - Polymerase Chain Reaction
Reg - Regeneration
RFE - Recursive Feature Elimination
RNA - Ribonucleic Acid
TPM - Transcripts per kilobase million
TSS - Transcription Start Site
Up - Upregulated
UTR - Untranslated Region
Wt - Wild-type

Introduction

I

The noncoding genome

One of the distinguishing hallmarks of eukaryotic genomes is their large size and low protein-coding content. Less than 2% of the human genome consists of protein-coding genes.[1] The question then arises as to the composition and function (if any) of the remaining genome.

Much of the noncoding regions of the human genome have historically been called "*junk DNA*". Transcriptome genome-wide analyses over the past 18 years demonstrated that regions between protein-coding genes are frequently transcribed into RNA molecules of diverse lengths.[2–5] The various types of non-protein-coding loci can be classified according to its length into: **1)** short (< 200 nucleotides) and **2)** long noncoding RNAs (> 200 nucleotides).

1. **Short noncoding RNAs**: carry out relative well-defined functions in cells, and are already accepted as fundamental players in gene regulation;[6,7] these include: microRNAs (miRNAs), small nucleolar RNAs (snoRNAs), Piwi-interacting RNAs (piRNAs), small nuclear RNAs (snRNAs), tRNAs, and rRNAs. Conversely, short noncoding RNAs represent a tiny fraction of the human, mouse, and fruit fly genomes (see Table 1). Usually short noncoding RNAs are recognized by 3D conformations by various proteins forming ribonucleoprotein complexes.[6,8]

2. **Long noncoding RNAs**: are the most common class of noncoding RNAs. Long noncoding RNAs (lncRNAs) are defined as RNAs longer than 200 nucleotides with no apparent coding potential. This poor definition encompasses a large and heterogeneous class of transcripts that differ in their biogenesis and genomic location, this poor definition comes from our limited understanding of lncRNAs. The majority of lncRNAs are transcribed by RNA polymerase II (Pol II) and often capped by 7-methyl guanosine (m^7G) at their 5′ ends, polyadenylated at their 3′ ends, and spliced similarly to protein-coding genes (PCGs).[9,10] It is worthwhile highlighting that enhancer regions are also transcribed into

enhancer RNAs (eRNAs).[9,11]

Organism	Gene number	Genomic coverage (Kb)	Genome sequence covered	Annotation
Human	8,130	783	0.027%	GENCODE[12]
Mouse	6,656	568	0.031%	GENCODE[12]
Fruit fly	1,019	161	0.134%	FlyBase[13]

Table 1: Short noncoding RNAs in the human, mouse and fruit fly genomes. Statistics are based on the following short noncoding RNAs: miRNAs, rRNAs, snoRNAs, snRNAs, and tRNAs.

LncRNAs in contrast with short noncoding RNAs, are highly abundant and except for a few lncRNAs their function remains elusive; even with the constant efforts by reference annotations of coding and noncoding genes, including GENCODE[12] or FlyBase[13] projects. Over the previous decades, the lncRNA literature has dramatically changed, from studying one single-lncRNA-locus to genome-wide analyses; perturbing several thousands of lncRNAs or their regulatory sequences with the aim to observe a phenotype and linking lncRNAs with a molecular function. This dramatic change was mainly ignited after the culmination of the human, mouse and fruit fly genome projects. Surprisingly, results from large genomic consortiums such as the Encyclopedia of DNA Elements (ENCODE) consortium have unveiled that most of the human genome is actively transcribed, whether it encodes a protein or not.[14,15] After ∼200 experiments conducted in humans by the ENCODE consortium estimated that ∼80% of the human genome is actively transcribed. Among these transcripts, ∼1%-2% mapped to protein-coding exons, whereas the rest mapped either to noncoding genes or protein-coding introns (where genic intronic lncRNAs are transcribed).[1,14,15] Similar results were obtained by the Functional Annotation of the Mammalian Genome (FANTOM) consortium.[16]

These results fomented a deeper study of lncRNAs in diverse model organisms, developmental stages, tissues, and human conditions. In the next section, we are going to study the infancy of lncRNA biology, from *H19* locus (the first uncovered lncRNA) to nowadays with the aim to give us a framework for future discoveries and perspectives.

I.1. LncRNA history: pre and post-genomic era

I.1.1. Early lncRNA discoveries

In the late 1980s, the first discovered eukaryotic lncRNA, *H19*, was characterized in the pre-genomic era, even though at that time *H19* was classified as a PCG[8] (Figure 1). LncRNA *H19* is a spliced, ∼2.3 Kb long transcript, with high sequence conservation across mammals, and localized in the cytosol. *H19* is involved in the control of cell-growth during early mammal embryonic development.[8] However, the function of *H19* as a lncRNA remained a mystery until the functional characterization of the second discovered eukaryotic lncRNA, *X-inactive specific transcript* (*Xist*).

LncRNA *Xist* shortly discovered after *H19* (Figure 1), is involved in chromosome X inactivation in female mammals. In mammals, dosage compensation of X-linked genes between females (XX) and males (XY) is achieved through X-chromosome inactivation (XCI), from which *Xist* is the master regulator.[17] LncRNA *Xist* is upregulated in one of the two X chromosomes in females at early embryonic stages, and its RNA spreads *in cis* along the entire X chromosome.

Xist recruits the Polycomb repressive complex 2 (PRC2) triggering the inactivation of the X chromosome.[8] Interestingly, *Xist* is a very long lncRNA (∼17 Kb) with six domains (A-F), and sometimes classified as macro and/or very-long lncRNA.[8]

The lncRNA relevance is not restricted to mammalian genomes, lncRNAs: *roX1* and *roX2* have a key role in fruit fly dosage compensation, and are another case of lncRNA functionality before the arrival of the genomic era (Figure 1). In *D. melanogaster*, dosage compensation involves the upregulation of X-linked genes in males to match the gene expression from the two X chromosomes in females.[18]

The male-specific lethal (MSL) ribonucleoprotein complex, composed of five MSL proteins and the lncRNAs *roX1* and *roX2*, is involved in the upregulation of genes located in the X chromosome of *Drosophila* males.[18] The MSL subunits coat the male X chromosome and bring about histone acetylation (H4K16ac), resulting in increased male transcription.[19] Remarkably, *roX1* and *roX2* report differences in size and sequence, but act redundantly to allow the binding of MSL2 and other subunits to target the male X chromosome.[20]

I.1.2. The dawn of the genomic era

First cDNA sequencing efforts uncovered thousands of newly discovered lncRNAs in the human, mouse and fruit fly genomes.[21–23] Remarkably in the early 2000s, the FANTOM consortium pioneered the genome-wide discovery of lncRNAs, publishing a set of 34,030 lncRNAs in the mouse genome.[22] Despite this explosion in the number of newly discovered lncRNAs, only a handful had been clearly characterized.

Previous studies were based on deep transcriptome sequencing, nonetheless, in 2009 Guttman *et al.* used chromatin signatures to identify and validate ∼1,600 and 100 long intervening RNAs (lincRNAs), respectively across four mouse cell types; with many lincRNAs bearing signs of purifying selection.[3] The team realized that genes transcribed by Pol II are marked by H3K4me3 at their promoters and H3K36me3 at the transcript end, then the so-called "*K4-K36 domain*" was used to identify lincRNAs genome-wide.

A relevant discovery regarding the noncoding genome was made in 2010; when it was shown that enhancers are actively transcribed.[24,25] The product of this transcription is termed eRNA, and its role has been the source of great debate and speculation. The role of most eRNAs has remained enigmatic, leading to suggest that enhancer transcription is the "*noisy byproduct*" of the transcriptional machinery. Nevertheless, a growing number of studies suggest diverse roles for eRNAs, including promotion of enhancer-promoter interactions, and gene regulation.[11,26]

In 2012, Djebali *et al.*, and Derrien *et al.* results pinpointed the well-known lncRNA features including lncRNAs exhibit standard canonical splice site signals and alternative splicing, lncRNA loci are under weak selective constraints –in human lncRNAs many are primate-specific– lncRNA TSS histone profiles are similar to those of PCGs for several active histone marks (H3K4me2, H3K4me3, H3K9ac, H3K27ac) and report slightly excess of silencing histone marks (H3K27me3, H3K36me3), lncRNA display lower and tissue-specific expression relative to PCGs, and lncRNAs are enriched in the nucleus.[1,4]

In 2017, Lagarde *et al.* developed the RNA Capture Long Seq (CLS), which combines targeted RNA capture with short-read (Illumina) and long-read (PacBio) sequencing.[27] CLS method tackles lncRNAs low expression and low read coverage by capture-oligos designed to tile lncRNA loci. This work is notable for producing full-length transcript models enabling us to characterize lncRNA genomic features, including promoter, gene structure and protein-coding-potential. Nevertheless, CLS method relies on PacBio technology due to its high price limits its application to most

labs and other genomes. Moreover, CLS is tailored to uncover lincRNAs leaving overlapping lncRNAs aside.

Nowadays, although tens of thousands of new lncRNAs have been identified by different catalogs such as GENCODE,[12] NONCODE,[28] RefSeq,[29] MiTranscriptome,[30] and FANTOM-CAT[16] in different genomes, except for a handful of genes, the function of most lncRNAs remain elusive. In consequence, it is paramount to study and characterize lncRNA functions in different cell-specific contexts, using deep transcriptome sequencing to unveil new lncRNA loci, and functionally validate them searching for phenotypes after creating targeted mutations in candidate genes.

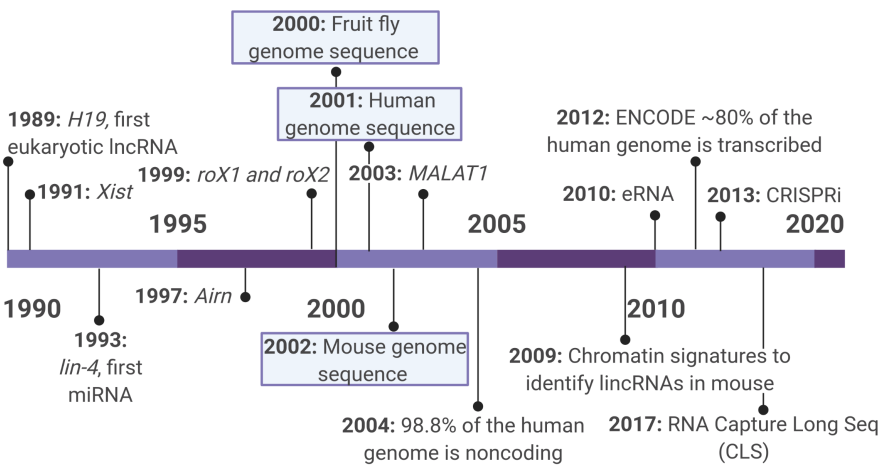

Figure 1: LncRNA discoveries timeline. Main discoveries in noncoding RNAs, in particular lncRNAs.

I.2. Long noncoding RNAs: a building block of biological processes

Based on lncRNAs genomic positions relative to neighboring PCGs, lncRNAs can be classified as intergenic, genic exonic, or genic intronic if lncRNA loci come from an intergenic region, overlaps a protein-coding exon, or intron,[31] respectively. LncRNAs are highly abundant in many organisms,[12,13] such as humans (17,948 genes and 48,741 transcripts), mice (13,186 genes and 18,833 transcripts), and fruit flies (2,545 genes and 3,047 transcripts; see Figure 2), but other lncRNA annotations such as NONCODE[28] estimates 96,411, 87,890, and 15,543 lncRNA genes for human, mouse, and fruit fly, respectively.

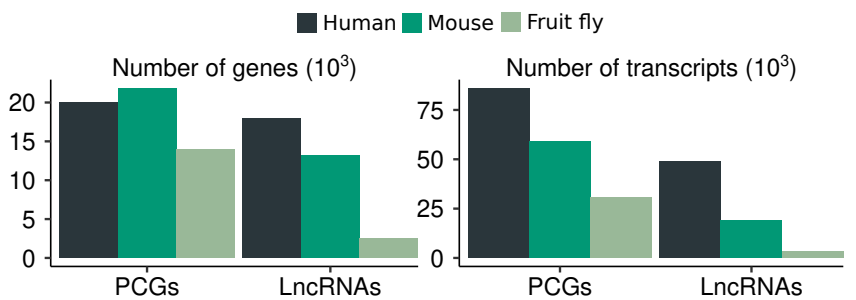

Figure 2: Statistics in the human, mouse and fruit fly genomes. Shown are the gene (left) and transcript (right) numbers for the PCG (left) and lncRNA (right) gene types. Inspired by.[4]

Mouse number of lncRNAs is mildly different from the human genome, however, it is unclear how much of this difference is biologically related rather than by the more mature status of the human genome annotation (Figure 2). For fruit fly *Drosophila melanogaster* (*D. melanogaster*) differences in the number of lncRNAs can be explained for the smaller *Drosophila* genome with approximately 120 megabases, compared to the human and mouse genomes with 3,100 and 2,700 megabases,[32] respectively. Moreover, fewer differences for PCGs are observed among the human, mouse and fruit fly genomes, prompting the notion that PCGs are better annotated and conserved.

There are at least three factors that make lncRNAs challenging to study. First, lncRNAs are poorly expressed compared to PCGs, meaning that lncRNA transcripts are underrepresented in any transcriptomic analysis, such as RNA sequencing (RNA-seq), expressed sequence tags (EST), tiling microarrays, and cap analysis of gene

expression (CAGE) data.[4,33] Second, lncRNAs show tissue-specific and condition-specific expression patterns, making it challenging to compare to other expression datasets.[1,10] Third, lncRNAs tend to have little primary sequence conservation, meaning that ortholog and paralog analyses are challenging to implement.[12,28] See Table 2 for further lncRNA and PCG comparisons.

Feature	LncRNA	PCG	Reference
H3K4me1	Low	Low	8
H3K4me3	High	High	8
H3K36me3	Moderate/high	High	8
H3K27ac	High	Low	8
Subcellular location	Nucleus Cytosol Mitochondria Other organelles, *e.g.* exosomes	Cytosol	9
Transcript length	Human: 714 bp median Mouse: 1087 bp median Fruit fly: 646 bp median	Human: ∼2.7 Kb median Mouse: ∼2.5 Kb median Fruit fly: ∼1.6 Kb median	12 12 13
RNA stability	Variable, overall lower than PCG Highly unstable: eRNA	Variable	8

Table 2: Comparison of lncRNA and PCG features. Only the longest transcript for each gene was considered; using the following gene annotations: the GENCODE Human, GENCODE Mouse, and/or Flybase reference annotations, versions: 37, M26, and r6.29, respectively (release: 2021).

In contrast with PCGs and short noncoding RNAs, the vast majority of lncRNAs functions remain enigmatic. LncRNA function has been subject of controversy, with few hundreds (or $\leq$ 1%) of experimentally validated or disease-associated lncRNAs.[33] Suggesting that lncRNA mere existence or production does not automatically imply functionality. Nevertheless, it is well documented that a growing number of lncRNAs are associated with relevant biological processes.[9,10,34] Additionally, lncRNAs are predominantly localized in the nucleus and several lncRNAs control the expression of nearby genes (*cis-acting lncRNAs*) by affecting their transcription and chromatin features. Other several lncRNAs function away from their loci (*trans-acting lncRNAs*); their functions can be structural, involvement in signaling pathways, and regulation of PCGs including splicing and translation.

Consequently, lncRNAs interact with several paramount cellular functions that

are of great importance, and alteration of their expression is inherent to numerous diseases such as neuronal disorders, hematopoiesis and immune response, cancer, etc. Thus, lncRNAs constitute a major gene class and unraveling their function will constitute a better understanding of our genome.

