Scuola Internazionale Superiore di Studi Avanzati
Ph.D. course in Functional and Structural Genomics

Transposons acting as ceRNAs (TAC) hypothesis: initial evidence from *in-silico* analyses of LINE1 overexpression contexts

Thesis submitted for the degree of "*Philosophiae Doctor*"

Candidate:
Mauro Esposito

Supervisor:
Prof. Remo Sanges

Academic year 2021/2022

Table of Contents

ABSTRACT 1

1 INTRODUCTION 2

1.1 Transposable elements: the jumping genes 2

1.1.1 Classification 2

1.1.1.1 The class of DNA transposons 4

1.1.1.2 The class of retrotransposons 5

1.1.2 The LINE1 elements 7

1.1.2.1 Structure and activity 7

1.1.3 The transposable elements can be co-opted during the evolution 11

1.1.3.1 Influence at transcriptional level 13

1.1.3.2 Influence at the post-transcriptional level 14

1.1.4 Diseases caused by dysregulated LINE1 activity 15

1.2 Mechanisms repressing LINE1 19

1.2.1 Control of LINE1 transcript levels at DNA-level 19

1.2.1.1 The DNA methyltransferases family 22

1.2.1.2 ATRX: the enzyme for remodelling nucleosomes 24

1.2.2 Control of LINE1 transcript levels at RNA-level 26

1.2.2.1 MicroRNAs 27

1.2.2.2 miRNAs controlling LINE1: the let-7 and miR128 cases 32

1.3 Competitive endogenous RNAs 35

1.3.1 Factors regulating the ceRNA dynamics 38

1.3.2 The RIDL hypothesis: LincRNA-RoR and BACE1-AS cases 40

1.3.3 The quantification of TEs: challenges and solutions 42

2 AIM OF THE PROJECT 44

3 MATERIALS AND METHODS 45

3.1 Data collection and pre-processing 45

3.2 Analysis of locus-specific TE expression 46

3.3 Detection of LINE1 autonomous transcription 46

3.4 Gene expression analysis 47

3.5 Functional enrichment analysis 47

3.6 Overlap analysis 48

3.7 Analysis of miRNA target-sites sharing 48

3.8 Identification of miRNAs sequestered by LINE1s 49

3.9 miRNA-gene networks identification 49

3.10 Association analysis of miRNA-LINE1 expression levels 50

3.11 Analysis of TE expression at consensus level 50

4 RESULTS AND DISCUSSION 51

4.1 LINE1s deregulation upon DNMT1-KO 51

4.1.1 Evaluation of TE transcription 52

4.1.2 Tool for identifying autonomous LINE1 transcription 54

4.1.3 Analysis of LINE1 autonomous transcription 56

4.1.4 LINE1s deregulation in a not-autonomously transcribed LINE1 dataset 57

4.1.4.1 Evaluation of TE transcription 57

4.1.4.2 Analysis of LINE1 non-autonomous transcription 59

4.2 Effects of LINE1 dysregulation upon DNMT1-KO 61

4.2.1 Analysis of deregulated genes 61

4.2.2 Overlap analysis 62

4.2.3 Identification of decoyed miRNAs 64

4.2.4 The miR-128 case 66

4.3 Support for a putative LINE1 ceRNA activity 69

4.3.1 TE quantification....69
4.3.2 Active LINE1 transcription analysis....70
4.3.3 miRNA target-sites sharing analysis....70
4.4 ceRNA activity depend on autonomous LINE1 transcription and AGO2....72
4.4.1 Evaluation of TE deregulation....72
4.4.2 Analysis of autonomous LINE1 transcription....75
4.4.3 miRNA target-sites sharing analysis....75
4.4.4 AGO2 levels analysis....76
5 CONCLUSIONS....78
6 BIBLIOGRAPHY....82

List of Figures

Figure 1: The structure of a full-length LINE1 retrotransposon.....7
Figure 2: A representation of the relationships of the various LINE1 subfamilies.....8
Figure 3: Scheme of the retrotransposition mechanism of LINE1.....10
Figure 4: Transposable elements can influence gene expression at different levels....12
Figure 5: LINE1 dysregulation and diseases.....16
Figure 6: The DNA methylation reactions.....23
Figure 7: MicroRNA biogenesis.....28
Figure 8: Model for controlling LINE-1 transcripts by let-7.....34
Figure 9: The Basis of the ceRNA mechanism.....35
Figure 10: Factors influencing ceRNA dynamics.....38
Figure 11: Functional classification of TE insertions.....41
Figure 12: Evaluation of TE transcription upon the DNMT1-KO.....53
Figure 13: A schematic view of an Illumina paired-end read.....54
Figure 14: The KO of DNMT1 leads to autonomous LINE1 transcription.....56
Figure 15: Evaluation of TE transcription in quiescent naive $CD4^+$ T cells.....58
Figure 16: The quiescent $CD4^+$ T cells overexpress LINE1 non-autonomously.....60
Figure 17: LINE1 transcripts might act as ceRNA.....63
Figure 18: let-7 miRNA family might be decoyed by LINE1.....65
Figure 19: The miR-128 levels are associated to LINE1 transcripts.....67
Figure 20: LINE1 could act as ceRNA when artificially overexpressed.....71
Figure 21: Evaluation of TE transcription upon the ATRX-KO.....74
Figure 22: ceRNA activity might depend on autonomous transcription and AGO2....77

List of Tables

Table 1: A snapshot of the human genome transposable elements.....3
Table 2: Information about the analyzed dataset.....51
Table 3: Biological processes and WikiPathways enriched upon DNMT1-KO.....61

List of Abbreviations and Symbols

TE	Transposable element
LTR	Long terminal repeats
HERV	Human endogenous retrovirus
LINE	Long interspersed nuclear elements
SINE	Short interspersed nuclear elements
RNP	Ribonucleoprotein complex
HUSH	Human Silencing Hub
piRNA	Piwi-interacting RNA
AGO	Argonaute
MECP2	Methyl-CpG-binding protein 2
RISC	miRNA-induced silencing complex
DNMT	DNA methyltransferases
ATRX	Alpha Thalassemia/mental Retardation syndrome X-linked
MiRNA	Micro RNA
AGS	Aicardi-Goutieres Syndrome
CeRNA	Competitive endogenous RNA
RIDL	Repeat insertion domains of lncRNAs

Abstract

LINE1 are transposable elements that can replicate within the genome by passing through RNA intermediates. The vast majority of LINE1 copies in the human genome are inactive and just between 100/150 copies are full length and still potentially capable to mobilize. During the evolution, they could have been positively selected for cellular beneficial functions. Nonetheless, LINE1 deregulation can be detrimental to the cell causing diseases like cancer. The activity of miRNAs represents a fundamental mechanism for controlling transcript levels in somatic cells. These are a class of small non-coding RNAs that cause degradation or translational inhibition of their target transcripts. Beyond this, competitive endogenous RNAs (ceRNAs), mostly made by circular and non-coding RNAs, have been observed to compete for the binding of the same set of miRNAs targeting protein coding genes. In my PhD project, I have explored the possibility that autonomously transcribed LINE1s may act as ceRNAs. I observed that genes sharing miRNA target sites with LINE1 have a tendency to be upregulated when LINE1 are overexpressed suggesting that LINE1 might act as ceRNAs. This finding will help in the interpretation of transcriptomic responses in contexts characterized by specific activation of transposons.