I.2.1. LncRNA conservation

Comparative analyses of genes across species can be a powerful tool for understanding their functions and action modes. For instance, the miRNA *let-7* is conserved from humans to nematodes.[35] Comparative analyses require two main inputs: sets of genes or genomes that can be compared, and bioinformatic tools for evaluating the conservation. Applying comparative analyses to lncRNAs is challenging for two main reasons. First, only a few lncRNAs had been annotated in species other than human, mouse, and fruit fly. Second, lncRNAs lack long conserved sequences or regions with strong conserved structures, which are important features for conservation algorithms. Consequently, lncRNA loci from various species can be compared in the three following levels:

1. **Primary sequence conservation:** The first approach is to apply whole-genome multiple alignments (*e.g.* those available in the *UCSC* genome browser) or directly align the query lncRNA with lncRNA databases from other species using BLAST/BLAT or other alignment tools. Sequence conservation results demonstrated that lncRNA exons are less conserved than PCG exons.[36,37] Interestingly, lncRNA exons on average are more conserved than PCG introns and random intergenic sequences.[38,39] However, there are two main drawbacks of using multiple alignments, as shown in Table 2 the length of lncRNAs are shorter than PCGs, and the violation of the key assumption that lncRNA exons in one species align to lncRNA exons in the other species, in many cases lncRNA loci are homologous to non-conserved sequences in the other species.[36,40]

2. **Structure conservation:** when comparing lncRNAs across more-distant species, sequence conservation might not be the best approach. An open debate is whether secondary structure plays an important role in lncRNA biology, as it does in short noncoding RNAs[6] (such as miRNA, tRNA, etc.). Two observations support the secondary structure importance, the rapid rate of lncRNA evolution and lncRNA ability to fold into secondary structures, many of which are stable, nevertheless forming secondary structures *per se* does not imply function. A successful usage of structure conservation is the detection of

distant homologs on lncRNAs *roX1* and *roX2* in *Drosophila* species[41] (Table 3). Quinn *et al.* identified 43 new *roX* orthologs in diverse *Drosophila* species across ∼40 million years of evolution distance despite limited sequence similarity. In Pegueroles *et al.* study in 4 nematode species, a higher number of lncRNA orthologs were identified using secondary structures.[42] Unfortunately, the currently available secondary structure predicting tools are not accurate enough for long sequences as lncRNAs,[43] thus prediction should be considered with caution. Additionally, there is no correlation between the amount of secondary structure and overall sequence conservation.[44,45]

3. **Positional conservation:** it has been proposed that in some cases, lncRNA function acts through transcription *per se* instead of transcript displaying a function for itself.[7,9,10] For instance in mice, one of the functions of the lncRNA *Airn* (a genic-intronic lncRNA overlapping the PCG *Igf2r*) can be explained for its position and not *Airn* transcript itself, repressing *Igf2r* by both transcription interference and DNA methylation.[7,46,47] In such lncRNAs, we would expect that the position of the region that is transcribed would be conserved, whereas the exon positions would evolve neutrally. The lncRNA *PVT1* can serve as an example of transcribed region conservation, *PVT1* shows deep positional conservation (Table 3) but the transcript length and exon-intron architecture evolved rapidly.[48] The *LincOFinder* pipeline added its own worth by uncovering 16 homologous lncRNAs between very evolutionary distant species, humans and amphioxus by position conservation.[49] Although, positional conservation has shown promising results unveiling homologous lncRNAs, its approach is only applied to lincRNAs, leaving aside overlapping lncRNAs. Moreover, positional conservation deeply relies on orthologous PCGs discarding lincRNAs within low coding-gene content.

Given the different levels of lncRNA conservation based on probability of conserved functionality, proximity to PCGs, overlap with transposable elements, tissue specificity, and expression levels; it has been proposed three levels of lncRNA classification: class I, class II, and class III.[50] In class I, lncRNA exon-intron structure and multiple sequences along the lncRNA locus are conserved across species, lncRNAs: *MALAT1*, *NEAT1*, and *NORAD* can be classified as class I (Table 3). In class II, lncRNAs are those in which the act of transcription and some RNA elements are conserved, whereas the majority of lncRNA locus experienced drastic changes in exon-intron structure and length, *lnc-ONECUT1* (*LINC02490*) can serve as a class II example (Table 3). In class III, lncRNAs show promoter sequence conservation and the act of transcription on the specific region, for example, the lncRNA *FENDRR* display

promoter conservation.

LncRNA	Conservation level	Mechanism
roX1-roX2[51]	Conserved in *Drosophila* species	Fruit fly dosage compensation
PVT1[48]	Deep positional conservation	Function as an oncogene in different cancers
MALAT1[9]	Multiple conserved sequence and e-i* structure	Involved in structural functions
NEAT1[9,50]	Multiple conserved sequence and e-i* structure	A scaffold lncRNA of paraspeckles
NORAD[50]	Multiple conserved sequence and e-i* structure	Promotes PCG stability for genome integrity
lnc-ONECUT1[50]	Transcription and some elements are conserved	NA
FENDRR[50]	Transcription and promoters are conserved	Is an essential regulator of heart

Table 3: Example of conserved lncRNAs. e-i*= exon-intron structure.

I.2.2. Small Open Reading Frames (smORFs) within lncRNA genes

By definition, lncRNAs lack coding potential. Surprisingly, 98% of annotated lncRNAs contain at least one small Open Reading Frame (smORF) in the human, mouse and fruit fly genomes with a median of six smORFs per lncRNA.[52] In consequence, Couso *et al.* results challenge the current definition of lncRNAs.

smORFs contain 10 to 100 codons, and millions of smORF sequences are found in eukaryotic genomes.[52,53] The putative function of these peptides is, however, often neglected and the genes that encode them remain listed as noncoding. The *tal* gene can be used as an example, which was previously annotated as noncoding, *tal* gene encodes 4 small peptides of 11 amino acids.[52] Nevertheless, examples of small-functional-peptides have been described functioning as regulators of membrane-associated proteins, or as components of ancient protein complexes.[54]

There are six smORF classes based on their RNA type, median codon size, translation rate, coding features, and function. smORFs within lncRNAs represent the third most abundant class of smORFs with a low translation efficiency.[52] These results highlight our poor lncRNA definition, and the need for more classification parameters in addition to length cutoff.

II

LncRNA roles and mechanisms of action

LncRNA roles and mechanisms of action, to this day, still struggle to keep pace with the ever-growing lncRNA catalogs: of the thousands of currently discovered lncRNA loci, less than 500 have robustly assigned cellular function.[55] Functional lncRNAs can be classified as "*cis-acting lncRNAs*", when they influence the expression, splicing and/or chromatin state of nearby genes, or "*trans-acting lncRNAs*", which act far from their locus.[10] Based on our current understanding, functional lncRNAs can influence gene expression at three main levels: **1)** chromatin regulation, **2)** transcriptional regulation, and **3)** post-transcriptional regulation.[9,34]

II.1. Chromatin regulation

Two famous lncRNAs *Xist* and *Airn* involved in chromatin regulation were discovered before the Human Genome Project (see Figure 1). *Xist* involved in mammalian dosage compensation, and *Airn* antisense to the imprinted *Igf2r* gene.[17,47]

Airn was uncovered by Wutz *et al.* as the first lncRNA in regulating the imprinted expression of neighboring PCGs.[56] *Airn* is an intronic antisense lncRNA overlapping the PCG *Igf2r*. Additionally, *Airn* functions as *trans-acting lncRNA* placing on the promoters of two distal imprinted target genes, *Slc22a2* and *Slc22a3*. Once there *Airn* recruits PRC2, which catalyzes H3K27me3 leading to gene silencing in mouse stem cells.[46]

As these two early unveiled lncRNAs were implicated in chromatin regulation, their discovery raised expectations that chromatin regulation might be a common feature of lncRNAs. Since then, several lncRNAs have been associated in displaying direct interaction with chromatin *in cis* and *in trans*, in the recruitment of chromatin modifiers, and acting as a decoy of chromatin modifiers. (See Table 4 to have a summarized view of lncRNAs involved in chromatin regulation).

LncRNA	Interacting with	Mechanism	Sequence features
Xist[17]	PRC2,YY1,hnRNPK,etc.	Silences X-linked genes	Long range interaction
Airn[46]	PRC2	Silences *Slc22a2* and *Slc22a3* genes	NA
TARID[57]	*GADD45A*	Forms R-loops and recruits *GADD45A*	Interacts with GC-rich seq.
ANRIL[58]	PRC1 and PRC2	Regulate distal genes *in trans*	*Alu* retroelements motifs
HOTTIP[59]	*WDR5-MLL*	Activates HOXA genes	NA
lncPRESS1[60]	*SIRT6*	Functions as *SIRT6* decoy	NA
APOLO[61]	*LHP1*	Functions as *LHP1* decoy	Two TTCTTC boxes

Table 4: LncRNAs involved in chromatin regulation

II.1.1. Direct interaction with chromatin

Dueva *et al.* conclusions that the negative charge of RNA can neutralize the positively charged histone tails and numerous lncRNAs localized in the chromatin where lncRNAs can interact with proteins, suggest a rapid switch of gene expression.[62] The well-studied lncRNAs *Xist* and *Airn* can serve as examples of lncRNAs with direct interaction with chromatin acting *in cis* and *in trans*, respectively (see Early lncRNA discoveries for further *Xist* mechanistic details).[17,46]

Moreover, lncRNAs can form RNA-DNA hybrids such as R-loops, by interacting with DNA. The lncRNA *TARID* mechanism of action is explained through R-loops with *GADD45A* locus, which drives the methylation of the *TCF21* promoter and consequently silences *TCF21* expression.[57] Holdt *et al.* work reported that the lncRNA *ANRIL* interacts with chromatin as a *trans-acting lncRNA* through *Alu* motifs, which drives *ANRIL* recruitment of PCR1 and PCR2 to distal genes leading to increased cell proliferation, increased cell adhesion and decreased apoptosis.[58]

II.1.2. Recruitment of chromatin modifiers

LncRNAs can interact with chromatin modifiers and recruit them to target PCG regulatory elements to activate or inactivate their locus expression *in cis*, or *in trans*. The lncRNA *HOTTIP* is one of the several lncRNAs that regulate the *HOXA* gene cluster, *HOTTIP* is localized upstream of the *HOXA* cluster, and *HOTTIP* expression contributes to the maintenance of chromatin organization in *HOXA* region.[63]

HOTTIP recruits the mixed-lineage leukemia (MLL; also known as KMT2A) complex, which is a chromatin modifier, to activate the expression of the *HOXA* genes

through H3K4me3 chromatin mark and playing as a notable regulator of mouse hematopoietic stem cells.[59]

II.1.3. Acting as a decoy of chromatin modifiers

In addition to interacting with chromatin and recruitment of chromatin modifiers, lncRNAs may function as decoys of chromatin modifiers by sequestering them from the DNA regulatory regions of target genes. For example, the lncRNA *lncPRESS1* acts as a decoy of *SIRT6* chromatin modifier.[60]

LncPRESS1 supports the pluripotency of human embryonic stem cells by sequestering *SIRT6* from the promoters of numerous pluripotency genes by maintaining active H3K56ac and H3K9ac chromatin marks. During p53-mediated differentiation or *lncPRESS1* depletion, *SIRT6* localizes to the chromatin and inhibits the expression of pluripotency genes.[60]

The *APOLO* gene is another lncRNA that functions as a decoy of chromatin modifiers.[61] In *Arabidopsis thaliana* (*A. thaliana*), *APOLO* acts as a decoy of *LHP1* during auxin response. Normally, *APOLO* and auxin target genes are silenced by H3K27me3 and the presence of the Polycomb factor-like heterochromatin 1 (*LHP1*). Then, in response to auxin *APOLO* is expressed and acts *in trans* to target its target gene promoters forming R-loops and acting as a decoy of *LHP1*, thereby allowing target-gene expression.[64]

II.2. Transcriptional regulation

The non-random genomic arrangement of lncRNAs throughout genomes could represent a key determinant for lncRNAs to regulate PCGs transcription. Moreover, Seila *et al.* reported antisense and bidirectional lncRNA transcription to be evolutionarily conserved, this could represent an evolutionary adaptation of genes to regulating their own transcription in a context-specific manner.[65]

Under Luo *et al.* results, we analyzed lincRNAs with a locus-locus distance from their closest neighboring PCG lower of 5 Kb for the human and mouse genomes, and 1 Kb for the fruit fly genome. Our observations are in agreement with Luo *et al.* study, where divergent lincRNAs are the most common lincRNA class in the human, mouse and fruitlfy genomes[66] (Figure 3). In addition, we observed fewer differences between the divergent lincRNA class and the rest of the lincRNA classes

within the fruit fly genome; this could explained by lower levels of bidirectionality in *D. melanogaster*.[67] Consequently, these non-random genomic arrangements of divergent lincRNAs suggest lncRNAs play a pivotal role in regulating nearby PCGs transcription.

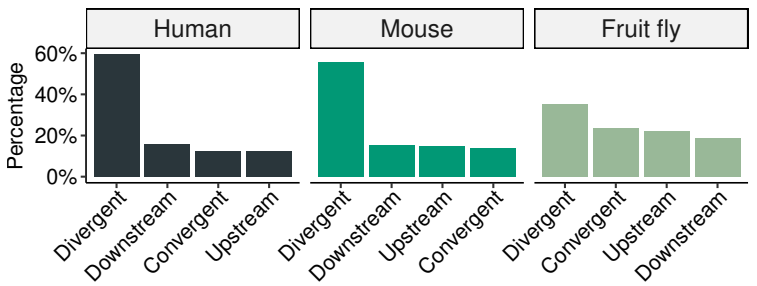

Figure 3: LincRNA classification for the human, mouse and fruit fly genomes.. Percentage of lincRNA classification for genes with a distance < 5 Kb between lincRNA locus and the closest neighboring PCG for human and mouse, and < 1 Kb for fruit fly. Inspired by.[66]

LncRNA regulates PCG transcription by two main mechanisms, and non-mutually exclusive **1)** transcript-dependent: the lncRNA transcript for itself can regulate PCG loci (*in cis* or *in trans*), or **2)** transcript-independent: the act of transcription of the lncRNA can generate a steric impediment or chromatin state that influence the expression of nearby genes.

II.2.1. Transcript-dependent regulation

The *cis-acting* lncRNA *ANRASSF1* is an antisense genic-exonic of the PCG *RASSF1*, which is a tumor suppressor gene in different cancers. The lncRNA *ANRASSF1* can serve as an example of transcript-dependent regulation, *ANRASSF1* is transcribed from the opposite strand of the *RASSF1* locus and is responsible for recruiting PRC2 to the *RASSF1* promoter region, leading to the H3K27me3 repressive marks. *ANRASSF1* transcript has a function for itself, forming an RNA/DNA hybrid and recruiting PRC2 to the *RASSF1* promoter.[9]

In *A. thaliana*, the *cis-acting* lncRNA *COOLAIR* is an antisense genic-exonic of the *FLC* locus, which is a regulator of the transition to reproduction. *COOLAIR* transcript is cold-induced and is involved in the epigenetic silencing of the PCG *FLC* through changed H3K36me3/H3K27me3 dynamics. Cold strongly upregulates *COOLAIR* transcript, which lingers at its site of transcription and coats the locus to promote

PRC2-dependent H3K27me3 leading to *FLC* silencing.[9]

As a *trans-acting* lncRNA with transcript functionality, we can highlight the lncRNA *HOTAIR*. *HOTAIR* is an antisense genic-intronic lncRNA of the *HOXC* locus. *HOX* transcription factors (TFs) encoded from four *HOX* gene clusters (*HOXA*, *HOXB*, *HOXC*, and *HOXD*) are deeply conserved and involved in positional identity and differentiation.[9,68] *HOTAIR* is required to maintain repressive chromatin marks at the distant *HOXD* locus through interactions between *HOTAIR* and components of PRC2.[7,10] Depletion of *HOTAIR* with small interfering RNAs (siRNAs) resulted in transcriptional activation of *HOXD* genes with an associated decrease in the repressive chromatin mark H3K27me3.[9]

II.2.2. Transcript-independent regulation

LncRNAs can suppress gene expression by interfering with the transcription machinery, which leads to alteration of the recruitment of TFs or Pol II at the inhibited promoter, alteration of histone modifications, and reduction of chromatin accessibility. Several transcript-independent mechanisms have been proposed (reviewed in[7,9,10]), but in this thesis work we will analyze three main mechanisms: **1)** RNA polymerase collision, **2)** regulatory elements embedded within lncRNA loci, and **3)** the increasing roles of eRNAs.

II.2.2.1. RNA polymerase collision

LncRNA transcription can regulate neighboring PCG expression after transcriptional initiation by transcriptional interference that occurs co-transcriptionally. This mechanism can be mediated by direct RNA polymerase collision by "*sitting-duck*" interference[1] or by one RNA polymerase acting as a "*roadblock*" for other incoming elongating polymerase.[7] If a gene is simultaneously transcribed in both directions, this leads to RNA polymerase collision (Figure 4A).

Nonetheless, *in vitro* phage polymerases that act in both directions are able to bypass each other; this is not the case for more complex bacterial or eukaryotic RNA polymerases.[69] Additionally, transcriptional interference by direct polymerase collision is most likely when two strong convergent genes are present, conversely, it is unlikely for two weak convergent genes.[7]

In *D. melanogaster*, the lncRNA *bsAS* (FlyBaseID=*CR44811*) regulates its PCG by

[1] When an elongating polymerase removes another that is already attached to a gene promoter.

polymerase collision. *bsAS* is an antisense genic-intronic of the PCG *bs*, which is involved in wing development and formation.[70] *bsAS* is involved in the regulation of *bs* isoform usage in flies in a tissue-specific manner, by the transcription of *bsAS*.[71] Expression of *bsAS* occurs specifically in wing intervein regions and impairs the transcription of the *bs* long isoforms, thus promoting the expression of the short isoform. Pérez-Lluch *et al.* proposed the RNA polymerase collision mechanism (Figure 4A) to explain the inhibition of the *bs* long isoform.