1 Introduction

1.1 Transposable elements: the jumping genes

Transposable elements (TEs) are complex and interspersed DNA repeats that can change their position within the genome. Discovered by Barbara McClintock (McClintock, 1956) in 1948, they have been considered for a long time as "junk" DNA. Also, together with long non-coding RNAs they are referred to as the "dark matter" of the genome. In the last years, because of the improvement of sequencing technologies, the functions and the evolution of the "dark matter" of the human genome have become clearer to the scientific community (Kim, Lee and Han, 2012). This made it possible to distinguish between the presence of elements that have been tolerated by the host cells and those that were not preserved during the evolution because of their detrimental effects on the cells.

1.1.1 Classification

Due to the repetitive and interspersed nature, TEs represent ~50% of the human genome (Table 1) even though the vast majority of them are inactive copies that are no longer able to mobilize. The traditional classification of TEs divides them based on the mechanism of transposition in Class 1 or retrotransposons and Class 2 or DNA transposons. While the firsts are able to replicate within the genome with a "copy-and-paste" mechanism, the second ones use a "cut-and-paste" mechanism that enables them to change their position without directly increasing the genome size. In addition, each class includes autonomous and non-autonomous TEs (Kapitonov, Pavlicek and Jurka, 2006). A TE is considered autonomous when it is able to encode the set of proteins needed for the transposition which is the mobilization process. Conversely, a non-autonomous TE is defined as such when its activity depends on the protein machinery produced by autonomous TEs.

Human Genome ~3200 Mb		# of Copies (×1000)	Total Length (Mb)	% of Genome	Active
LINEs		**868**	**558.8**	**20.42**	
LINE1[1]		516	462	16.89	Active
LINE2		315	88.2	3.22	
LINE3		37	8.4	0.31	
SINEs		**1558**	**359.6**	**13.29**	
Alu[1]		1090	290.1	10.6	Active using L1 RT
MIR		393	60.1	2.2	
MIR3		75	9.3	0.34	
SVA[1]		2.76	4.2	0.15	Active using L1 RT
LTR retro-transposons		**443**	**227**	**8.29**	
ERV class I		112	79.2	2.89	
ERV (K) class II		8	8.5	0.31	
ERV (L) class III		83	39.5	1.44	
MaLR		240	99.8	3.65	
DNA transposons		**294**	**77.6**	**2.84**	
hAT	Charlie	182	38.1	1.39	
	Zaphod	13	4.3	0.16	
Tc-1	Tigger	57	28	1.02	
	Tc2	4	0.9	0.03	
	Mariner	14	2.6	0.1	
PiggyBac-like		2	0.5	0.02	
Unclassified		22	3.2	0.12	

Table 1: A snapshot of the human genome transposable elements.TEs represent ~50% of the human genome; the autonomous known active elements belong exclusively to the LINE1 family (Mandal and Kazazian, 2008).

1.1.1.1 The class of DNA transposons

The DNA transposons constitute the ~3% of the human genome and are considered "fossil" DNA since no element has demonstrated to be still active in human cells so far. Being "cut-and-paste" TEs, they are excised from the original genomic locus and then reinserted in a new position through the enzymatic activity of the transposase protein machinery. The function of this enzyme is to catalyze hydrolysis and transesterification reactions similar to V(D)J recombination process (Oettinger *et al.*, 1990).

In the human genome, the most numerous superfamilies of this class of transposons are hAT and TC1/mariner (Pace and Feschotte, 2007) transposons. hAT elements represent about two-thirds of the DNA transposons and, because of their ~3000-bp length, they should be able to encode a ~500-aa transposase. However, the majority of them are non-autonomous TEs since they accumulated deletions that partially or completely erase the sequence encoding the transposase. The remaining one-third of the DNA transposons is mainly composed of the TC1/mariner elements from which a noteworthy artificial application arose, the *Sleeping Beauty* element. This engineered TE is able to translocate from one to another DNA site allowing also the movement of an artificial sequence (Ivics *et al.*, 1997). Particularly, it is composed of a transposase sequence that was synthetically reconstructed to be functional. In addition to this, the element is coupled with the sequence of interest that is recognized by the translated protein and then inserted within the genome. With this system, this non-viral vector finds large employment in gene delivery experiments.

1.1.1.2 The class of retrotransposons

The class of retrotransposons, differently from the DNA transposons, is characterized by the capability to directly increase the genomic DNA amount upon transposition through its replicative mechanism. Typically, they are first transcribed in an RNA intermediate and then reverse transcribed in DNA into a new locus of the genome. Two are the broad classes that comprise the majority of retrotransposons: long terminal repeats (LTR) and non-LTR retrotransposons (Luning Prak and Kazazian, 2000).

The LTR-retrotransposons

The LTR-retrotransposons are non-autonomous elements thought to have originated from past retroviral infections (Lerat and Capy, 1999). It is believed that their origin resides in exogenous retroviruses that, infecting germline human cells and after having lost the capability to be infective, have become human endogenous retroviruses (HERVs). From the structural point of view, the LTR-retrotransposons share many similarities with the retroviruses. Indeed, they are composed of the proviral genes *gag* (group-specific antigen), *pol* (polymerase) and *env* (envelope) flanked on both ends by long terminal repeats (LTR) (Havecker, Gao and Voytas, 2004). Due to their "LTR-Sequence-LTR" structure, it has been proposed that, for many copies of these elements, homologous recombination events involving the LTRs have caused the removal of inserted pro-viruses, resulting in solo LTR elements. For this reason, the HERVs that are still retaining the original internal sequence flanked by LTRs are the smallest fraction. In total, HERVs represent about 7% (Smit, 1999) of the human genome and, accordingly to similarities with exogenous retroviruses, they are classified into three classes: Class I, II and III. Among these, the Class II seems to include the elements most likely to be functional, those belonging to the HERV-K subfamily. The discovery of retrovirus-like particles found in tumor-derived cell lines (Herbst, Sauter and Mueller-Lantzsch, 1996) suggests that the elements of this subfamily might not be fully silenced but more research is required in order to understand if they are still active in humans in healthy conditions.

The non-LTR retrotransposons

The non-LTR elements are the other broad class of retrotransposons that makes up about 34% of the human genome (Smit, 1999). Historically, they are classified considering the length of their sequences into the long interspersed nuclear elements (LINEs) and the short interspersed nuclear elements (SINEs). Indeed, while the firsts are ~6000 bp, the seconds are shorter with a length of just ~300 bp.

The SINEs are non-autonomous elements mainly composed of the Alu family in human. They are represented by more than one million copies making them the most abundant family of TEs in the human genome. Even though the vast majority carry inactivating mutations making them fossils, some active elements exist. Indeed, several Alus can currently participate in retrotransposition events taking advantage of LINE protein machinery. This is made possible because the Alu sequences are evolutionarily derived from the 7SL RNA, a scaffold RNA molecule which is a component of the signal recognition particle (SRP). The function of this complex is to bind newly synthesized peptides as soon as they emerge from the ribosomes for delivering the protein toward the endoplasmic reticulum. Since Alus can bind SRP proteins for the similarities with the 7SL RNA, it has been proposed that their RNA may be strategically positioned near ribosomes to hijack the retrotransposition proteins encoded by LINE elements (Boeke, 1997).

The LINE are autonomous elements composed by 3 different families: LINE1, LINE2 and LINE3. The last two families represent ~4% of the human genome and they are supposed to be originated from CR1-like retroelements, a widely distributed retrotransposon found in invertebrates as well as in mammals. LINE2 and LINE3 are nowadays inactive elements because of their old age and high frequency of 5' truncation events (Kapitonov, Pavlicek and Jurka, 2006).

1.1.2 The LINE1 elements

With about 500 thousand copies, LINE1 is a family of non-LTR retrotransposons that accounts for the 17% of the human genome (Lander *et al.*, 2001).

1.1.2.1 Structure and activity

A canonical full-length LINE1 copy (Figure 1) is ~6 kbp long (Scott *et al.*, 1987), lacks splicing signals and encodes for a bi-cistronic mRNA. Entering into details, the coding portion is composed by the ORF1 and the ORF2: while the first encodes for an RNA-binding protein (Kolosha and Martin, 1997), the second one for a protein with endonuclease (Feng *et al.*, 1996) and reverse transcriptase domains (Mathias *et al.*, 1991). The autonomous transcription is possible because of a bidirectional polII promoter contained in the 5'UTR (Swergold, 1990; Mätlik, Redik and Speek, 2006), while the final part is formed by the 3'UTR embedding a polyA tail. Among the huge amount of LINE1 elements, it has been estimated that the human genome contains just about 5000 full-length LINE1 elements (Sassaman *et al.*, 1997).

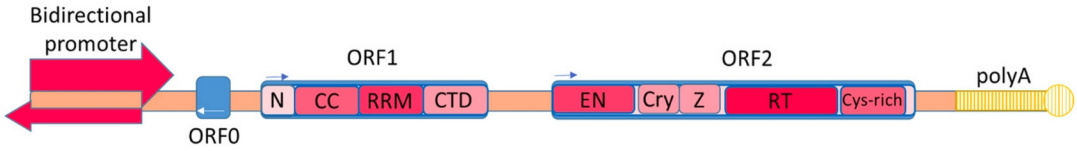

Figure 1: The structure of a full-length LINE1 retrotransposon. ORF1 consists of an RNA recognition motive (RRM) while the ORF2 is composed of endonuclease (EN) and retrotransposase (RT) domains. The autonomous transcription starts from the promoter region located in the 5'UTR region (Protasova, Andreeva and Rogaev, 2021).

Despite the high number of full-length copies, the vast majority of them are inactive and they are unable to mobilize within the genome. The reason of this lack of activity resides in several mutations and truncation events that have affected these LINE1 copies during the evolution (Mills *et al.*, 2007). ORF1 and ORF2 proteins, which are essential for the retrotransposition process, are unable to be translated properly as a result of these mutations. Nevertheless, about 100/150 LINE1 copies are found to be both full-lenght and potentially able to propagate throughout the human genome (Sassaman *et al.*, 1997) since they contain ORF sequences that are intact.

Subfamily classification

In the Smit *et al.* study (Smit *et al.*, 1995), the 3'UTR of LINE1s was analyzed to categorize the annotated copies in 47 different subfamilies. In their study, the LINE1 (L1) were classified as L1M, L1P or L1HS based on the distribution of the element in Mammalian, Primate or Human (Human-Specific) respectively. A letter was added to the name for defining the subdivision based on the 3'UTR structure (e.g. L1PA), while a final number starting from the most recently active source gene was used within each group (e.g. L1PA1).

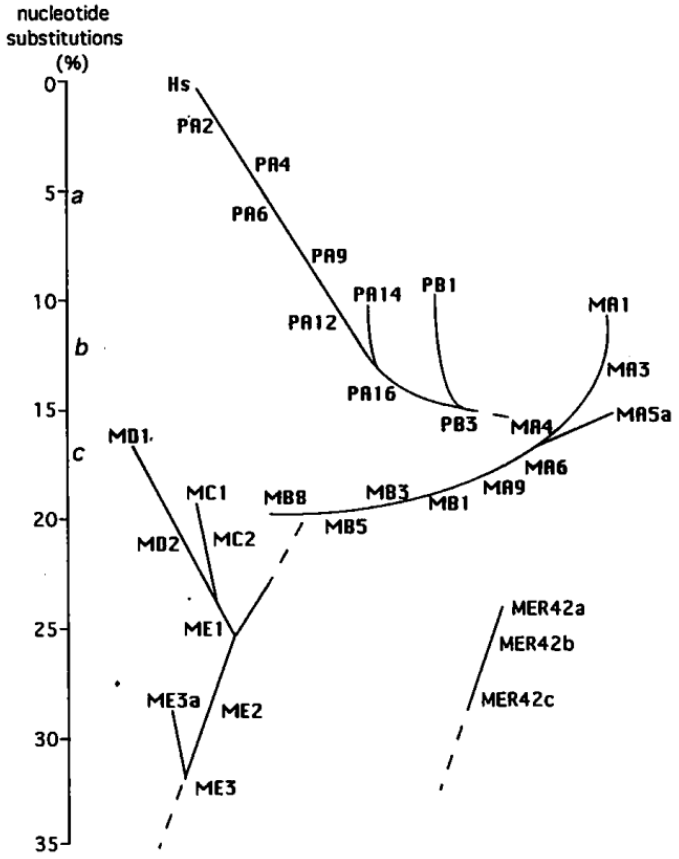

Figure 2: A representation of the relationships of the various LINE1 subfamilies. The position of each subfamily along the y-axis reflects their time of appearance. An idea of the age is indicated from the letters a, b and c that indicate the divergence of the human genome from the New World monkeys, the prosimians and other mammals, respectively.