Furthermore, the lncRNAs *Airn*[2] and *Chaserr* mechanisms of action are explained by polymerase collision. *Chaserr* is a conserved lncRNA and is located upstream of the *Chd2* gene, which is a chromatin remodeler implicated in neurological disorders.[72]

II.2.2.2. Regulatory elements embedded within lncRNA loci

As described above, functional DNA elements within lncRNA loci can activate the expression of neighboring genes (Figure 4B). The lncRNA *Bendr* regulates *in cis* its neighboring gene, *BEND4*, through the presence of enhancer elements in its locus. The enhancer element is activated by *Bendr* transcription.[9,73]

The lincRNA *p21* provides another instructive example of regulatory elements within lncRNAs. *p21* is a nuclear-localized transcript that neighbors the *CDKN1A* gene in humans and mice. Genetic analyses of the lincRNA *p21* uncovered that its locus contains *cis-regulatory* DNA elements that modulate *CDKN1A* expression.[9] Other lncRNAs have been reported with similar roles in the activation of proximal enhancers,[73] such as the lncRNA *Uph*.[74]

II.2.2.3. eRNAs

Active enhancers can be transcribed into two main types of noncoding RNAs: eRNA and enhancer-associated lncRNAs (elncRNAs).[26] The main distinction between eRNAs and elncRNAs is their genomic features. elncRNAs are mostly unidirectional, polyadenylated, spliced, longer (up to 4 Kb) and transcribed from higher-activity enhancers. By contrast, eRNAs are bidirectional capped transcripts, non-polyadenylated, unspliced, shorter (< 2 Kb), unstable and transcribed from H3K4me1 marked enhancers.[9,26,75,76] Moreover, the general features of eRNA and elncRNA are highly conserved from humans to flies.[75]

[2] See LncRNA conservation for a detailed *Airn cis-acting* mechanism.

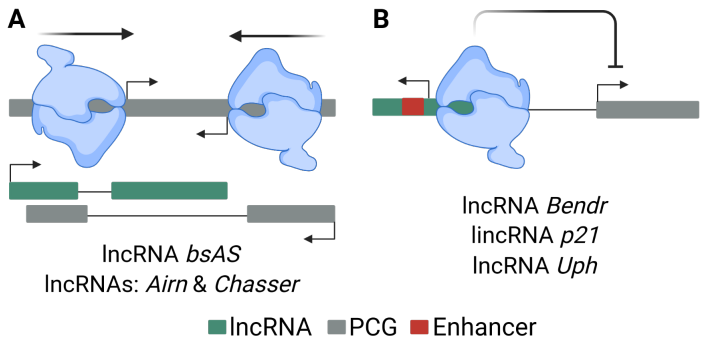

Figure 4: Transcript-independent mechanisms. **(A)** RNA polymerase collision. **(B)** Regulatory elements embedded within lncRNA loci.

The literature overall supports a model wherein eRNAs contribute to enhancer action by interacting with nuclear proteins to promote enhancer-promoter looping, and gene regulation.[9,26] For instance, the lncRNA *eNRIP* is transcribed into an eRNA, which recruits cohesin to form enhancer-promoter looping. Thus, promoting contact between *NRIP1* and *TFF1* promoters leads to loci expression of these genes, this mechanism is regulated by estrogen receptor activation.[9]

II.3. Post-transcriptional regulation

In addition to their roles in chromatin and transcriptional regulation, lncRNAs can act through their ability to establish interactions with proteins and nucleic acids regulating PCGs post-transcriptionally.[10,77] Here, we highlight a few of the many different modes lncRNA functions as post-transcriptional regulators, mainly focusing on: **1)** lncRNAs as a source of miRNAs and **2)** lncRNAs regulating PCG splicing.

II.3.1. LncRNAs as a source of miRNAs

miRNAs are short noncoding RNAs (∼22 nucleotides), which play a relevant role in the post-transcriptional regulation of gene expression.[78] In many cases, miRNAs are derived from the introns or exons of larger genes ("*host*"). If the miRNA is processed from the host exonic sequence, the processing reaction typically leads to rapid exonucleolytic degradation of the host. By contrast, if the miRNA is processed from the host intronic sequence the host RNA stability is typically not affected.[6]

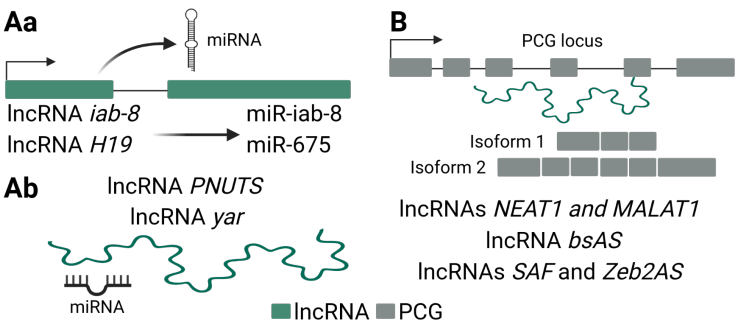

Figure 5: Post-transcriptional regulation of lncRNAs. **(Aa)** LncRNAs as a source of miRNAs. **(Ab)** LncRNAs acting as miRNA "*sponges*". **(B)** LncRNAs regulating isoform usage.

In *D. melanogaster*, the lncRNA *iab-8* acts as a source of miRNAs. Once transcribed, lncRNA *iab-8* is processed into three miRNAs transcripts that are collectively called miR-iab-8, these miRNAs are processed from lncRNA *iab-8* intronic sequence. These miRNAs are known to target and downregulate the homeotic genes *abd-A* and *Ubx*, as well as their cofactors *hth* and *exd*.[79] Knocking down lncRNA *iab-8* expression results in male and female sterility.[80]

In mammals, several lncRNAs have been described as precursors of miRNAs. A well-studied case is the maternally-imprinted *H19* locus. During skeletal muscle differentiation and regeneration in mice, the lncRNA *H19* is processed into the miRNA miR-675, which is embedded in *H19* first intron. This miRNA functions by directly downregulating the *Smad* TF.[81] In parallel, *H19* is highly present in fetal tissues, where it is found to be processed into miR-675, which limits placental growth by targeting, among others, the PCG *Igf1r*.[82]

II.3.1.1. LncRNAs acting as "*sponge*" of miRNAs

Some lncRNAs contain miRNA complementary sites that can regulate gene expression as competitive endogenous RNAs or "*sponges*" of miRNAs, thereby reducing miRNA availability to target PCGs.[83,84]

For instance, the lncRNA *PNUTS* serves as a miRNA sponge of the microRNA miR-205. In tumors, the pre-mRNA *PNUTS* generates the lncRNA *PNUTS* through alternative splicing, the lncRNA locus contains seven miR-205 binding sites, decreasing the availability of miR-205 to bind and suppress its target genes (*ZEB1* and *ZEB2*).[85] In *D. melanogaster*, the lncRNA *yar* contains ∼33 miRNA binding sites, its

cytoplasmic location, and its incapacity to affect transcription of neighboring genes suggest *yar* may function as a miRNA sponge[86]. Although, the exact mechanism remains enigmatic.

II.3.2. LncRNAs regulating pre-mRNA splicing

Recently, certain lncRNAs have been shown to play a crucial role in regulating pre-mRNA alternative splicing (AS) in response to several stimuli or diseases.[77,87] The main mechanisms involving lncRNAs in AS modulation can be classified in two ways: **1)** lncRNAs interacting with splicing factors (SFs), and **2)** lncRNAs forming RNA-RNA duplexes with pre-mRNA molecules.

II.3.2.1. LncRNAs interacting with splicing factors

Using genome-wide screenings, the intergenic lncRNAs *NEAT1* and *MALAT1* (or *NEAT2*) were among the first lncRNA loci implicated to interact with SFs in mouse and human cells.[77] *NEAT1*[9,77] is localized in paraspeckles[3], whereas *MALAT1*[9,77] is part of the polyadenylated component of nuclear speckles[4]. More recently, more lncRNAs were reported to modulate PCG AS (*e.g. SAF*, *GOMAFU*, and *LINC01133*).

Serine-arginine-rich (SR) proteins are part of a conserved protein family involved in splicing.[77] SR proteins are commonly localized in the nucleus (although several of them are known to shuttle between the nucleus and the cytoplasm) and their function in splicing is linked to its phosphorylation status.[77]

During adipocyte differentiation, *NEAT1* modulates the AS profile of the *PPARγ* pre-mRNA into *PPARγ-1* or *PPARγ-2* isoforms. *NEAT1* modulates *SRp40* phosphorylation status by interacting with the *Clk* kinase.[88] Phosphorylated *SRp40* promotes the processing of the *PPARγ* pre-mRNA into the *PPARγ-2* isoform. By contrast, the dephosphorylation of *SRp40* promotes the *PPARγ-1* isoform expression.[89] *PPARγ* encodes for the major TF implicated in adipocyte differentiation, in consequence *NEAT1* modulation plays a relevant for cell viability and function. Additionally, *NEAT1* depletion causes a decrease of *PPARγ-1* and *PPARγ-2* isoforms, in particular *PPARγ-2* isoform.

The lncRNA *MALAT1* acts as an oncogene and its abnormal transcription is implicated in the development and progression of many cancers.[90,91] Results in human

[3]Paraspeckles: nuclear domains that control sequestration of related proteins.
[4]Nuclear speckles: nuclear domains enriched in pre-mRNA splicing factors.

cells demonstrate that *MALAT1* regulates splicing by modulating SR splicing factors distribution and phosphorylation dynamics[92]. Depletion of *MALAT1* enhances the dephosphorylated pool of SR proteins resulting in the mislocalization of speckle components and changes in AS of pre-mRNAs. The control of the levels of phosphorylated SR proteins impacts not only AS but also other SR post-transcriptional mechanisms, including RNA export, translation and nonsense-mediated decay.[77] The exact *MALAT1* mechanism by which *MALAT1* depletion alters the ratios of phosphorylated to dephosphorylated SR proteins in the cell remains elusive. However, it is possible *MALAT1* regulates the action of the *SRPK1* kinase and the *PP1/2* phosphatase, which modify SR proteins.[77]

II.3.2.2. LncRNAs forming RNA-RNA duplexes with pre-mRNA molecules

Overlapping lncRNAs in antisense represent 31.8%, 27.1% and 33.7% of the human, mouse, and fruit fly genomes[12,13], respectively (overlapping in the coding and in the noncoding regions). Krystal *et al.* detected RNA-RNA duplexes *in vivo*, when the team was studying the oncogene *N-myc* and its overlapping gene in antisense.[93] Thus, it was postulated that RNA-RNA duplexes can modulate pre-mRNA splicing.

In *D. melanogaster*, the lncRNA *bsAS* controls its PCG isoform usage[71] (see RNA polymerase collision for *bsAS* mechanism). In mammals, the lncRNA *SAF* is linked with apoptosis and cancer through the interaction between the *FAS* receptor and its ligand. In human cell lines, *SAF* is transcribed from the first intron of the *FAS* locus. *SAF* interacts with the exon 6 of the *FAS* pre-mRNA, forming RNA-RNA duplexes. *SAF* recruits the splicing factor *SPF45* facilitating AS and exclusion of exon 6. The exclusion of exon 6 from the *FAS* pre-mRNA leads to producing soluble *FAS*, which lacks the transmembrane domain rendering cell less sensitive to *FAS*-mediated apoptosis.[94,95]

Epithelial-mesenchymal transition (EMT) can be highlighted as another biological context where lncRNAs regulate PCG isoform usage through RNA-RNA duplexes. The genic-exonic *Zeb2AS* overlaps in antisense with the *Zeb2* locus. After EMT, the *Snail* TF induces the transcription of the lncRNA *Zeb2AS* in epithelial cells. A specific RNA-RNA duplex around the 5′ splice site of the 5′ UTR intron prevents the binding of the spliceosome.[96] Thus, favoring *Zeb2* translation. In absence of *Zeb2AS* transcription, the resulting mRNA contains a stable secondary structure before the first codon, which is able to block *Zeb2* translation.[77,96]

II.4. Conservation of lncRNA functions

The percentage of lncRNA conservation is increasingly regarded as a key feature in evaluating the impact of a studied lncRNA. If a lncRNA is involved in a human illness, it is relevant to know whether it can be studied in a model organism. Conversely, if a lncRNA is uncovered in a model organism, evidence of conservation is important to establishing relevance to human biology.

One paramount riddle is –if conserved lncRNAs also function in similar mechanisms in other species?– Several studies have found that lncRNA tissue specificity as well as specific expression patterns, are generally highly conserved.[48,50] Thus, conserved lncRNA could act in similar contexts in different species. For instance, the lncRNA *CARMEN* is required for cardiomyogenesis for both human and mouse cells,[50] *XIST* is required for X inactivation in humans and mice,[50] and the lncRNA *NEAT1* causes loss of paraspeckles across species (Table 5).

LncRNA	Conservation level	Mechanism
CARMEN[50]	Conserved lof* phenotype in mouse and human	Required for cardiomyogenesis
XIST[17,50]	Conserved across mammals	X inactivation in female mammals
NEAT1[9,50]	Multiple conserved sequence and e-i* structure	A scaffold lncRNA of paraspeckles

Table 5: LncRNAs with conserved functions. e-i*= exon-intron structure; lof*= loss-of-function.

III

High-throughput screens to uncover functional lncRNAs

After conducting a genome-wide transcriptome study comparing two biological conditions, we obtain a list of differentially expressed genes – among them lncRNAs. However, this approach explains little or nothing about lncRNA biology and its mechanisms of action. Several reverse-genetics assays[97–101] have been successfully used to uncover lncRNA functions, searching for phenotypes after creating targeted mutations (*e.g.* knockout and knockdown experiments) in candidate loci.

For PCGs a single insertion or deletion can abolish the PCG functionality. In contrast, for lncRNAs this approach does not apply due to our limited understanding of lncRNA functional domains. Consequently, other strategies are used including full-length deletion of lncRNA locus or deletion of lncRNA promoter regions.[102,103] These constraints condition the loss-of-function approaches implemented in lncRNAs. Moreover, it is advisable to minimize the removal of DNA regions for lncRNA functional analyses.

Reverse-genetics methods can be broadly classified according to their targets, for instance acting directly at the lncRNA locus level (*e.g.* DNA cleavage or local recruitment of silencing histone marks at the lncRNA TSS) or at the lncRNA transcript level (*e.g.* RNA knockdown through RNAi).[103,104] Acting at the RNA-level represents the most direct method for assessing lncRNA functionality without confounding factors caused by disruptions of DNA regulatory elements. RNA interference (RNAi) and antisense oligos (AOs) represent the most implemented methods for studying lncRNAs acting at the transcript level, with more than 1,500 studies unveiling lncRNA functionality in diverse cellular contexts[103] (see Table 6). RNAi and AOs knockdown their target lncRNAs through RISC and RNase H mediated mechanisms, respectively.[103,104] The lncRNAs *Neat1*, *SPRY4-IT1*, *DGCR5*, and other lncRNAs have been reported to show phenotypic consequences using RNAi and AOs methodologies.[105–107]

Nonetheless, RNAi and AOs methods present important disadvantages including the inability of genome-wide screens.[104] Additionally, Stojic *et al.* work demonstrated considerable off-target defects and sequence-dependent nature applying RNAi and AOs technologies within the HeLa cell line transcriptome.[108] Moreover, RNAi incapacity to knockdown nuclear lncRNAs is a well-known drawback, hampering the analysis of a large fraction of lncRNAs.[103,104,108] Finally, these hurdles have paved the way for the usage of CRISPR-related systems.

III.1. CRISPRi: genome-wide lncRNA screening

The bacterial Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-Cas9 nuclease system is a highly adaptable technique, and has been used in many genome-wide editing studies.[109,110] Briefly, the CRISPR-Cas9 system works through the guiding of the Cas9 protein to a target sequence through a single guide RNA (sgRNA), the sgRNA directs the enzyme to bind DNA, where the nuclease induces a DNA double-strand break (DSB). Upon cleavage, the DNA repair machinery is recruited to the DSB, often inducing point mutations or frameshift mutations at the target locus to functionally knockout the PCG.[109,110]

CRISPR has been further modified for modulating locus expression without modifying the genomic sequence through the use of a nuclease-dead Cas9 (dCas9), which binds the target site without cleaving the DNA.[111] The CRISPR-dCas9 has been adapted for both gene inhibition (CRISPRi[112]) and activation (CRISPRa[113,114]). These inhibition and activation CRISPR-systems have been successfully applied in high-throughput screens in many different cells to improve the understanding and characterization of lncRNAs.[97,98,101] Further, the newly discovered Cas13 enzyme, which binds and modifies the RNA rather than the DNA, shows potential for high-throughput lncRNA analysis at the transcriptional level[115] (see Table 6).