Calculating the sequence divergence of each subfamily from the respective consensus (Figure 2), in this study was hypothesized the time elapsed since the appearance of the subfamily. With this procedure, they predicted that members of subfamily L1PA6 and older can be found in New World Monkeys, from L1PA15 and older in prosimians and L1MA4 and older in mammals.

In the Brouha *et al.* work, the ~150 LINE1 full-lenght copies potentially able to retrotranspose in the human genome were seen to belong mostly to L1PA1, L1PA2, and L1HS subfamilies (Brouha *et al.*, 2003).

Retrotransposition mechanism

The pool of active LINE1 elements are able to retrotranspose (i.e. mobilize) through a target-site primed reverse transcription mechanism (Cost *et al.*, 2002), as shown in Figure 3. In this process, a LINE1 RNA molecule is transcribed starting from the promoter located in its 5'UTR (Athanikar, Badge and Moran, 2004). Following the polyadenylation step, the new copy of the LINE1 is exported from the nucleus to the cytoplasm (Dai *et al.*, 2012). Here, the two ORFs are translated in proteins allowing the formation of the LINE1 ribonucleoprotein complex (RNP) (Doucet *et al.*, 2010). This complex is assembled by aggregating the LINE1 transcript and the two encoded proteins, most likely due to the ORF1 function. At this stage, the membrane-associated endosomal sorting complex is required for the transport of the LINE1 transcript from the cytoplasm back to the nucleus (Horn *et al.*, 2017). An interesting aspect of this process part is the timing that LINE1 transcript benefits to translocate into the nucleus. While it has been shown in cancer cell models that the LINE1 RNP complex translocates during the mitosis (Mita *et al.*, 2018), in other tissues the association with a specific cell cycle stage is different. Even more intriguing results the comprehension of this mechanism in non-dividing cells such as the neurons (Sanchez-Luque *et al.*, 2019). Once entered into the nucleus, the ORF2 endonuclease activity generates a single-strand DNA break into the preferentially recognized cleavage site (dTn-dAn) (Sultana *et al.*, 2019), thought to produce a particular secondary structure (Cost and Boeke, 1998). The pairing of the LINE1-polyA to this site allows the reverse transcription of the LINE1 RNA transcript by prolonging the genomic 3' hydroxyl group previously created (Doucet *et al.*, 2015). At this point, host DNA repair proteins

are fundamental to integrate into the genome the reverse complement strand of the new LINE1 copy (Flasch *et al.*, 2019). In order to preserve the genomic DNA integrity, the repair proteins are also supposed to be the cause of 5'UTR truncation events in the newly integrated copies (Zingler *et al.*, 2005). The final steps of the retrotransposition mechanism are provided by DNA replication/repair proteins that break the second DNA strand and fix the gap by using the cDNA LINE1 strand as template (Flasch *et al.*, 2019). Ligation and LINE1-RNA degradation are steps likely mediated by host enzymes.

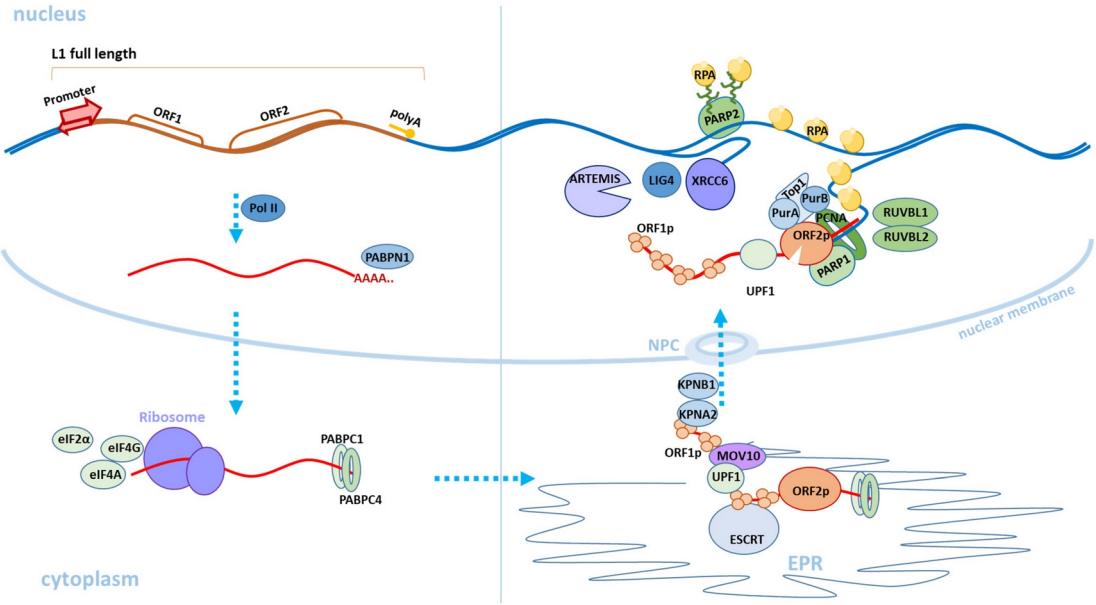

Figure 3: Scheme of the retrotransposition mechanism of LINE1. After the expression of a full-length copy of LINE1 (red line) in the nucleus, the transcript is transported to the cytoplasm to be translated and form the LINE1 RNP. Through the endoplasmic reticulum, the complex is transported back to the nucleus and LINE1 DNA copy is formed and integrated taking advantage of the reverse transcription mechanism (Protasova, Andreeva and Rogaev, 2021).

The overall mechanism generates new copies of LINE1 that, in according to the ORF2 recognition pattern, have been frequently found in A+T-rich DNA sequences typical of non-coding regions (Cost and Boeke, 1998). In addition, this mechanism seldom leads to the transposition of the entire LINE1 sequence (Zingler *et al.*, 2005), explaining why the majority of the elements are composed just of 3' portions shorter than 1 kbp. From the evolutionary perspective, these features made these elements a group of successful genomic parasites (Luning Prak and Kazazian, 2000) because it is simpler

for the genome to accept short insertions that do not occur in coding regions. On the other side, the positive selection of LINE1s inserted into genes occurred and this raises the need to investigate any potential useful role that these TEs may play (Kazazian and Moran, 1998).

1.1.3 The transposable elements can be co-opted during the evolution

Three are the traits proposed which make the TEs sequences more likely to undergo positive selection during the evolution: the ability to spread throughout the genome, the embedding of regulatory sequences and the capability to sequester silencing factors for transcriptional repression (Ali, Han and Liang, 2021). Because of these features, TE sequences have impacted both on the function and on the transcriptional regulation of genes. Indeed, many instances of domestication events involve TEs that were fixed for regulating the genic expression both at the transcriptional and at the post-transcriptional level (Feschotte, 2008) (Figure 4).