CRISPRi is based on the use of dCas9 protein fused to the Krüppel-associated box (KRAB) transcriptional repression domain.[112] The CRISPRi system inhibits transcription in part through the dCas9 ability to sterically hinder RNA polymerase binding, and in part through the KRAB domain to place the repressive histone mark H3K9me3 at its target TSS[104,112,116] (see Figure 6). Gilbert *et al.* demonstrated that the use of dCas9 and KRAB domain improved the knockdown of gene targets significantly compared to dCas9 alone.[112] In addition, the repression effect of CRISPRi is transient, and the effect is diminished until elimination at six to fourteen days after transfection.[117] As shown in Figure 6, the CRISPRi system deeply relies on the cor-

rect lncRNA TSS annotation, which in many times is not either complete or accurate leading to diminished CRISPRi effectiveness.

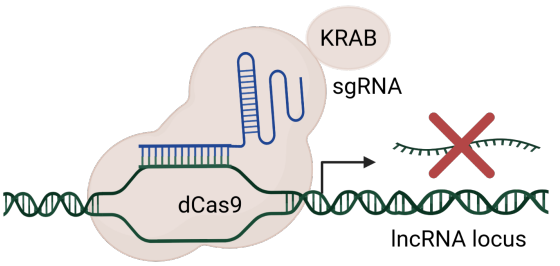

Figure 6: CRISPRi repression mechanism. Blue and green ribbons denote sgRNA and lncRNA locus, respectively.

Recently, additional repressive domains have been added to the dCas9-KRAB cassette to further improve the knockdown capabilities of the CRISPRi system. These additional repressive domains were selected by screening multiple domains from DNA-binding proteins. Notably, SIN3-interacting domain (SID), ZIM3, and methyl CpG binding protein 2 (MeCP2) were reported to improve the efficiency of repression by CRISPRi.[118,119] Moreover, for achieving long-term repression effects, the domains DNMT3A, DNMT3L, and Tet1 have been fused to dCas9, which are specifically designed to alter the DNA methylation states.[120]

The CRISPRi system presents lower off-target defects compared to RNAi and AOs.[103,104,108] Further, CRISPRi shows decreased sequence-dependent off-target effects suggesting CRISPRi-mediated loci inhibition is highly specific, and comparisons between cells treated with different sgRNAs can be safely performed.[108]

Technique	Target	Outcome	Mechanism	Limitation
RNAi[104,108]	RNA	Knockdown	RISC	Inefficient for nuclear lncRNAs
AOs[103,108]	RNA	Knockdown	RNase H	Elevated off-targets
CRISPRi[112]	DNA	Knockdown	dCas9 & KRAB domain	Requires accurate lncRNA TSS
CRISPRa[104]	DNA	E.t.	dCas9 & VPR domain	Cannot discern *in cis* and *trans*
CRISPR-Cas13[104]	RNA	Knockdown	2 HEPN endoRNase domains	Limitated sgRNA portals

Table 6: Techniques to explore lncRNA functions. E.t.= Enhanced transcription.

Nonetheless, CRISPRi-mediated inhibition is far from perfect. It has been reported that chromatin accessibility has a major impact on the success of CRISPRi. Importantly, although CRISPRi was designed to create a barrier leading to the col-

lapse of the RNA polymerase complex in a local and transient manner,[112] in some cases CRISPRi may also lead to changes in methylation states and hence to the silencing of neighboring genes.[121] This may be particularly relevant for overlapping and divergent lncRNAs, which are found within or in close proximity to functional PCGs.

III.2. Cases of use of CRISPRi

In recent years, CRISPRi platforms have been adopted for functional screenings of regulatory elements (*e.g.* enhancers and promoters) and noncoding RNAs.[97–101] For large-scale screening of perturbations, Liu *et al.* developed sgRNA pooled libraries; their pooled library targeted the TSS of 16,401 lncRNAs, with ten sgRNAs per TSS.[97] For the generation of this comprehensive library, the authors used three gene catalogs (Ensembl, MiTranscriptome, and Rinn/Broad) and obtained their expression values from seven human cell lines. Using their libraries, the authors then screened for lncRNAs affecting fitness; they found nearly 500 different lncRNAs significantly affecting cell-growth. An important finding of this pioneering screening was that most functional lncRNAs displayed a cell type-specific effect, while similar experiments targeting PCGs displayed that between one-third and half of the identified essential genes are shared between multiple cell types.[97]

More recently, Haswell *et al.* generated a pooled sgRNA CRISPRi library targeting 12,611 lncRNA transcripts expressed in human embryonic stem cells (hESCs), using 10 sgRNAs per transcript.[98] The authors screened for genes affecting hESC differentiation; they identified sixty functional lncRNAs, of which several were functionally validated. Notably, among the twenty-three positive PCG controls in the library, only six were identified as positive hits.[98] This finding emphasizes that CRISPRi remains limited in terms of sensitivity, suggesting that the number of functional lncRNAs may be significantly greater than what is currently reported.

IV

The role of lncRNAs in regeneration

LncRNAs are implicated in diverse biological contexts including development, neuronal disorders, immune response, cancer, etc. However, in this section we are going to focus on studies of lncRNAs that play a function in regeneration and their mechanism of action, across distinct model organisms and regeneration types (Table 7).

IV.1. Regeneration

Regeneration is the replacement of single-cells, tissues or body parts in homeostasis or following trauma, and regeneration capacity can vary widely among species, tissues, and life stages. Regeneration encompasses both the cellular self-renewal of a particular tissue throughout the organism's life ("*tissue homeostasis*" or "*physiological regeneration*"), and the restoration of injured tissues or lost body parts ("*reparative regeneration*").[122,123]

In mammals, an example of physiological regeneration is the cellular replacement of endometrium, epidermis, gut lining, and red blood cells. Cellular self-renewal in adult organs involves stem cell differentiation or transdifferentiation of existing cells.[124] Conversely, reparative regeneration can be either incomplete, with only partial restoration of structure and function, or complete. Incomplete regeneration includes regeneration of digital tips of fetal and juvenile mice, and fingertips of children – a process involving blastema[5] formation.[125,126] Complete regeneration includes the axolotl ability to regenerate their limbs. Axolotl amputation stimulates the formation of blastema from remaining cells, which is similar to a limb stump. Next, blastema cells grow and are patterned into mature skeletal elements.[127]

In the early 1900's, Morgan coined the terms "*epimorphosis*" and "*morphallaxis*" to refer to regenerative phenomena in which cellular proliferation takes part, and

[5]Blastema: a mass of proliferative cells that form after amputation (*e.g.* in salamander limb stump), and ultimately gives rise to new structures.

to refer to re-patterning of existing tissue (with limited cellular proliferation), respectively.[128] Currently, reparative regeneration can be classified as follows:

1. **Blastema-mediated epimorphic regeneration:** repair occurs via blastema formation. Wound-healing after an extreme injury such as limb regeneration in urodele amphibians (*e.g.* salamanders), full-thickness skin recovery in mice, or tissue regeneration after physical fragmentation in *Drosophila* imaginal discs can be classified as blastema-mediate epimorphic regeneration.[123,129]

2. **Epimorphic regeneration:** recovery takes place from a precursor-independent process that requires direct recruitment and cellular proliferation of differentiated cells, this repair is observed in hepatocytes, and in zebrafish hearts.[130,131]

3. **Morphallaxis regeneration:** is observed in invertebrates and occurs through the re-patterning of existing tissues. *Hydra* is one example where morphallaxis takes place.[122]

Additionally, another classification has been proposed for regeneration based on the multiple levels of biological organization, ranging from cells to tissues, organs, structures, and whole-body regeneration.[132]

Across the animal kingdom, there is a remarkable diversity of regeneration capacity, not only from one species to another, but also between tissues and organs or between developmental/life stages of the same species. For instance, whereas planarians can regenerate their whole-body from tiny fragments, certain Platyhelminthes cannot regenerate their heads after amputation.[122] Similarly, the capacity for skin regeneration has evolved differently between the mouse lab model (*Mus musculus*) and the African spiny mouse (*Acomys*). While the African spiny mouse can regenerate the entire dermis, as well as the underlying connective tissues, the mouse lab is unable to regenerate and instead forms fibrotic scars.[122,133] Notably, the same is observed in heart regeneration in teleost species, where the heart regeneration is not common to all species. Although hearts in other cyprinids such as the goldfish (*Carassius auratus*) and the giant danio (*Devario aequipinnatus*) regenerate successfully, those in medaka (*Oryzias latipies*) scar instead.[133]

In mammals, including humans, some tissues have elevated regenerative capacity throughout life, such as blood cells, intestinal epithelium, liver, skeletal muscle and skin up to a certain threshold of damage or loss. In contrast, several organs including the brain, spinal cord, heart and joints possess minimal regeneration capacity.[127] These deviations highlight the great diversity of regeneration between tissues and organs in the same species.

Moreover, regeneration also depends on the developmental stage or age of the individual. For example, aging negatively affects regenerative capacity as a result of cellular senescence[6], telomere shortening, impaired cell differentiation, and increased metabolic stress.[123] In addition, aging impairs peripheral nerve regeneration in mammals, and in all vertebrates regeneration capacity is increased in younger animals. In mammals, fetuses and newborns have a relatively higher regeneration potential, which is lost in adulthood.[122] The same negative correlation between age and regeneration is observed in *Drosophila* imaginal discs and adult male zebrafish, which are unable to regenerate their pectoral fins due to the localized growth of breeding ornaments.[122,129,133]

IV.2. *Drosophila* imaginal discs: a model to study regeneration

Many model organisms are used in the study of tissue regeneration, but in this thesis work we are going to focus on regeneration studies in *Drosophila* imaginal discs. Additionally, we are going to discuss other model systems to study regeneration:

1. **Planarians:** certain planarians can regenerate their whole-body from a tissue fragment, through stem cells termed "*neoblasts*". This is a robust model for interrogating stem cell involvement in regeneration. Although, stable transgenesis for planarians has been challenging to develop,[127] however its arrival will enable us to study gain-of-function in living cells.

2. ***Hydra*:** also exhibit whole-body regeneration like certain planarians,[122] nonetheless with less tissue-complexity and more rudimentary transgenesis techniques.[127]

3. **Salamanders:** possess a high regeneration potential. Newts and axolotls have the remarkable ability to regenerate limbs. Genome data are now available for certain salamander species, which could facilitate the study of genome-wide salamander regeneration capacities. Further, axolotls have the shortest generation times (∼12 months) and are amenable to transgenesis and gene-editing techniques.[127]

[6]Cellular senescence: is a process in which cells cease dividing and undergo distinctive phenotypic alterations.

4. **Zebrafish:** one of the most studied models for regenerative biology. Several mutant strains have been identified and multiple genetic tools, originally pioneered in *Drosophila* and in mice, have been successfully adapted in zebrafish. Zebrafish have a remarkable capacity to regenerate different organs, including all seven fins and scales, as well as tissues with therapeutic relevance, such as brain, heart, kidney, liver, pancreatic β-cells, retinae, and the spinal cord.[134] The main drawback of the zebrafish model is its elevated generation time of $\sim$3 months.[130]

5. **Mice:** although its limited regeneration potential, mouse models have been essential to understand hepatocyte and satellite cells (muscle-specific stem cells) function in liver and skeletal muscle regeneration, respectively.[131,135,136] For instance, the rodent partial hepatectomy (PHx) model, where two thirds of the rodent liver are removed surgically, has been one of the most significant sources of liver regeneration knowledge.[131] Moreover, *Mus musculus* researchers have a plethora of genetic tools at their disposal, such as loss-of-function and gain-of-function techniques, genome editing (*e.g.* CRISPR), and well-characterized phenotypes.

The fruit fly (*D. melanogaster*) along with the zebrafish (*Danio rerio*), the frog (*Xenopus laevis*) and the mouse (*Mus musculus*) has been instrumental in providing fundamental insights not only in tissue regeneration but into a wide variety of biological processes. *Drosophila* as a model organism provides major features for probing the function and regulation of genes during development, regeneration, physiological, and pathological processes. Such relevant features include a life cycle well-studied at the gene and cellular level, tissues with regenerative capacity (*e.g.* imaginal discs), complex and well-characterized morphology, abundant gene-editing tools, well-documented genomic sequence, lower genome complexity compared to vertebrates, and RNA-seq data from different biological contexts and tissues.

More importantly, the major biological processes are highly conserved between fruit flies and humans. In fact, $\sim$77% of known human disease genes have homologs in the fruit fly genome.[137] However, no fly homologs have been uncovered for lncRNAs involved in human diseases.

Imaginal discs are epithelial sacs with two cellular layers (*columnar epithelium* and *squamous peripodial epithelium*) that are the primordia of adult appendages and other cuticular structures. Imaginal discs are capable of regenerating after damage, and thus can serve as a model to study regeneration.[122,129,138] Damage can be induced physically (physical fragmentation or X-ray irradiation) or genetically, by genetic in-

duction of cell-death.[129]

Genetic ablation takes advantage of the *Gal4/UAS* system to target a pro-apoptotic gene (*e.g. egr*, *rpr*, *debcl*, *hid*) to a defined region of the imaginal disc and the temperature-sensitive version of *Gal80* (*Gal-80^{ts}*) to restrict the ablation to a specific time frame across normal imaginal disc development[129,138,139] (see Figure 7A). Inducing cell-death genetically offers three advantages over physical damage. First, since the disc is ablated *in situ*, adult structures can be generated from imaginal discs, offering the extent of studying regeneration in living organisms. Second, specifically induce cell-death in the *spalt major* (*salm*) domain of the wing pouch (Figure 7B). Third, genetic ablation is far less laborious. Interestingly, discrepancies are shown in the response to ablation with different pro-apoptotic genes.[129] These variances may reflect different signaling pathways triggered by each pro-apoptotic gene.

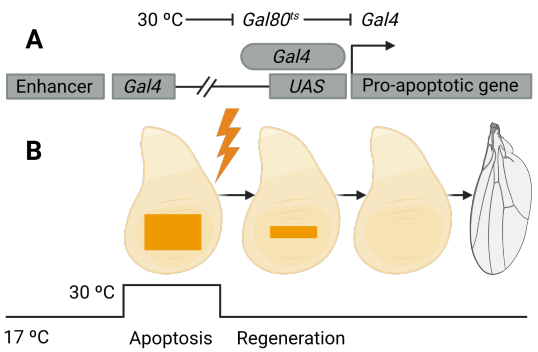

Figure 7: Regeneration in *Drosophila* wing imaginal disc. **(A)** Induction of cell-death using the *Gal4/UAS* system, 30ºC inhibits *Gal-80^{ts}* permitting pro-apoptotic gene expression. **(B)** Regeneration progress of wing imaginal disc, cell-death induction occurs on the wing pouch. Inspired by.[129]

IV.3. LncRNAs involved in regeneration

Several recent studies have described roles of chromatin structures, DNA regulatory elements (enhancers and promoters), transcription factors, PCGs, and signaling pathways in regeneration.[122,138–140] Nonetheless, lncRNAs function and their mechanisms of action in regeneration remain poorly explored, mostly limited to performing transcriptome analyses for PCGs and leaving lncRNAs as an appendix. Some lncRNAs have been unveiled to act in more than one regenerative type. For instance, *H19*

and *MALAT1* are involved in skeletal muscle and liver regeneration,[131,136] whereas *Sirt1AS* is implicated in muscle and cardiac regeneration.[136,141]

Following injury, skeletal muscle can regenerate from muscle-specific stem cells, termed satellite cells (SCs), which proliferate and differentiate into myotubes.[135] *H19* is highly expressed in SCs, and *H19* targeted deletion leads to 50% loss of SCs in adult mice.[142] The mechanism is unknown but it may be linked with the *Igf2-Igf1r* signaling pathway.[136] Moreover, the pro-proliferative *Cdc6* and *Smad* genes are repressed by two miRNAs produced from the first intron of *H19*.[143] In liver regeneration, *H19* is upregulated and contributes to the increased expression of the *CcnD1* gene and DNA synthesis, leading to hepatocyte proliferation.[131]

MALAT1 is upregulated after 2 hours of liver wound-healing and acts as a regulatory factor in the cell cycle. *MALAT1* participates in the activation of the *Wnt/β-catenin* signaling pathway by inhibiting the *Axin1* and *APC* loci.[144] Additionally, *MALAT1* participates in muscle differentiation, acting as a sponge for miR-133; preventing miR-133 from inhibiting its target PCG, such as *SRF*. In consequence, *SRF* is expressed and able to promote terminal differentiation of the muscle progenitor cells.[145]

The lncRNA *Sirt1AS* is transcribed from the antisense strand of the PCG *Sirt1*, which is a NAD-dependent class III protein deacetylase. *Sirt1AS* interacts with the 3'UTR of *Sirt1* forming a RNA-RNA duplex to protect *Sirt1* transcript from degradation mediated by miR-34a.[136,141] Thus, *Sirt1* stability and pro-proliferation ability are augmented. During muscle regeneration, *Sirt1AS* transcription sustains muscle-progenitor-cell proliferation by increasing the expression of cyclins B, D, and E.[146,147] Moreover, loss-or-function results in mice suggest *Sirt1AS* is required and sufficient to induce cardiomyocyte proliferation (mechanism needed for heart regeneration). Additional results in cardiac regeneration demonstrated that *Sirt1AS* overexpression enhances survival rate, improves cardiac function, and inhibits fibrosis after myocardial infarction.[141]

Additional lncRNAs have been uncovered to function in different regeneration types including muscle, cardiac, liver, and nerve regeneration in diverse model organisms (see Table 7). Acting in a wide-range of mechanisms, for instance acting as a source of miRNAs (*H19*), acting as a sponge of miRNAs (*MALAT1*, *NR-045363*, *LUCAT1*, *CAREL*), encoding functional peptides from smORFs (*LINC00961*), promoting chromatin loops (^{ce}*eRNA*), forming RNA-RNA duplexes (*Sirt1AS*), inhibiting or activating the expression *in cis* or *in trans* of neighboring PCGs (*Dum*, *SRA*, *CPR*, *lncPHx2*, *lncHand2*, *Silc1*), and activating signaling pathways (*ECRAR*, *lncDACH1*, *LALR1*). See[127,131,135,136,141] for further details.