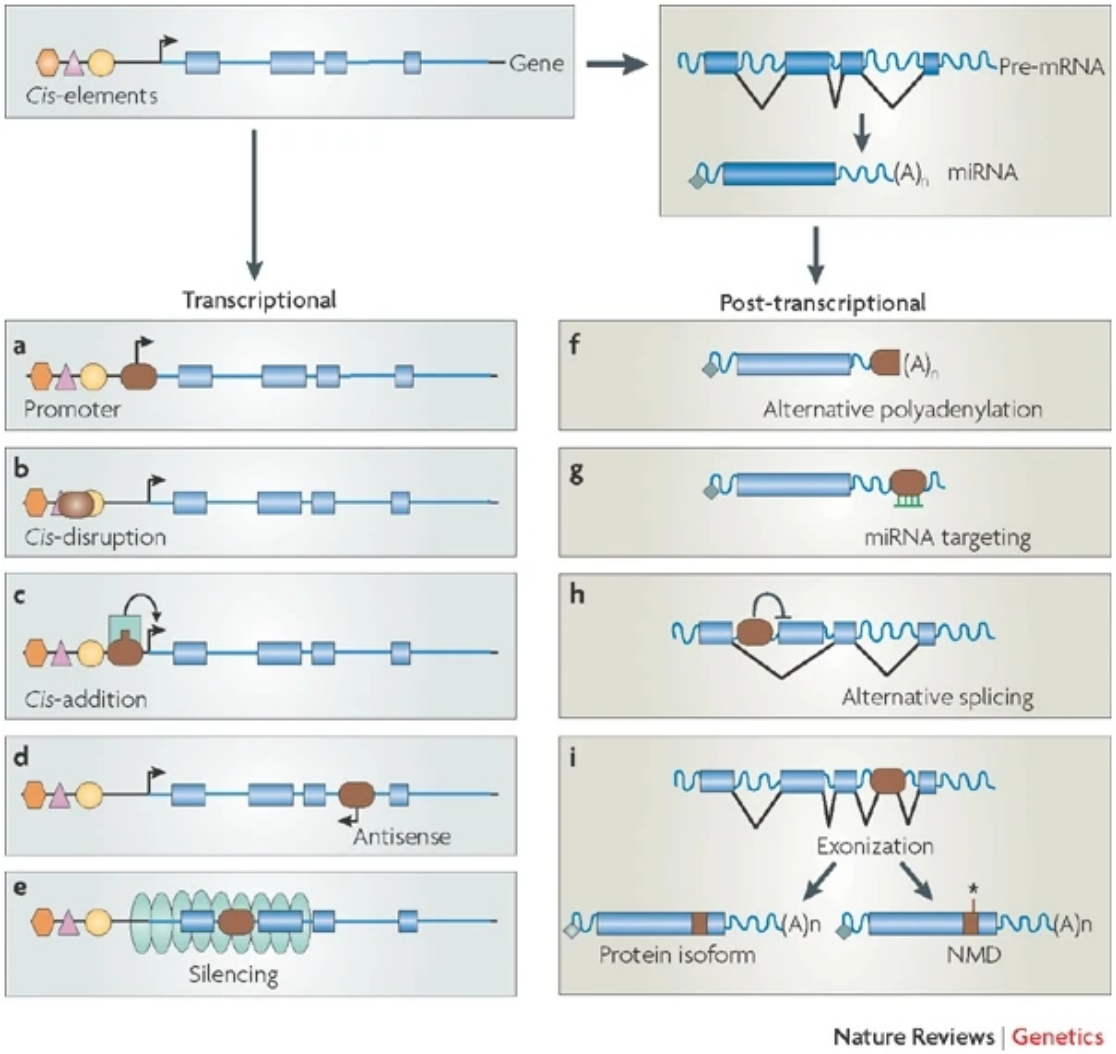

Figure 4: Transposable elements can influence gene expression at different levels. At the transcriptional level, a TE can introduce an alternative transcription start site (**A**), disrupt/introduce cis-regulatory elements (**B, C**), drive antisense transcription (**D**) or potentially sequester silencing factors (**E**). At the post-transcriptional level, a TE in the 3′ UTR can introduce an alternative polyadenylation site (**F**) or a miRNA target-site (**G**). Within an intron, instead, a TE can interfere with the splicing events (**H**) or can be incorporated as an alternative exon (**I**) (Feschotte, 2008).

1.1.3.1 Influence at transcriptional level

In the work by Jordan *et al.* was observed that ~25% of human promoters contain TE sequences (Jordan *et al.*, 2003). Given that the 5' untranslated region of LINE1s contains a sense and an antisense PolII promoter, the transcription of a canonical gene can putatively undergo interference in its expression upon a TE insertion in its *locus* (Speek, 2001) (Figure 4a). An example of an alternative promoter provided by a LINE1 element is found in the Kim *et al.* (Kim and Hahn, 2011) work. A specific L1HS element was observed to be inserted in the CHRM3 gene, a muscarinic cholinergic receptor that mediates many acetylcholine effects in the nervous system. In this study, specific transcripts were shown to derive from the antisense promoter of the LINE1 element, demonstrating the capability for the TE to start the expression of the hosting gene with its own promoter sequence. At the same time, the intergenic insertion of a TE can disrupt or provide *cis*-regulatory elements for the downstream gene. In this way, the TE become directly or indirectly part of the genomic features that regulate the expression of the gene (Figure 4b,c). In according to this, about 25% of DNAseI hypersensitive sites in human CD4+ T cells overlap with annotated TEs (Mariño-Ramírez and Jordan, 2006) suggesting a significant contribution in providing *cis*-regulatory sequences. Moreover, different works showed the co-option of TEs for supplying enhancers (Bejerano *et al.*, 2006) or target-sites for specific transcription factors (Sundaram *et al.*, 2014). Considering that the 5'UTR of L1HS elements contains target-site for the widely expressed YY1 (Yin Yang 1) transcription factor (Becker *et al.*, 1993; Athanikar, Badge and Moran, 2004), the co-option of these elements probably acquired a strong relevance during the evolution. Going on with the ways that co-opted TEs can employ to control the gene expression at transcriptional level, it has been observed that TEs, and especially LINE1s, can drive antisense transcription potentially impairing the transcription of hosting genes (Speek, 2001) (Figure 4d). Another important TE feature is that they experience a strong transcriptional repression inside the cells. Because of this, their insertion may indirectly serve as a nucleation center for the formation of heterochromatin, preventing or reducing the transcription of nearby and hosting genes (Figure 4e). An instance of this was observed in the Liu *et al.* (Liu *et al.*, 2018) work in which the HUman Silencing Hub (HUSH) complex was discovered to interact with young full-length

LINE1s. Reshaping the local chromatin organization with histone repressive marks, the HUSH complex directly causes the decreased expression of neighboring genes.