LncRNA	Regeneration type	Mechanism
H19[143]	Skeletal muscle and liver regeneration	Acts as a source of miRNAs
MALAT1[144]	Skeletal muscle and liver regeneration	Acts as a sponge for miR-133
Sirt1AS[141]	Skeletal muscle and cardiac regeneration	Inhibits *Sirt1* degradation
Dum[136]	Skeletal muscle regeneration	Inhibits *Dppa2* expression
^{ce}eRNA[136]	Skeletal muscle regeneration	Increases *MyoD* expression
SRA[136]	Skeletal muscle regeneration	Co-activator of *MyoD*
LINC00961[136]	Skeletal muscle regeneration	Contains a smORF that encodes for *SPAR*
ECRAR[141]	Cardiac regeneration	Promotes cardiomyocytes to re-enter cell-cycle
NR-045363[141]	Cardiac regeneration	Acts as a sponge for miR-216
LUCAT1[141]	Cardiac regeneration	Acts as a sponge for miR-612
lncDACH1[141]	Cardiac regeneration	Bounds to *PP1A* subunit
CPR[141]	Cardiac regeneration	Inhibits *MCM3* expression
CAREL[148]	Cardiac regeneration	Acts as a sponge for miR-296
LALR1[131]	Liver regeneration	Activates the *Wnt/β-catenin* pathway
lncPHx2[131]	Liver regeneration	Activates *E2F1* and histone proteins expression
lncHand2[131]	Liver regeneration	Upregulates *c-Met* expression
Silc1[127]	Nerve regeneration	Upregulates *Sox11* expression

Table 7: Mechanisms of action of lncRNAs involved in regeneration

To the best of our knowledge, most of the work performed about tissue regeneration in *Drosophila* imaginal discs has been focused mainly on the chromatin, transcription factor, signaling pathways, and PCGs level.[122,138–140] And little work has been performed in the literature to characterize the role and mechanisms of action of lncRNAs in fruit fly discs during regeneration.

Objectives

The main objective of the present Thesis Project is to unravel the role of lncRNAs in two biological scenarios. The first, cell-growth in seven human cell lines (**Chapter I**: **??**). The second, after genetically inducing cell-death in *Drosophila* wing imaginal discs (**Chapter II**: **??**). Hence, the objectives of this Thesis Project are (see Figure 8 for a general overview of this thesis work):

1. To harness the richness of ever-growing public available genomic datasets by using nonlinear models, such as tree-based machine learning models, to generate a classifier to unveil functional lncRNAs in the context of cell-growth in human cell lines.

2. To understand the role of lncRNAs during regeneration, using *Drosophila melanogaster* wing imaginal disc as a regeneration-model, to generate a list of lncRNA candidates to perform experimental validations, and unveil their function in the context of regeneration.

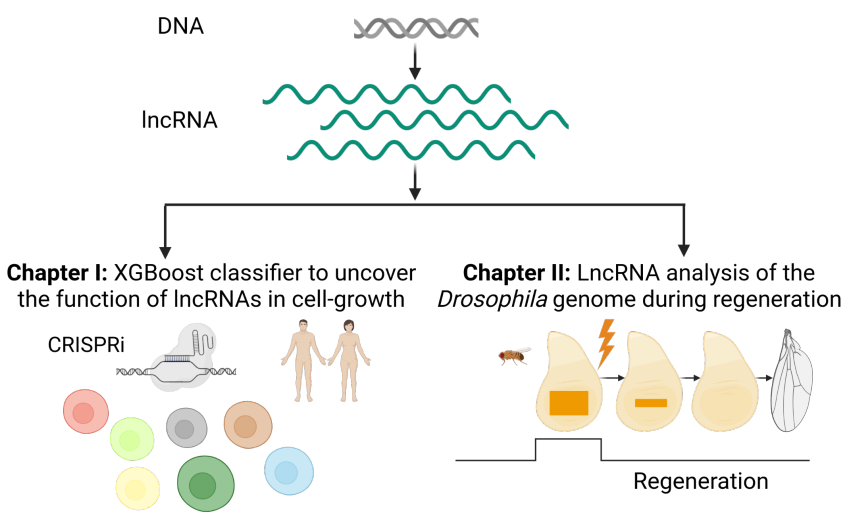

Figure 8: Thesis outline. Graphical abstract of the topics covered in this thesis work.

Materials and Methods

I

Materials

I.1. XGBoost classifier to uncover the function of lncRNAs in cell-growth

CRISPRi library was obtained from Liu *et al.*[97] work (**??**). Additionally, 124 transcription factors (TFs) were downloaded from ENCODE[**encode_2011_user**, 14] to train and test different machine learning (ML) models (**??**).

Cell line	Number of hits	Targeted loci	Cell line	Number of hits	Targeted loci
HEK293T	28	5,615	MCF7	117	5,725
HeLa	52	6,158	MDAMB231	44	5,725
iPSC	438	5,534	U87	88	5,689
K562	144	16,401			

Table 1: CRISPRi library

ARID3A	*ATF1*	*ATF2*	*ATF3*	*BACH1*	*BCLAF1*	*BHLHE40*
BRCA1	*CBX3*	*CBX8*	*CEBPB*	*CEBPZ*	*CHD1*	*CHD2*
CHD7	*CREB1*	*CTBP2*	*CTCF*	*CTCFL*	*CUX1*	*E2F1*
E2F4	*E2F6*	*EGR1*	*ELF1*	*ELK1*	*EP300*	*ESRRA*
ETS1	*EZH2*	*FOS*	*FOSL1*	*FOSL2*	*FOXA1*	*FOXM1*
GABPA	*GATA1*	*GATA2*	*GATA3*	*GTF2F1*	*HCFC1*	*HDAC1*
HDAC2	*HDAC6*	*HSF1*	*IKZF1*	*IRF1*	*JUN*	*JUND*
KDM1A	*KDM5A*	*KDM5B*	*MAFF*	*MAFK*	*MAX*	*MAZ*
MEF2A	*MTA3*	*MXI1*	*MYBL2*	*MYC*	*NANOG*	*NCOR1*
NFE2	*NFIC*	*NFYA*	*NFYB*	*NR2C2*	*NR2F2*	*NR3C1*

NRF1	*PHF8*	*PML*	*POLR2A*	*POU5F1*	*RAD21*	*RBBP5*
RCOR1	*RELA*	*REST*	*RFX5*	*RNF2*	*RXRA*	*SAP30*
SETDB1	*SIN3A*	*SIX5*	*SMARCA4*	*SMARCB1*	*SMARCC2*	*SMC3*
SP1	*SPI1*	*SREBF1*	*SREBF2*	*SRF*	*STAT5A*	*SUPT20H*
SUZ12	*TAF1*	*TAF7*	*TAL1*	*TBL1XR1*	*TBP*	*TCF12*
TCF7L2	*TEAD4*	*THAP1*	*TRIM28*	*UBTF*	*USF1*	*USF2*
YY1	*ZBTB33*	*ZBTB7A*	*ZC3H11A*	*ZKSCAN1*	*ZMIZ1*	*ZNF143*
ZNF217	*ZNF263*	*ZNF274*	*ZNF384*	*ZZZ3*		

Table 2: ENCODE TFs. 124 TFs from the ENCODE project.[encode_2011_user, 14]

I.2. LncRNA analysis of the *Drosophila* genome during regeneration

I.2.1. Characterization of cell-damage lncRNAs

Regeneration data was acquired from Vizcaya-Molina *et al.* study[138] under GEO accession number: GSE102841. **??** indicates type of genome-wide technique, organism, tissue, and condition. In this thesis work, terms 0h and early, 15h and mid, and 25h and late terms were used interchangeably.

Technique	Organism	Tissue	Condition	Reference
RNA-seq	*D. melanogaster*	Wing disc, 0h, 15h and 25h	Injured and Uninjured	[138]
H3K4me1 ChIP-seq	*D. melanogaster*	Wing disc, 0h	Injured and Uninjured	[138]
H3K27ac ChIP-seq	*D. melanogaster*	Wing disc, 0h	Injured and Uninjured	[138]
RNA Pol-II ChIP-seq	*D. melanogaster*	Wing disc, 0h	Injured and Uninjured	[138]
ATAC-seq	*D. melanogaster*	Wing disc, 0h, 15h and 25h	Injured and Uninjured	[138]

Table 3: Regeneration data

I.2.2. LncRNA developmental and tissue signatures

Developmental gene expression of *Drosophila melanogaster* (*D. melanogaster*) across embryonic, larval, white pre-pupal (WPP), and pupal stages were obtained from the modENCODE project[celniker_2009, modencode_2010] (**??**).

Condition	Time point	Organism	Technique	Reference
Development	Embryo, 0h-24h	*D. melanogaster*	RNA-seq	celniker_2009, modencode_2010
Development	L1, L2 and L3	*D. melanogaster*	RNA-seq	celniker_2009, modencode_2010
Development	WPP	*D. melanogaster*	RNA-seq	celniker_2009, modencode_2010
Development	Pupae, 12h-4days	*D. melanogaster*	RNA-seq	celniker_2009, modencode_2010

Table 4: *D. melanogaster* developmental data

In addition, leg and wing imaginal discs data was obtained from Pérez-Lluch *et al.*[71] work. Antenna and eye imaginal disc reads were obtained from the Roderic Guigó's lab at the Centre de Regulació Genòmica (CRG, Barcelona, Spain). Antenna, eyen, leg and wing imaginal disc data was produced in three *D. melanogaster* developmental time points L3, WPP and late pupae (4.5 days pupae, see **??**).

Imaginal disc	Time point	Organism	Technique	Reference
Antenna	L3, WPP and LP	*D. melanogaster*	RNA-seq	NA
Eye	L3, WPP and LP	*D. melanogaster*	RNA-seq	NA
Leg	L3, WPP and LP	*D. melanogaster*	RNA-seq	[71]
Wing	L3, WPP and LP	*D. melanogaster*	RNA-seq	[71]

Table 5: *D. melanogaster* imaginal disc data

I.2.3. Assessing the lncRNA:*CR40469* function during *D. melanogaster* imaginal-disc regeneration-process

The lncRNA *CR40469* knockout (KO) data contains the lncRNA *CR40469* knocked-out ($CR40469^{KO}$) and the lncRNA *CR40469* in wild-type ($CR40469^{Wt}$) within control and regeneration conditions both at the early time point (0h, **??**).

Genotype	Condition	Tissue	Organism	Technique	Reference
$CR40469^{Wt}$	Uninjured	Wing disc, 0h	*D. melanogaster*	RNA-seq	NA
$CR40469^{KO}$	Uninjured	Wing disc, 0h	*D. melanogaster*	RNA-seq	NA
$CR40469^{Wt}$	Injured	Wing disc, 0h	*D. melanogaster*	RNA-seq	NA
$CR40469^{KO}$	Injured	Wing disc, 0h	*D. melanogaster*	RNA-seq	NA

Table 6: *CR40469* knockout data

II

Methods

II.1. XGBoost classifier to uncover the function of lncRNAs in cell-growth

II.1.1. Data gathering and preprocessing

II.1.1.1. CRISPRi data

CRISPRi data was obtained from Liu *et al.*[97] targeting 16,401 lncRNA transcripts from seven human cell lines: iPSC, K562, U87, MCF7, MDA-MB-231, HeLa, and HEK293T; and 18 genomic features: expression level in $\log_2$(FPKM + 0.1), near FANTOM enhancer, near cancer associated SNP, number of exons, within *Pol2* loop, near super enhancer, within *CTCF* loop, near traditional enhancer, has mouse ortholog, locus is heterozygous deleted, is intergenic, transcript length, locus is amplified, is antisense, locus-nearest coding gene distance, TSS-nearest coding distance, near VISTA enhancer, and locus is homozygous deleted.

LncRNA hit is defined if inhibiting its transcriptional expression modifies cell-growth, either positively or negatively.[97] See **??** to have a general overview.

II.1.1.2. ENCODE TF ChIP-seq

We used ENCODE TF ChIP-seq data[**encode_2011_user**, 14] to determine transcription factor peak height within lncRNA promoters across five cell lines: HEK293T, HeLa, MCF7, K562 and H1-hESC, using 124 transcription factors (TFs).

We downloaded the bigBed narrowPeak files with optimal irreproducible discovery rate (IDR) thresholded peaks in hg19 assembly coordinates. We applied

a window of [-300; +100] bp upstream and downstream, respectively at the TSS to obtain lncRNA promoters, according to Dao *et al.*[dao_2017] Then using *BEDTools*[quinlan_2010_bedtools] intersect v2.27, TFs bigBed, and lncRNA promoters bed file the TF peak height was obtained. A 10% intersection cutoff between TF ChIP-seq and lncRNA promoter was used.

II.1.2. Model training

Stratified 10-fold cross-validation with 3 different randomizations in each repetition was adopted to train all supervised models, using the *RepeatedStratifiedKFold* class from *scikit-learn*[pedregosa_2011_scikit] version 0.24.1, with 90% and 10% for training and test, respectively (see **??**). Ensuring the training and the test sets to have the same hit proportion as the original dataset.

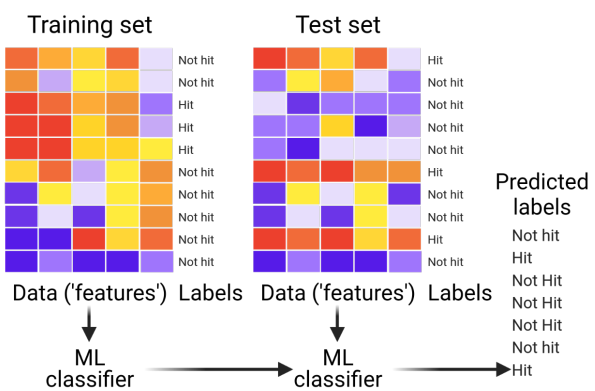

Figure 9: Process followed for model training. The functional screening based on CRISPRi and the ENCODE Transcription Factor datasets were splitted into 90% for the training set (adopting a stratified cross-validation) and 10% for the testing set; along with binary labels indicating whether the lncRNA locus is either a hit or not hit.

II.1.2.1. XGBoost

The *dmlc XGBoost* library (https://xgboost.readthedocs.io/en/latest/index.html) version 1.3.3 was used for implementing the XGBoost[chen_2016_xgboost] model. XGBoost is a type of gradient boosting decision tree method; its objective function is defined as follows:

$$L(\phi) = \sum_{n=1}^{n} loss(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where loss is the logistic regression for binary classification (*binary:logistic*), $\Omega(f_k)$is the complexity of the tree, and K is the number of trees in the model. The machine learning method, XGBoost, was tuned to search for an optimal sensitivity and specificity solution. To tune the hyper-parameters, we adopted the *GridSearchCV scikit-learn*[pedregosa_2011_scikit] class to improve the performance of the model, using a NVIDIA GPU *GeForce RTX-2060* (drivers version= 465.31, and CUDA version= 11.3). The hyper-parameters tuned for XGBoost control the growth and the robustness of the model and were the following:

- Growth: learning rate, max depth, and regularization lambda.
- Robustness: gamma.