1.1.3.2 Influence at the post-transcriptional level

If changes at transcriptional level of TE insertions are easier to figure out and identify, the effects that these elements may have at post-transcriptional level are more challenging to detect.

Among these, it has been observed that a TE inside a 3'UTR can potentially introduce an alternative polyadenilation site (Figure 4f) impacting on the post-transcriptional fate of the targeted transcript. In the work of Lee *et al.* (Lee, Ji and Tian, 2008), polyA regions of LINE1 elements were shown to generate polyA sites that are exploited by the hosting genes. Strictly related to TEs inserted into the 3'UTRs, different studies suggested how the evolution has guided the co-option of TEs for providing miRNA target-sites to the host transcript (Figure 4g). About 12% of all TE-derived miRNA target-sites located in the 3'UTR of human genes were observed to be located in LINE1s, as reported in the study of Spengler *et al.* (Spengler, Oakley and Davidson, 2014). In addition to the impact on the mRNA homeostasis, insertion of TEs can also induce change in the final protein product. As shown in Figure 4h, TE insertions may in fact introduce signals that interfere with alternative splicing events. Changing the canonical splicing pattern, phenomena such as intron retention and exon skipping can alter the final structure of the translated protein (Ni *et al.*, 2007). Ultimately, "exonization" events have also been reported (Figure 4i), they regard TEs containing cryptic splice sites that can be included as alternative exons in the mature form of the hosting transcripts. This may lead either to the production of a new protein isoform or to the induction of the nonsense-mediated decay (NMD) pathway if the TE contains a premature stop codon (Kaer *et al.*, 2011).

In light of the different ways for regulating the gene expression, it is crucial to highlight the evolutionary perspective of TEs and LINE1s as a big reservoir of regulatory sequences. As suggested by Feschotte (Feschotte, 2008), they became as such since they were prone to accumulate *de novo* mutations, transforming part of their sequence in regulatory ones. In addition to this, the pre-existence of regulatory

elements within TE sequences might have enhanced the possibilities for them to be co-opted immediately after the insertion. Nevertheless, despite selected beneficial functions, a deregulated TE activity can be detrimental to the cell, and it has been well established its involvement in diseases like cancer and neurodevelopmental disorders (Zhang, Zhang and Yu, 2020).

1.1.4 Diseases caused by dysregulated LINE1 activity

The deregulated activities of TEs, and particularly of LINE1s, have been observed in a wide spectrum of human diseases (Zhang, Zhang and Yu, 2020). Indeed, escaping from silencing mechanisms, LINE1 full-length elements are potentially able to rely on their own transcriptional promoter for their expression. Once the RNA is translated, the activity of the ORF1 and ORF2 proteins can result in the creation of a new genomic insertion. Therefore, as long as the transcriptional reactivation of LINE1s is limited to old mutated non-coding elements, the consequences can be imperceptible. Conversely, if the activated LINE1s are retrotransposition-competent, the outcome can be harmful (Burns, 2020). A direct instance of a detrimental impact that can originate from the deregulated LINE1 activity is provided by the first LINE1 disease-causing insertion discovered by Kazazian in a patient with hemophilia A (Kazazian *et al.*, 1988). In this case, a LINE1 was discovered to be inserted into an exon of the coagulation factor VIII gene, disrupting the gene causative of the genetic disorder. This study is a milestone of human genetics because for the first time demonstrated the existence of autonomous retrotransposition in the human genome.

Understanding how LINE1s can be reactivated is therefore essential to comprehend and prevent diseases caused by their retrotransposition. Recent evidences suggest that both environmental stimuli such as cellular stress (Mourier *et al.*, 2014) and natural cellular processes like senescence (De Cecco *et al.*, 2013) are characterized by destabilized epigenetic mechanisms that trigger the transcriptional activation of TEs (Chuong, Elde and Feschotte, 2017). Therefore, it is not surprising that one of the main driver for the transcriptional reactivation of LINE1 is directly related to changes into the DNA methylation patterns, an epigenetic modification generally associated to transcriptional repression (Lyko, 2018). As shown in Figure 5, hypomethylation states

are related to 4 different macro-groups of diseases: cancer, metabolic pathologies, neurological disorders, and autoimmunity.

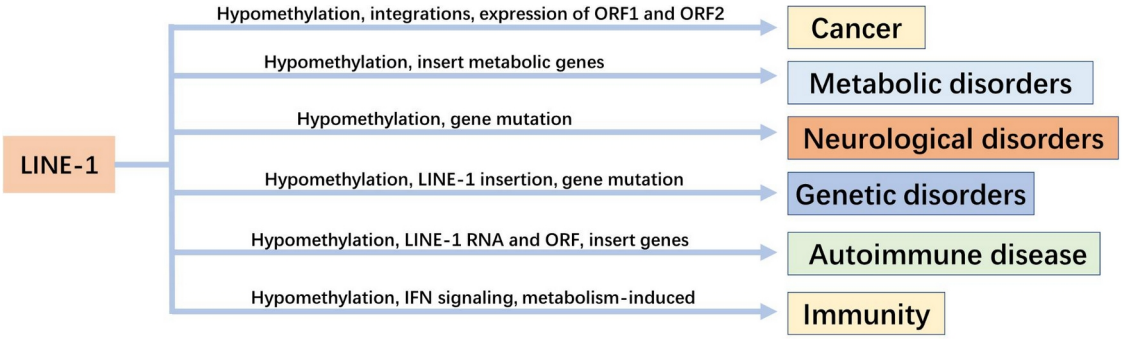

Figure 5: LINE1 dysregulation and diseases. Summary of different types of LINE1 dysregulation and the related caused disease. Cancer, metabolic pathologies, neurological disorders, and autoimmunity can be potentially caused by the hypomethylation state of LINE1 elements (Zhang, Zhang and Yu, 2020).

Cancer

A common feature of human malignancies is the DNA hypomethylation of LINE1 promoter regions (Wilson, Power and Molloy, 2007). Lacking one of the most important repressive layer of the LINE1 transcription, many cancers are characterized by a strong overexpression of this TE family. Following the LINE1 life-cycle, the aberrant transcription of the TEs is coherently associated to the increased translation of ORF1 and ORF2 proteins. On one hand, ORF1 levels represent a hallmark of many cancers and it is used as a diagnostic marker (Ardeljan *et al.*, 2017). In contrast, the ORF2 application in the diagnostic field is limited since its level is more difficult to detect since it is translated at a lower level (Ardeljan *et al.*, 2019). Nonetheless, its presence was observed to exacerbate the DNA damage because of its endonuclease activity (Kines *et al.*, 2014). The final outcome of the higher LINE1 expression is the rapid increase in their number of copies within the genome (Rodriguez-Martin *et al.*, 2020). Many insertions appear to be passengers and they result shared by cellular sub-populations of the tumoral mass. Nevertheless, genomic hot-spots (Burns, 2017) have been discovered and they might potentially confer selective advantages to the pre-neoplastic cells. Genic rearrangements such as oncogene amplification and oncosuppressor deletions are in fact phenomena that can result from LINE1 integration events (Rodriguez-Martin *et al.*, 2020).