II.1.2.2. Logistic regression

Scikit-learn[pedregosa_2011_scikit] *LogisticRegression* was implemented for the logistic regression model. *C values* and penalty hyper-parameters were tuned using *GridSearchCV*.

- *C value*: inverse of regularization strength, smaller c values means stronger regularization.
- Penalty: Lasso (l_1), and Ridge (l_2) applying square and absolute transformation on the model coefficients, respectively.

II.1.2.3. Balanced random forest

We used a random forest modification to perform data resampling on the bootstrap sample to change the class distribution. The *BalancedRandomForestClassifier* class from the *imbalanced-learn*[lemaitre_2017_imbalanced] python library, version 0.8.0, implements this and performs random under-sampling of the majority class (*i.e.* not hits) in each bootstrap sample. The balanced random forest was implemented with default parameters.

II.1.2.4. Cost-sensitive methods

As our dataset was unbalanced, the ratio of the minority positive class (hits) versus the majority negative class (not hits) was 1/55, we adopted the XGBoost *scale position weight* parameter to train a cost-sensitive XGBoost classifier for imbalanced data, 54.81 (default scale position weight) , 100, and 1000 values were used for grid search.

$$default\ scale\ postion\ weight = \frac{sum(majority\ negative\ class)}{sum(minority\ positive\ class)}$$

For the class-weight logistic regression model the inverse of the class distribution was used, by passing *balanced* as the input to the logistic regression *class_weight* parameter.

The final tuning results for cost-sensitive XGBoost and cost-sensitive logistic regression were the following:

Cost-sensitive XGBoost		
Learning rate= 0.05	Max depth= 5	Regularization lambda= 5.0
Scale position weight= 100	Gamma= 1.0	
Cost-sensitive Logistic regression		
C value= 1.0	Penalty= *l*2	Class weight =balanced

Table 7: Cost-sensitive parameters

II.1.2.5. Sampling methods

We adopted random majority under-sampling with and without replacement to resample our training and test sets, which reduced the impact of data imbalance. The python package *imbalanced-learn*[lemaitre_2017_imbalanced] version 0.8.0 was used to implement the random majority under-sampling method.

The following sampling strategy values were used: 3%, 4%, 5%, 10%, 20%, 30%, 40% and 50%.

II.1.2.6. Metrics

Further, to evaluate all model performance's, we measured the sensitivity (recall), specificity, precision, F1 score, AUROC, Brier score, and Brier skill score, using the stratified 10-fold cross-validation process described above (see **??**).

Sensitivity is the ratio of correctly predicted positive observations to all observations in a specific class, and aims to minimize the number of false negatives. It was calculated as follows in terms of the confusion matrix:

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity is the ratio of correctly predicted negative observations to all observations in a specific class, and it was obtained as:

$$Specificity = \frac{TN}{TN + FP}$$

Precision is the ratio of correctly predicted positive observations to total predicted positive observations, aims to minimize the number of false positives, and was calculated using the following equation:

$$Precision = \frac{TP}{TP + FP}$$

F1 score is the weighted average of precision and sensitivity, maximize both precision and sensitivity, and was calculated as follows:

$$F1\ score = 2 * \frac{Sensitivity * Precision}{Sensitivity + Precision}$$

Brier score is a mean square error criterion applied to binary data, and was measured as:

$$Brier\ score = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$\hat{y}_i$ is the predicted probabilities given to a set of n binary observations, and y_i taking on values 0 and 1. Brier score ranges between 0 and 1, with 0 being the score of a perfectly skilled classifier. Brier skill score is a relative metric used to compare models, a negative value means decreased performance compared to the reference. It was implemented as follows:

$$Brier\ skill\ score = 1 - (\frac{Brier\ score}{ref.\ Brier\ score})$$

II.1.3. Recursive feature elimination (RFE)

Recursive feature elimination (RFE) removing the lowest importance features, based on SHAP values[lundberg_2020, lundberg_2018] with stratified cross-validation was implemented using the *ShapRFECV* class from the *probatus* python library (https://ing-bank.github.io/probatus/index.html), version 1.8.4. The step was one feature per iteration, and using sensitivity and specificity as scoring metrics.

II.1.4. Model explainability and predictions

TreeExplainer from the SHapley Additive exPlanations[lundberg_2020, lundberg_2018] (SHAP) framework version 0.39.0 has been used to explain the output of our XGBoost model. Global and local explanations were obtained based on 10% of the data not used to train our algorithm. The SHAP framework is based on Shapley values[shapley_SHAP_values], which is a cooperative game theory concept introduced by Shapley.

To generate a list of lncRNA candidates for experimental evaluation, we used our cost-sensitive XGBoost model with 71 features to predict hit probabilities using the whole CRISPRi library.

II.1.5. Experimental evaluation

The lncRNA *LINC00879* was knocked-down using CRISPRi. Two synthetic guide RNAs (sgRNAs) were retrieved from Liu *et al.*[97] sgRNA table, and clone them into *pCRISPRia-v2.0* plasmid which includes the blue-fluorescent-protein (BFP).

For competitive growth assay, we mixed cells expressing mCherry and BFP containing the two sgRNAs targeting the lncRNA of interest or the non-targeting control at 50%. Flow cytometry was used to measure the change of BFP^+ cells fraction over 7 days. Three technical replicates were used and knockdown was validated using qPCR. (*These experiments were carried out by Joshua Hazan, from Assaf Bester's lab at Technion-Israel Institute of Technology; Haifa, Israel*). To assess differences between cells with sgRNAs and negative controls multiple paired *t-test* and *Bonferroni p-value* correction were used.

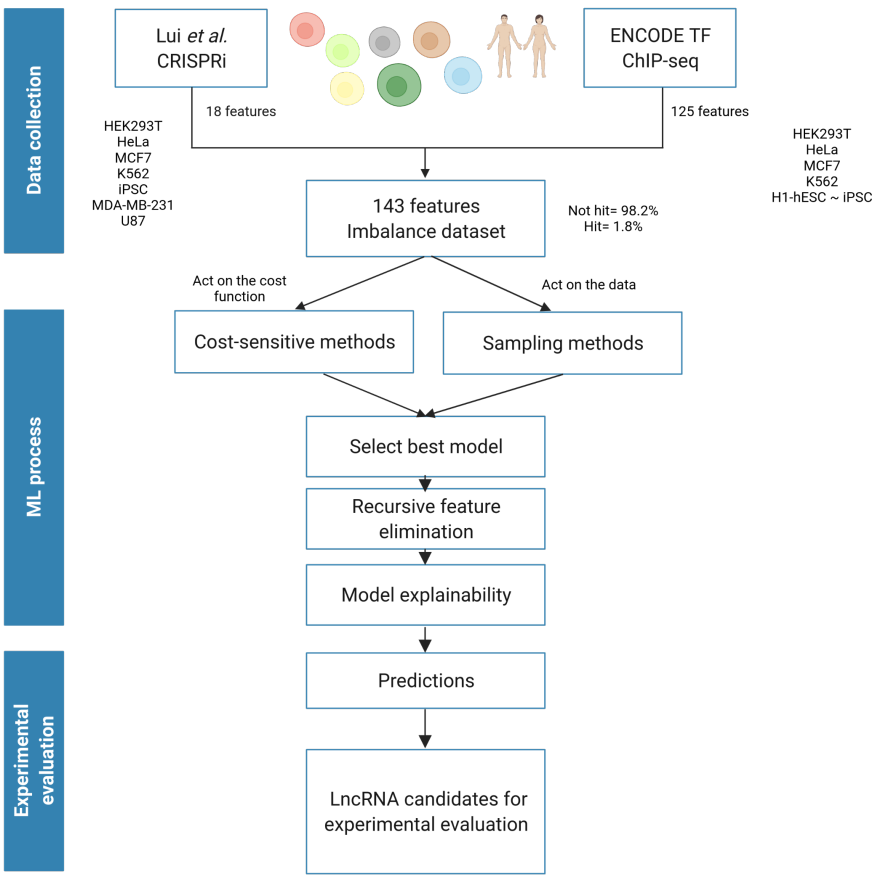

Figure 10: Machine learning (ML) workflow. Processes followed to build and evaluate machine learning models, and generate candidates for experimental validation.

II.2. LncRNA analysis of the *Drosophila* genome during regeneration

II.2.1. Characterization of cell-damage lncRNAs

II.2.1.1. Mapping and quantification

RNA-seq regeneration data was obtained from Vizcaya-Molina *et al.*[138] (**??**). Data was processed using the *grape-nf* pipeline (https://github.com/guigolab/grape-nf). RNA-seq reads were aligned to the fly genome (dm6) using *STAR*[dobin_2013_star] v2.4.0j with up to 4 mismatches per paired alignment using the FlyBase genome annotation r6.29 (04 2019). Only alignments for reads mapping to ten or fewer loci were reported. Genes and transcripts were quantified in TPMs using *RSEM*[li_2011_rsem] v1.2.21. The gtf version r6.29 contains a total of 16,412 genes; 13,957 protein coding genes (PCGs) and 2,455 long noncoding RNAs (lncRNAs). In our study the lncRNAs were defined as genes > 200 bp and aligned to canonical chromosomes. See **??** to have a general overview of the gene expression analysis in regeneration.

II.2.1.2. Quality control of BAM files

Quality control of alignment sequencing data was performed using *QualiMap*[garcia_2012_qualimap] v.2.2.1 and *Picard* v.2.6.0 (http://broadinstitute.github.io/picard/). Using *Qualimap* we obtained: number of reads, number of mapped reads, duplication rate, and GC percentage; and using *Picard* we obtained: dropout, and GC dropout.

Assessment of replicates reliability was measured with weighted correlation network analysis (WGCNA). WGCNA was implemented with the *R* package *WGCNA*[langfelder_wgcna] version 1.69. A cutoff of less than 2 standard deviations from a normal distribution was implemented to utilize a replicate.

II.2.1.3. Differential gene expression comparing: regeneration vs. control

Differential gene expression analyses between control and regeneration were performed separately on each time-point. Genes were filtered per time point, removing

all genes with a gene expression < 1 TPM. Analyses were run using *R* version 3.6.2, *DESeq2*[love_2014_deseq2], and a fold change[138] approach. All genes with an absolute fold change > 1.7 in both methods were considered differentially expressed.

In addition, the PCGs *rpr* and *Gadd45* were used as positive controls, and both were upregulated at the early time point. Positive controls were confirmed through qPCR (*Confirmation experiments were carried out by Carlos Camilleri, from Montserrat Coromina's lab at Universitat de Barcelona; Barcelona, Spain*).

II.2.1.4. Coding potential

The coding capability of lncRNAs was measured using *CPAT*[wang_2013_cpat] version 3.0.4. Following the developer's indications, we took a cut-off < 0.39 to classify them as noncoding RNAs.

II.2.1.5. ATAC-seq analysis

Uniquely aligned reads to canonical chromosomes from nucleosome-free data was retrieved from Vizcaya-Molina *et al.*[138] study. Aggregation plots around lncRNA TSS (± 400 bp) were produced using *bwtool*[pohl_bwtool] *summary* version 1.0. Bed6 files were used as input to *bwtool* to take into account gene strandness. LncRNA promoters were obtained using a 301 bp window centered on the main transcription start site (TSS).[batut_2017]

II.2.1.6. Genome-wide lncRNA classification

LncRNA genes were classified with respect to their genome location using the classification module of the *FEELnc*[31] pipeline. *FEELnc* received as input the 2,455 annotated lncRNAs from the gtf version r6.29 classifying the lncRNAs in three broad groups: **1)** intergenic (**??**A), **2)** genic intronic (**??**B), and **3)** genic exonic (**??**C). The classification was mutually exclusive in the following rank: genic exonic > genic intronic > intergenic. Genic exonic and genic intronic were subcategorized as: sense or antisense, and as: overlapping, nested or containing. Intergenic were subcategorized as: same strand, divergent or convergent.

To calculate the percentage of overlapping between the genic exonic and their overlapping PCGs, we took all genic exonic pairs and their overlapping PCGs. Then using *BEDTools*[quinlan_2010_bedtools] intersect v2.27, we obtained the number of base

pairs that overlapped between genic-exonic exons and PCG exons and divided by the total exon length.

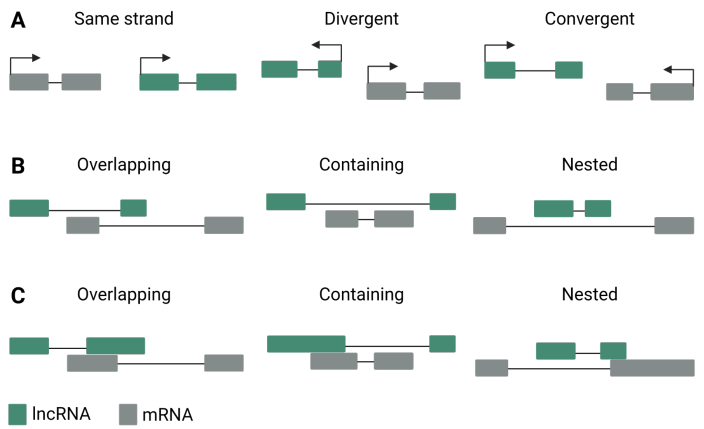

Figure 11: LncRNA classification. **(A)** Intergenic classification. **(B)** Genic intronic and **(C)** genic exonic: overlapping, containing, and nested in sense and in antisense. Figure inspired by Wucher *et al.*[31]

II.2.1.7. Gene ontology enrichment

For each differentially expressed lncRNA, the expressed set of neighboring PCGs were extracted (FlyBase version r6.29). For genic (exonic and intronic) all overlapping PCGs, and for intergenic PCGs with a distance $\leq$ of 5 Kb on each side were considered (**??**A).

Next, the *R* library *clusterProfiler*$^{\textbf{yu\_2012}}$ version 3.14.3 was used in combination with *Drosophila* annotations from the *R* library *org.Dm.eg.db*$^{\textbf{carlson\_2013}}$ version 3.10.0 to compute the gene ontology enrichment, using biological processes. *P-values* were adjusted using *FDR* multiple testing correction.

II.2.1.8. LncRNA:PCG co-expression analysis

To study the expression correlations between lncRNAs and PCGs, we used our lncRNA classification (genic exonic, genic intronic and intergenic) to automatically identify all lncRNA:PCG pairs. For long intervening noncoding RNAs (lincRNA; in this thesis work lincRNA and intergenic terms were used indistinctly), we kept

all pairs that showed a locus-locus distance $\leq$ of 5 Kb up and downstream of each lincRNA, and for genic lncRNAs all their overlapping PCGs (**??**A).

Next, we performed a regeneration and control specific analysis, removing lncRNA:PCG pairs that were expressed < 1 TPM, in control or in regeneration, and performed two analysis: **(1)** observe the DE status of lncRNA-PCG pairs, and **(2)** classify the expression patterns of lncRNA-PCG pairs.

The DE status of lncRNA-PCG pairs consisted in classifying them as concordant or discordant. Concordant cases where defined as positive directionality (*i.e.* lncRNA:upregulated and PCG:upregulated or lncRNA:downregulated and PCG:downregulated) and discordant cases were the opposite (*i.e.* lncRNA:upregulated and PCG:downregulated or lncRNA:downregulated and PCG:upregulated).

For the classification of expression patterns among lncRNA-PCG pairs, increasing, decreasing, peak and valley were the implemented classes (**??**B). We labeled as: increasing, if the lncRNA increased its expression during the three time-points; decreasing if it decreased its expression in all time-points; peak, if the maximum expression was at the mid time-point (15h); and finally valley, if the minimum expression was at the mid time-point.

After our classification, we retained the concordant (*i.e.* lncRNA:PCG: increasing-increasing; decreasing-decreasing; peak-peak; valley-valley) and discordant cases (*i.e.* lncRNA:PCG: increasing-decreasing; decreasing-increasing; peak-valley; valley-peak). In this analysis lncRNAs define the co-expression label, *e.g.* concordant-valley= lncRNA is valley and neighboring PCG is valley, discordant-valley= lncRNA is valley and neighboring PCG is peak.

II.2.1.9. LncRNA genomic features

GC content and length of: genes, promoters, and transcripts of all lncRNA were obtained using the *GC* class from *Biopython*[cock_2009] version 1.78, and gtf file version r6.29. Then, *Kruskal-Wallis* test was used to compare the *GC* percentage and length among: lncRNA differentially expressed (DE), lncRNA expressed (in regeneration and/or control), and the rest of annotated lncRNAs. For cases with a *p value* < 0.05 a pairwise Wilcoxon test was performed to obtain the *p value* of each comparison. The *FDR* correction method was used.