Metabolic disorders

In metabolic disorders, the LINE1 epigenetic status is observed to play a role as potential disease indicator. Indeed, it has been detected an association between LINE1 DNA methylation levels and the type 2 diabetes mellitus (María Martín-Núñez *et al.*, 2014) as well as LDL/HDL cholesterol levels (Pearce *et al.*, 2012). In addition to this, changes into the metabolic state are considered a part of the reprogramming of tumor cells (Liberti and Locasale, 2016). The involvement of LINE1 in metabolism appears to have a potential significant impact on tumors. In support of this consideration, LINE1 insertion in the FGGY (FGGY Carbohydrate Kinase Domain Containing) gene has been detected in lung squamous call carcinoma. Since this gene encodes for a protein that phosphorylates carbohydrates, disrupting mutations were demonstrated to change the metabolism, causing a poor prognosis in patients (R. Zhang *et al.*, 2019).

Neurological disorders

Neurological disorders represent another category of diseases in which LINE1s were observed to have a potential pathogenic role. In Rett syndrome, mutations in the methyl CpG binding protein 2 (MeCP2) leads to derepression and overexpression of LINE1s (Muotri *et al.*, 2010). In aging processes and in frontotemporal lobe degeneration, this family of TEs was found to be overexpressed (Li *et al.*, 2012). An increased number of LINE1 copy number in the genome possibly determined by retrotransposition following hypomethylation and reactivation of these elements was also observed to be a common feature of different psychiatric disorders such as schizophrenia and bipolar disorder (Li *et al.*, 2018). Considering the LINE1 activity during neuronal differentiation (Coufal *et al.*, 2009), the potential consequences of its deregulated transcription in the brain make this TEs family a potential cause of neurodevelopmental disorders.

Autoimmune disorders

The last group of diseases caused by dysregulated LINE1 activity is composed of autoimmune disorders. Pathologies like lupus erythematosus, Sjogren's syndrome and psoriasis are characterized by both hypomethylation of LINE1 loci as well as overexpression of this family of TEs (Yooyongsatit *et al.*, 2015; Mavragani *et al.*,

2016). In these diseases the presence of LINE1 DNA retrotranscribed in the cytoplasm seems to activate the innate immune response that is at the basis of the autoimmunity. In this field, the Aicardi-Goutieres syndrome (AGS) is a dramatic disease characterized by severe neurological impairment (Crow and Rehwinkel, 2009). The autoimmunity of this disorder appears to be caused by the inability to remove LINE1 retrotranscribed DNA molecules generated in the cytoplasm during the LINE1 life-cycle (Thomas *et al.*, 2017). The activation of the interferon signalling pathway in response to the accumulation of extrachromosomal DNA triggers the inflammatory autoimmune response. The reason why LINE1s carryout retrotranscription also in the cytoplasm remains still to be clarified.

1.2 Mechanisms repressing LINE1

Considering the harmful effect that the LINE1 dysregulation can have, it is essential for the cell to control the transcript levels of this TE family. For this reason, LINE1 RNA abundance is repressed at many stages of the retrotransposition process through the employment of several defense cellular mechanisms both acting at DNA as well as at RNA level.

1.2.1 Control of LINE1 transcript levels at DNA-level

Epigenetic modifications are the most used way the cell uses to repress the activity of retrotransposons at DNA-level. They are processes that alter the gene expression regulation without modifying the DNA sequence and include histone modifications and DNA methylation (Gujar, Weisenberger and Liang, 2019). Modifying the epigenome, the cell is able to control the nucleosome occupancy of genomic loci promoting or inhibiting the accessibility of PolII and transcription factors to the DNA sequence (Venkatesh and Workman, 2015). These modifications directly affect the gene expression.

The mechanisms that act at DNA-level for controlling LINE1 transcript levels, exert their function by decreasing the chromatin availability. Among the key proteins that act by promoting the formation of heterochromatin at LINE1 genomic loci (Van Meter *et al.*, 2014), there is SIRT6, a component of the Sirtuin genes family. It inhibits LINE1s transcription both by interacting with MeCP2 and by modifying the factor KAP1 (TRIM28). In the first mechanism, SIRT6 interacts with MeCP2 and activates the methylation process of the LINE1 promoter regions (Muotri *et al.*, 2010). In the second one, the ribosylation of KAP1 mediated by SIRT6 promotes the formation of heterochromatin in ancient LINE1 loci (Castro-Diaz *et al.*, 2014). On the other hand, genomic loci of young LINE1 copies were observed to experience the activity of the HUSH complex for the establishment of the heterochromatin state. This epigenetic complex is composed by the proteins TASOR, MPP8 and PPHLN1 (periphilin1). In the study of Robbez-Masson *et al.* (Robbez-Masson *et al.*, 2018), the TASOR component was shown to bind and repress evolutionary young LINE1 elements probably taking advantage of histone repressive marks. By reorganizing the local

chromatin state, the HUSH complex is moreover shown to affect also the expression of genes enriched in young LINE1s.

In addition to these proteins, cell cycle factors are also shown to play a role into the regulation of LINE1. An instance of this is p53, a transcription factor with oncosuppressor activities. Due to its role in response to various cellular stresses, it has been defined the "guardian of the genome". When p53 is activated as a result of DNA damage, cell cycle progression is arrested. Based on the extent and type of DNA damage, repair mechanisms or apoptosis are induced as final response to the stress stimuli (Hafner *et al.*, 2019). It has been observed that ~50% of the human cancers experience a LINE1 derepression (Rodić *et al.*, 2014). In response to the resulting DNA damage created by increased retrotransposition events, p53 is activated to first arrest the cell cycle and then to induce apoptosis (Haoudi et al., 2004). The cell cycle arrest results thus fundamental considering that LINE1s enter the nucleus during the M phase and retrotranspose during the S phase (Mita *et al.*, 2018). In addition to this, recently it has been discovered an additional regulatory layer of LINE1 transcripts level mediated by this cell cycle factor. Indeed, a p53 target-site in the 5'UTR promoter region of young LINE1 elements has been demonstrated to allow the establishment of repressive histone marks aiding the cell in the control of LINE1 activity at DNA-level (Tiwari *et al.*, 2020). As appreciated from the previous mechanisms, the heterochromatin state of the 5'UTR promoter region results pivotal to control the transcription of the LINE1s. In according to this, the binding of the transcription factor YY1 is an exciting discovery for its capability to act both as activator and as repressor of the gene expression. YY1 is a transcription factor widely expressed in mammalian cells that recognize and binds a small (10 nt) DNA sequence using its zinc-finger domain. Post-translational modifications of YY1 as well as the interplay with different cofactors and chromatin modifiers have been proposed to explain the capability of the protein to act as a positive or negative transcriptional modulator (Verheul *et al.*, 2020). In the context of LINE1 control, YY1 was observed both to be required for the initiation of LINE1 transcription (Athanikar, Badge and Moran, 2004) and to facilitate the methylation of young LINE1 promoters for transcriptional repression (Sanchez-Luque *et al.*, 2019). This latter type of epigenetic

modification represents the most powerful tool used by the cell for repressing the LINE1 expression.