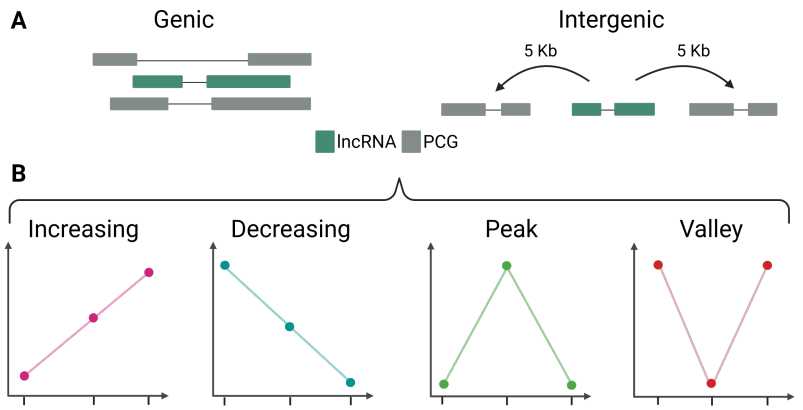

Figure 12: LncRNA:PCG co-expression analysis. **(A)** LncRNA-PCG pair selection strategy. **(B)** Co-expression classification; *y-axis*= gene expression, and *x-axis*: 0h, 15h, and 25h time points.

II.2.1.10. Sequence conservation

LncRNA sequence conservation was obtained using the *dm6* 27-way multiple alignment (23 *Drosophila* sequences, house fly, *Anopheles* mosquito, honey bee and red flour beetle) from the *UCSC genome browser*.[tyner_2017_ucsc] Next, *maf_parse* from *PHAST*[hubisz_2011_phast] v1.4 was used to do multiple alignments, and finally the maximum alignment score was taken with its respective number of aligned sequences and number of conserved species.

Two analyses were done, the first one was using the gene sequence (*i.e.* exons and introns), and the second one was using the exons and then calculating the mean conservation for exons by gene. Percentage of conservation was obtained dividing the length of aligned sequences by the length of the genomic feature.

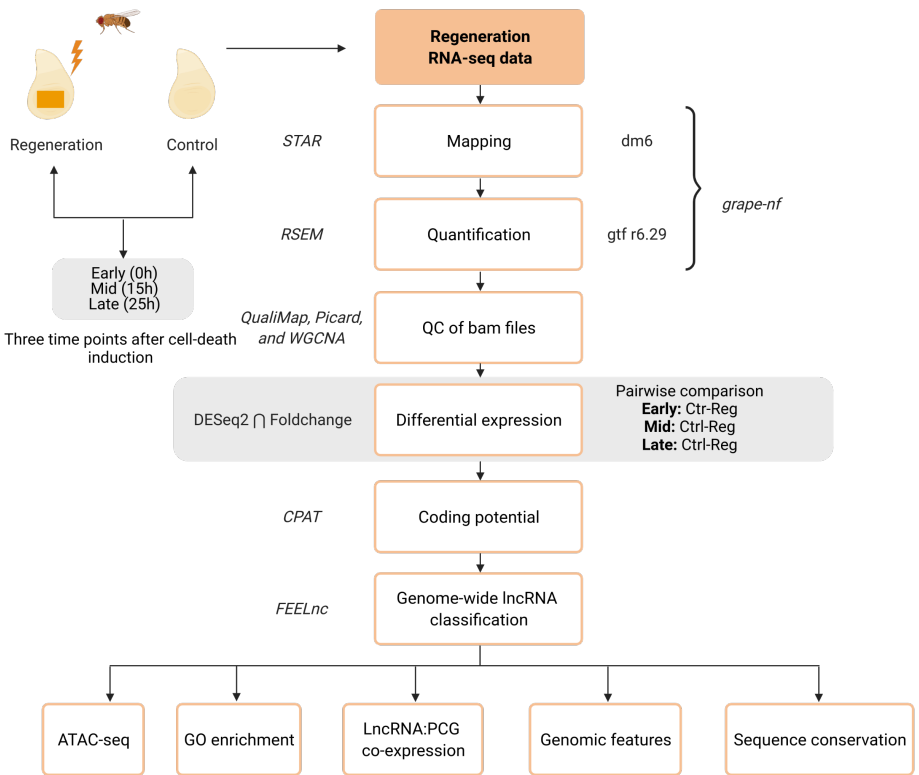

Figure 13: Gene expression analysis workflow. LncRNA analysis of the *Drosophila* genome during regeneration.

II.2.2. LncRNA developmental and tissue signatures

II.2.2.1. Mapping, quantification, and QC

Developmental RNA-seq data of *D. melanogaster* was obtained from the modENCODE project[celniker_2009, modencode_2010] (http://modencode.org/), 21 distinct developmental stages were analyzed from embryonic to pupal stage. These stages included 12 embryonic stages divided at 2h intervals from 0h to 24h, 3 larval stages, and 6 pupal stages, with an average of 3 replicate for each stage.

D. melanogaster leg and wing imaginal disc reads were obtained from Pérez-Lluch *et al.*[71] study. Eye-antenna imaginal discs were dissected into two separated antenna and eye imaginal discs, and subsequently antenna and eye imaginal discs were individually sequenced. (*These experiments were carried out by Marina Ruiz-Romero, from Roderic Guigó's lab at CRG; Barcelona, Spain*). Antenna, eye, leg and wing imaginal disc data was produced for three developmental stages L3, WPP and late pupae, with 2 replicates for each imaginal disc and developmental stage.

Developmental and imaginal disc datasets were mapped, quantified, and QC analyzed exactly as the regeneration dataset (see **??** and **??**).

II.2.2.2. K-means clustering

For the cluster analysis the developmental dataset was divided in two groups: the first group (embryo-larvae group) contained the 12 embryonic stages and the three larval stages, and the second group (pupae group) contained the 6 pupal stages.

Only lncRNAs expressed in at least one condition for the embryo-larvae group or for the pupae group were selected for the cluster analysis based on gene expression. Then, TPMs were $\log_{10}$ transformed and scaled before doing the clustering.

We iteratively implemented the k-means algorithm using the *R* function *kmeans* and run the algorithm 10 times with random initialized centroids. Following, the clusters were filtered to remove elements with a PCA distance from the cluster centroid above cluster mean distance. Finally, we recalculated the clusters until we reached robust clusters for the embryo-larvae group and for the pupae group.

II.2.3. Assessing the lncRNA:*CR40469* function during *D. melanogaster* imaginal-disc regeneration-process

II.2.3.1. *CR40469* knockout and induction of cell-death

The lncRNA *CR40469* was knocked-out (KO) by homozygous deletion using ends-out homologous recombination.[baena_2013] *CR40469* KO ($CR40469^{KO}$) deletion was confirmed via genomic qPCR. We used $CR40469^{KO}$ and *CR40469* wild-type ($CR40469^{Wt}$; Wt= wild-type) genotypes in combination with induction of cell-death at the early time point (regeneration 0h) and without induction of cell-death at the early time point (control 0h) to study the effects on gene expression. Obtaining four combinations: **(1)** $CR40469^{KO}$ in regeneration at 0h, **(2)** $CR40469^{KO}$ in control at 0h, **(3)** $CR40469^{Wt}$ in regeneration at 0h, and **(4)** $CR40469^{Wt}$ in control at 0h (see **??**).

Cell-death was induced using the expression of the pro-apoptotic *rpr* gene according to.[138,139] Regeneration experiments were performed for 16h at the L3 stage in the *salm* domain. In our study, control samples without *rpr* expression were treated in parallel. (*These experiments were carried out by Carlos Camilleri, from Montserrat Coromina's lab at Universitat de Barcelona; Barcelona, Spain*).

II.2.3.2. RNA-seq library preparation, sequencing, and processing

A total of 40 wing imaginal discs were dissected for each genotype ($CR40469^{KO}$ and $CR40469^{Wt}$) and cell-death condition (regeneration and control). Three technical replicates and three independent biological replicates were performed per condition. All libraries were sequenced on Illumina HiSeq at the Ultra sequencing unit of the Centre for Genomic Regulation (CRG, Barcelona, Spain). (*These experiments were carried out by Carlos Camilleri, from Montserrat Coromina's lab at Universitat de Barcelona; Barcelona, Spain*).

Mapping, quantification, and quality control analyses were carried out using the same process described above. See **??** and **??** for further details.

II.2.3.3. Differential gene expression

We used the statistical methods implemented in the *DESeq2*[love_2014_deseq2] package version 1.26.0. Only genes expressed at least 1 TPM in at least one sample were selected for this analysis. The two factors with interaction approach was implemented,

using the following design matrix:

$$design\ matrix = model.matrix(\sim genotype + condition + genotype : condition)$$

where genotype is $CR40469^{KO}$ or $CR40469^{Wt}$ and condition is regeneration or control. All genes with an absolute fold change > 1.7 and an adjusted *p-value* < 0.05 were considered differentially expressed. The *Benjamini-Hochberg* correction method was used.

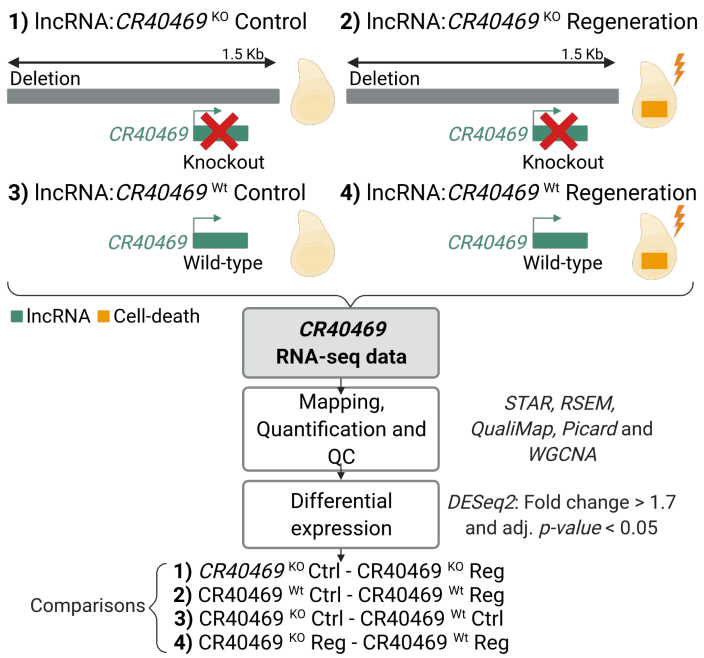

Figure 14: *CR40469* KO analysis workflow. Description of the four types of samples at the early time point, pipeline used to process raw reads, and the four comparisons performed for the differential expression analysis.

Bibliography

1. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (2012).
2. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316,** 1484–1488 (2007).
3. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458,** 223–227 (2009).
4. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research* **22,** 1775–1789 (2012).
5. Brown, J. B. *et al.* Diversity and dynamics of the Drosophila transcriptome. *Nature* **512,** 393–399 (2014).
6. Ulitsky, I. Interactions between short and long noncoding RNAs. *FEBS letters* **592,** 2874–2883 (2018).
7. Pelechano, V. & Steinmetz, L. M. Gene regulation by antisense transcription. *Nature Reviews Genetics* **14,** 880–893 (2013).
8. Jarroux, J., Morillon, A. & Pinskaya, M. History, discovery, and classification of lncRNAs. *Long Non Coding RNA Biology*, 1–46 (2017).
9. Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews Molecular Cell Biology* **22,** 96–118 (2021).
10. Kopp, F. & Mendell, J. T. Functional classification and experimental dissection of long noncoding RNAs. *Cell* **172,** 393–407 (2018).
11. Kim, T.-K., Hemberg, M. & Gray, J. M. Enhancer RNAs: a class of long noncoding RNAs synthesized at enhancers. *Cold Spring Harbor perspectives in biology* **7,** a018622 (2015).
12. Frankish, A. *et al.* GENCODE 2021. *Nucleic acids research* **49,** D916–D923 (2021).
13. Thurmond, J. *et al.* FlyBase 2.0: the next generation. *Nucleic acids research* **47,** D759–D765 (2019).
14. Consortium, E. P. *et al.* The ENCODE (ENCyclopedia of DNA elements) project. *Science* **306,** 636–640 (2004).
15. Rao, M. R. S. *Long Non Coding RNA Biology* (Springer, 2017).
16. Hon, C.-C. *et al.* An atlas of human long non-coding RNAs with accurate 5 ends. *Nature* **543,** 199–204 (2017).
17. Loda, A. & Heard, E. Xist RNA in action: Past, present, and future. *PLoS genetics* **15,** e1008333 (2019).
18. Kim, M., Faucillion, M.-L. & Larsson, J. RNA-on-X 1 and 2 in Drosophila melanogaster fulfill separate functions in dosage compensation. *PLoS genetics* **14,** e1007842 (2018).
19. Gelbart, M. E., Larschan, E., Peng, S., Park, P. J. & Kuroda, M. I. Drosophila MSL complex globally acetylates H4K16 on the male X chromosome for dosage compensation. *Nature structural & molecular biology* **16,** 825–832 (2009).
20. Meller, V. H. & Rattner, B. P. The roX genes encode redundant male-specific lethal transcripts required for targeting of the MSL complex. *The EMBO journal* **21,** 1084–1091 (2002).
21. Ota, T. *et al.* Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nature genetics* **36,** 40–45 (2004).
22. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *science* **309,** 1559–1563 (2005).
23. Stapleton, M. *et al.* The Drosophila gene collection: identification of putative full-length cDNAs for 70% of D. melanogaster genes. *Genome research* **12,** 1294–1300 (2002).
24. Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465,** 182–187 (2010).
25. De Santa, F. *et al.* A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS biology* **8,** e1000384 (2010).
26. Arnold, P. R., Wells, A. D. & Li, X. C. Diversity and emerging roles of enhancer RNA in regulation of gene expression and cell fate. *Frontiers in cell and developmental biology* **7,** 377 (2020).
27. Lagarde, J. *et al.* High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nature genetics* **49,** 1731–1740 (2017).
28. Zhao, L. *et al.* NONCODEV6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. *Nucleic Acids Research* **49,** D165–D171 (2021).
29. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research* **44,** D733–D745 (2016).
30. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nature genetics* **47,** 199–208 (2015).
31. Wucher, V. *et al.* FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic acids research* **45,** e57–e57 (2017).

32. Adams, M. D. *et al.* The genome sequence of Drosophila melanogaster. *Science* **287,** 2185–2195 (2000).

33. Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R. & Johnson, R. Towards a complete map of the human long non-coding RNA transcriptome. *Nature Reviews Genetics* **19,** 535–548 (2018).

34. Li, K. *et al.* Insights into the Functions of LncRNAs in Drosophila. *International journal of molecular sciences* **20,** 4646 (2019).

35. Pasquinelli, A. E. *et al.* Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* **408,** 86–89 (2000).

36. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505,** 635–640 (2014).

37. Haerty, W. & Ponting, C. P. Mutations within lncRNAs are effectively selected against in fruitfly but not in human. *Genome biology* **14,** 1–16 (2013).

38. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development* **25,** 1915–1927 (2011).

39. Marques, A. C. & Ponting, C. P. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome biology* **10,** 1–12 (2009).

40. Washietl, S., Kellis, M. & Garber, M. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome research* **24,** 616–628 (2014).

41. Quinn, J. J. *et al.* Rapid evolutionary turnover underlies conserved lncRNA–genome interactions. *Genes & development* **30,** 191–207 (2016).

42. Pegueroles, C. *et al.* Transcriptomic analyses reveal groups of co-expressed, syntenic lncRNAs in four species of the genus Caenorhabditis. *RNA biology* **16,** 320–329 (2019).

43. Novikova, I. V., Dharap, A., Hennelly, S. P. & Sanbonmatsu, K. Y. 3S: shotgun secondary structure determination of long non-coding RNAs. *Methods* **63,** 170–177 (2013).

44. Managadze, D., Rogozin, I. B., Chernikova, D., Shabalina, S. A. & Koonin, E. V. Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome biology and evolution* **3,** 1390–1404 (2011).

45. Yang, J.-R. & Zhang, J. Human long noncoding RNAs are substantially less folded than messenger RNAs. *Molecular biology and evolution* **32,** 970–977 (2015).

46. Schertzer, M. D. *et al.* lncRNA-induced spread of polycomb controlled by genome architecture, RNA abundance, and CpG island DNA. *Molecular cell* **75,** 523–537 (2019).

47. Santoro, F. & Pauler, F. M. Silencing by the imprinted Airn macro lncRNA: Transcription is the answer. *Cell cycle* **12,** 711–712 (2013).

48. Hezroni, H. *et al.* Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell reports* **11,** 1110–1122 (2015).

49. Herrera-Úbeda, C. *et al.* Microsyntenic clusters reveal conservation of lncRNAs in chordates despite absence of sequence conservation. *Biology* **8,** 61 (2019).

50. Ulitsky, I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nature Reviews Genetics* **17,** 601–614 (2016).

51. Flintoft, L. Structure and function for lncRNAs. *Nature reviews genetics* **14,** 598–598 (2013).

52. Couso, J.-P. & Patraquim, P. Classification and function of small open reading frames. *Nature reviews Molecular cell biology* **18,** 575–589 (2017).

53. Andrews, S. J. & Rothnagel, J. A. Emerging evidence for functional peptides encoded by short open reading frames. *Nature Reviews Genetics* **15,** 193–204 (2014).

54. Pueyo, J. I., Magny, E. G. & Couso, J. P. New peptides under the s (ORF) ace of the genome. *Trends in biochemical sciences* **41,** 665–678 (2016).

55. Quek, X. C. *et al.* lncRNAdb v2. 0: expanding the reference database for functional long noncoding RNAs. *Nucleic acids research* **43,** D168–D173 (2015).

56. Wutz, A. *et al.* Imprinted expression of the Igf2r gene depends on an intronic CpG island. *Nature* **389,** 745–749 (1997).

57. Arab, K. *et al.* GADD45A binds R-loops and recruits TET1 to CpG island promoters. *Nature genetics* **51,** 217–223 (2019).

58. Holdt, L. M. *et al.* Alu elements in ANRIL noncoding RNA at chromosome 9p21 modulate atherogenic cell functions through trans-regulation of gene networks. *PLoS genetics* **9,** e1003588 (2013).

59. Luo, H. *et al.* HOTTIP lncRNA promotes hematopoietic stem cell self-renewal leading to AML-like disease in mice. *Cancer Cell* **36,** 645–659 (2019).

60. Jain, A. K. *et al.* LncPRESS1 is a p53-regulated LncRNA that safeguards pluripotency by disrupting SIRT6-mediated de-acetylation of histone H3K56. *Molecular cell* **64,** 967–981 (2016).

61. Mas, A. M. & Huarte, M. lncRNA–DNA hybrids regulate distant genes. *EMBO reports* **21,** e50107 (2020).

62. Dueva, R. *et al.* Neutralization of the positive charges on histone tails by RNA promotes an open chromatin structure. *Cell chemical biology* **26,** 1436–1449 (2019).

63. Wang, K. C. *et al.* A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472,** 120–124 (2011).

64. Ariel, F. *et al.* R-loop mediated trans action of the APOLO long noncoding RNA. *Molecular cell* **77,** 1055–1065 (2020).

65. Seila, A. C. *et al.* Divergent transcription from active promoters. *science* **322,** 1849–1851 (2008).

66. Luo, S. *et al.* Divergent lncRNAs regulate gene expression and lineage differentiation in pluripotent cells. *Cell stem cell* **18,** 637–652 (2016).

67. Core, L. J. *et al.* Defining the status of RNA polymerase at promoters. *Cell reports* **2,** 1025–1035 (2012).

68. Carnesecchi, J., Pinto, P. B. & Lohmann, I. Hox transcription factors: an overview of multi-step regulators of gene expression. *International Journal of Developmental Biology* **62,** 723–732 (2018).

69. Hobson, D. J., Wei, W., Steinmetz, L. M. & Svejstrup, J. Q. RNA polymerase II collision interrupts convergent transcription. *Molecular cell* **48,** 365–374 (2012).

70. Nussbaumer, U., Halder, G., Groppe, J., Affolter, M. & Montagne, J. Expression of the blistered/DSRF gene is controlled by different morphogens during Drosophila trachea and wing development. *Mechanisms of development* **96,** 27–36 (2000).

71. Pérez-Lluch, S. *et al.* bsAS, an antisense long noncoding RNA, essential for correct wing development through regulation of blistered/DSRF isoform usage. *PLoS genetics* **16,** e1009245 (2020).

72. Rom, A. *et al.* Regulation of CHD2 expression by the Chaserr long noncoding RNA gene is essential for viability. *Nature communications* **10,** 1–15 (2019).

73. Engreitz, J. M. *et al.* Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539,** 452–455 (2016).

74. Anderson, K. M. *et al.* Transcription of the noncoding RNA upperhand controls Hand2 expression and heart development. *Nature* **539,** 433–436 (2016).

75. Mikhaylichenko, O. *et al.* The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. *Genes & development* **32,** 42–57 (2018).

76. Syed, K. M. & Hon, C.-C. Heterogeneity among enhancer RNAs: origins, consequences and perspectives. *Essays in Biochemistry* (2021).

77. Romero-Barrios, N., Legascue, M. F., Benhamed, M., Ariel, F. & Crespi, M. Splicing regulation by long noncoding RNAs. *Nucleic acids research* **46,** 2169–2184 (2018).

78. Siomi, H. & Siomi, M. C. Posttranscriptional regulation of microRNA biogenesis in animals. *Molecular cell* **38,** 323–332 (2010).

79. Garaulet, D. L. *et al.* Homeotic function of Drosophila Bithorax-complex miRNAs mediates fertility by restricting multiple Hox genes and TALE cofactors in the CNS. *Developmental cell* **29,** 635–648 (2014).

80. Maeda, R. K. *et al.* The lncRNA male-specific abdominal plays a critical role in Drosophila accessory gland development and male fertility. *PLoS genetics* **14,** e1007519 (2018).

81. Kallen, A. N. *et al.* The imprinted H19 lncRNA antagonizes let-7 microRNAs. *Molecular cell* **52,** 101–112 (2013).

82. Keniry, A. *et al.* The H19 lincRNA is a developmental reservoir of miR-675 that suppresses growth and Igf1r. *Nature cell biology* **14,** 659–665 (2012).

83. Cesana, M. *et al.* A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* **147,** 358–369 (2011).

84. Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P. P. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* **146,** 353–358 (2011).

85. Grelet, S. *et al.* A regulated PNUTS mRNA to lncRNA splice switch mediates EMT and tumour progression. *Nature cell biology* **19,** 1105–1115 (2017).

86. Soshnev, A. A. *et al.* A conserved long noncoding RNA affects sleep behavior in Drosophila. *Genetics* **189,** 455–468 (2011).

87. He, R.-Z., Luo, D.-X. & Mo, Y.-Y. Emerging roles of lncRNAs in the post-transcriptional regulation in cancer. *Genes & diseases* **6,** 6–15 (2019).

88. Jiang, K. *et al.* Akt2 regulation of Cdc2-like kinases (Clk/Sty), serine/arginine-rich (SR) protein phosphorylation, and insulin-induced alternative splicing of PKCβII messenger ribonucleic acid. *Endocrinology* **150,** 2087–2097 (2009).

89. Cooper, D. R. *et al.* Long non-coding RNA NEAT1 associates with SRp40 to temporally regulate PPARγ2 splicing during adipogenesis in 3T3-L1 cells. *Genes* **5,** 1050–1063 (2014).

90. Wang, X. *et al.* LncRNA MALAT1 promotes development of mantle cell lymphoma by associating with EZH2. *Journal of translational medicine* **14,** 1–14 (2016).

91. Malakar, P. *et al.* Long noncoding RNA MALAT1 promotes hepatocellular carcinoma development by SRSF1 upregulation and mTOR activation. *Cancer research* **77,** 1155–1167 (2017).

92. Tripathi, V. *et al.* The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Molecular cell* **39,** 925–938 (2010).

93. Krystal, G. W., Armstrong, B. & Battey, J. N-myc mRNA forms an RNA-RNA duplex with endogenous antisense transcripts. *Molecular and cellular biology* **10,** 4180–4191 (1990).

94. Villamizar, O., Chambers, C. B., Riberdy, J. M., Persons, D. A. & Wilber, A. Long noncoding RNA Saf and splicing factor 45 increase soluble Fas and resistance to apoptosis. *Oncotarget* **7,** 13810 (2016).

95. Villamizar, O. *et al.* Fas-antisense long noncoding RNA is differentially expressed during maturation of human erythrocytes and confers resistance to Fas-mediated cell death. *Blood Cells, Molecules, and Diseases* **58,** 57–66 (2016).

96. Beltran, M. *et al.* A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial–mesenchymal transition. *Genes & development* **22,** 756–769 (2008).

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

98. Haswell, J. R. *et al.* Genome-wide CRISPR interference screen identifies long non-coding RNA loci required for differentiation and pluripotency. *bioRxiv* (2021).

99. Liu, S. *et al.* Wnt-regulated lncRNA discovery enhanced by in vivo identification and CRISPRi functional validation. *Genome medicine* **12,** 1–22 (2020).

100. Cai, P. *et al.* A genome-wide long noncoding RNA CRISPRi screen identifies PRANCR as a novel regulator of epidermal homeostasis. *Genome research* **30,** 22–34 (2020).

101. Liu, S. J. *et al.* CRISPRi-based radiation modifier screen identifies long non-coding RNA therapeutic targets in glioma. *Genome biology* **21,** 1–18 (2020).

102. Perry, R. B.-T. & Ulitsky, I. The functions of long noncoding RNAs in development and stem cells. *Development* **143,** 3882–3894 (2016).

103. Gao, F., Cai, Y., Kapranov, P. & Xu, D. Reverse-genetics studies of lncRNAs—what we have learnt and paths forward. *Genome biology* **21,** 1–23 (2020).

104. Morelli, E. *et al.* in *Long Non-Coding RNAs in Cancer* 189–204 (Springer, 2021).

105. Clemson, C. M. *et al.* An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Molecular cell* **33,** 717–726 (2009).

106. Khaitan, D. *et al.* The melanoma-upregulated long noncoding RNA SPRY4-IT1 modulates apoptosis and invasion. *Cancer research* **71,** 3852–3862 (2011).

107. Meng, Q. *et al.* The DGCR5 long noncoding RNA may regulate expression of several schizophrenia-related genes. *Science translational medicine* **10** (2018).

108. Stojic, L. *et al.* Specificity of RNAi, LNA and CRISPRi as loss-of-function methods in transcriptional analysis. *Nucleic acids research* **46,** 5950–5966 (2018).

109. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339,** 819–823 (2013).

110. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339,** 823–826 (2013).

111. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152,** 1173–1183 (2013).

112. Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154,** 442–451 (2013).

113. Maeder, M. L. *et al.* CRISPR RNA–guided activation of endogenous human genes. *Nature methods* **10,** 977–979 (2013).

114. Konermann, S. *et al.* Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517,** 583–588 (2015).

115. Guo, X. *et al.* Transcriptome-wide Cas13 guide RNA design for model organisms and viral RNA pathogens. *Cell Genomics,* 100001 (2021).

116. Larson, M. H. *et al.* CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nature protocols* **8,** 2180–2196 (2013).

117. Nuñez, J. K. *et al.* Genome-wide programmable transcriptional memory by CRISPR-based epigenome editing. *Cell* **184,** 2503–2519 (2021).

118. Raffeiner, P. *et al.* An MXD1-derived repressor peptide identifies noncoding mediators of MYC-driven cell proliferation. *Proceedings of the National Academy of Sciences* **117,** 6571–6579 (2020).

119. Alerasool, N., Segal, D., Lee, H. & Taipale, M. An efficient KRAB domain for CRISPRi applications in human cells. *Nature methods* **17,** 1093–1096 (2020).

120. Nuñez, J. K. *et al.* Genome-wide programmable transcriptional memory by CRISPR-based epigenome editing. *Cell* **184,** 2503–2519 (2021).

121. Thakore, P. I. *et al.* Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nature methods* **12,** 1143–1149 (2015).

122. Vizcaya-Molina, E., Klein, C. C., Serras, F. & Corominas, M. *Chromatin dynamics in regeneration epithelia: Lessons from Drosophila imaginal discs* in *Seminars in cell & developmental biology* **97** (2020), 55–62.

123. Iismaa, S. E. *et al.* Comparative regenerative mechanisms across different mammalian tissues. *NPJ Regenerative medicine* **3,** 1–20 (2018).

124. Kopp, J. L., Grompe, M. & Sander, M. Stem cells versus plasticity in liver and pancreas regeneration. *Nature cell biology* **18,** 238–245 (2016).

125. Han, M., Yang, X., Farrington, J. E. & Muneoka, K. Digit regeneration is regulated by Msx1 and BMP4 in fetal mice (2003).

126. Lehoczky, J. A. & Tabin, C. J. Lgr6 marks nail stem cells and is required for digit tip regeneration. *Proceedings of the National Academy of Sciences* **112,** 13249–13254 (2015).

127. Goldman, J. A. & Poss, K. D. Gene regulatory programmes of tissue regeneration. *Nature Reviews Genetics* **21,** 511–525 (2020).

128. Morgan, T. H. Regeneration and liability to injury. *Science* **14,** 235–248 (1901).

129. Hariharan, I. K. & Serras, F. Imaginal disc regeneration takes flight. *Current opinion in cell biology* **48,** 10–16 (2017).

130. González-Rosa, J. M., Burns, C. E. & Burns, C. G. Zebrafish heart regeneration: 15 years of discoveries. *Regeneration* **4,** 105–123 (2017).

131. Sergeeva, O., Sviridov, E. & Zatsepin, T. *Noncoding RNA in Liver Regeneration—From Molecular Mechanisms to Clinical Implications* in *Seminars in liver disease* **40** (2020), 070–083.

132. Bely, A. E. & Nyberg, K. G. Evolution of animal regeneration: re-emergence of a field. *Trends in ecology & evolution* **25,** 161–170 (2010).

133. Chen, C.-H. & Poss, K. D. Regeneration genetics. *Annual review of genetics* **51,** 63–82 (2017).

134. Pfefferli, C. & Jaźwińska, A. The art of fin regeneration in zebrafish. *Regeneration* **2,** 72–83 (2015).

135. Baghdadi, M. B. & Tajbakhsh, S. Regulation and phylogeny of skeletal muscle regeneration. *Developmental biology* **433,** 200–209 (2018).

136. Gonçalves, T. J. & Armand, A.-S. Non-coding RNAs in skeletal muscle regeneration. *Non-coding RNA research* **2,** 56–67 (2017).

137. Ji, J.-Y., Han, C. & Deng, W.-M. Understanding human diseases using Drosophila. *Journal of genetics and genomics= Yi chuan xue bao* **46,** 155–156 (2019).

138. Vizcaya-Molina, E. *et al.* Damage-responsive elements in Drosophila regeneration. *Genome research* **28,** 1852–1866 (2018).

139. Santabárbara-Ruiz, P. *et al.* ROS-induced JNK and p38 signaling is required for unpaired cytokine activation during Drosophila regeneration. *PLoS genetics* **11,** e1005595 (2015).

140. Blanco, E. *et al.* Gene expression following induction of regeneration in Drosophila wing imaginal discs. Expression profile of regenerating wing discs. *BMC developmental biology* **10,** 1–14 (2010).

141. Dong, X. *et al.* Non-coding RNAs in cardiomyocyte proliferation and cardiac regeneration: Dissecting their therapeutic values. *Journal of Cellular and Molecular Medicine* **25,** 2315–2332 (2021).

142. Venkatraman, A. *et al.* Maternal imprinting at the H19–Igf2 locus maintains adult haematopoietic stem cell quiescence. *Nature* **500,** 345–349 (2013).

143. Dey, B. K., Pfeifer, K. & Dutta, A. The H19 long noncoding RNA gives rise to microRNAs miR-675-3p and miR-675-5p to promote skeletal muscle differentiation and regeneration. *Genes & development* **28,** 491–501 (2014).

144. Li, C. *et al.* The role of lncRNA MALAT1 in the regulation of hepatocyte proliferation during liver regeneration. *International journal of molecular medicine* **39,** 347–356 (2017).

145. Han, X., Yang, F., Cao, H. & Liang, Z. Malat1 regulates serum response factor through miR-133 as a competing endogenous RNA in myogenesis. *The FASEB Journal* **29,** 3054–3064 (2015).

146. Wang, G.-q. *et al.* Sirt1 AS lncRNA interacts with its mRNA to inhibit muscle formation by attenuating function of miR-34a. *Scientific reports* **6,** 1–13 (2016).

147. Wang, Y. *et al.* Identification, stability and expression of Sirt1 antisense long non-coding RNA. *Gene* **539,** 117–124 (2014).

148. Cai, B. *et al.* The long noncoding RNA CAREL controls cardiac regeneration. *Journal of the American College of Cardiology* **72,** 534–550 (2018).