1.2.1.1 The DNA methyltransferases family

The most common DNA methylation occurs in the cytosine that are in the 5'-CpG-3' conformation (Breiling and Lyko, 2015). Considering the effects on the gene expression regulation, it is commonly accepted that the DNA methylation in promoters both affects the DNA binding by transcription factors and help in the recruitment of repressive factors (Tate and Bird, 1993), resulting in transcriptional repression. On the other side, the DNA methylation of gene bodies is associated with active transcription since it can stabilize the transcriptional elongation by inhibiting events such as spurious transcription (Neri *et al.*, 2017).

Since it controls a variety of different cellular processes like gene imprinting (Henckel and Arnaud, 2010) and X-chromosome inactivation (Riggs, 1975), the DNA methylation must be finely regulated and maintained during the different processes carried out in the entire life of an organism. Indeed, during the development, waves of DNA demethylation and metylation occur (Smallwood and Kelsey, 2012) providing temporal windows for the transcription of TEs. On the other hand, in somatic dividing cells the DNA methylation patterns must be conserved and stably inherited for contributing to the peculiar transcriptional profile of the cell type and for protecting the cells from TEs activities.

The enzymes required for this function are members of the DNA methyltransferases (DNMT) family (Lyko, 2018). This group includes DNMT1, which is involved into the maintenance of DNA methylation after DNA replication, and DNMT3a/DNMT3b which are instead selected for the establishment of *de-novo* DNA methylation in unmethylated genomic loci (Xie *et al.*, 1999) (Figure 6). Nevertheless, the DNMT functions are not exclusive. Indeed, on one hand, DNMT3a and DNMT3b were shown to assist DNMT1 in maintaining DNA mehtylation at repetitive elements (Liang *et al.*, 2002), while on the other hand, DNMT1 was observed to have a role into the *de-novo* deposition of methylation marks (Fatemi *et al.*, 2002).

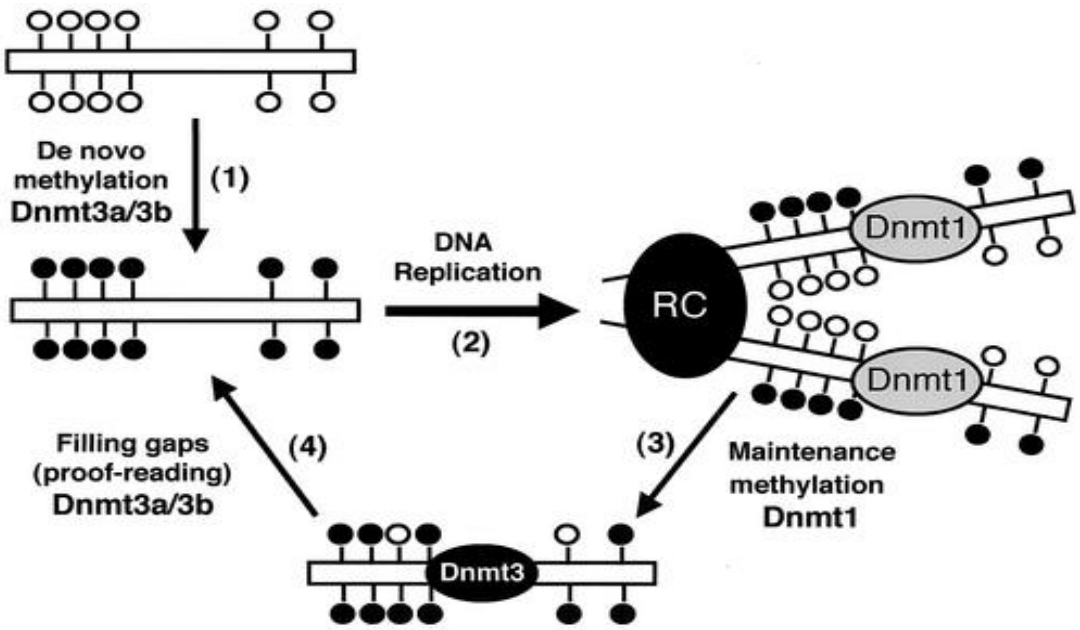

Figure 6: The DNA methylation reactions. **1)** Dnmt3a and Dnmt3b establish new DNA methylation patterns by de-novo methylation of CpG. **2)** Upon DNA replication, the new synthesized DNA becomes hemimethylated. **3)** Dnmt1 restores the full methylation (Chen *et al., 2003)*.

The function of DNMT1 will be evaluated at a higher depth in this thesis, since the effects of its KO will be discussed in the following paragraphs. Particularly, DNMT1 is in charge of the establishment of the DNA methylation profiles from the old and methylated parental to the novel and unmethylated daughter DNA strand during the cell division. The enzyme is composed by a C-terminal catalytic domain and an N-terminal regulatory domain. The first, consists of the methyltransferase subdomain and the target-recognition domain (TRD) capable to recognize hemimethylated cytosines (Zhang *et al.*, 2015). The N-terminal regulatory domain, instead, is composed of the tandem bromo-adjacent homology (BAH1/2) domain that inhibits *de-novo* methylation for maintaining high fidelity of the enzyme towards the hemimethylated regions of the DNA (Song *et al.*, 2011).

In agreement to its function, the lack of DNMT1 activity leads to a global genome-wide DNA demethylation. From past literature, it has been observed that mouse embryonic stem cells with DNMT1-deficient activity show a normal phenotype until differentiation. The massive death of cells occurring at differentiation is in line with the dramatic methylation changes that take place during these cellular processes (Lei

et al., 1996). Thus, because of its essential function, known DNMT1 mutations in human are both rare and almost always heterozygous. Mutations have been described in hereditary sensory and autonomic neuropathy type 1E (Klein *et al.*, 2011) and in cancer (Baylin and Jones, 2011).

Despite the difficulties of developing a human cellular model for studying the lack of DNMT1 activity, in the work by Jonsson *et al.* (Jönsson *et al.*, 2019) the authors were able to homozygously disrupt the DNMT1 gene in human neuronal progenitor cells (hNPCs) using the CRISPR-Cas9 technology. Upon the mutation, the global DNA demethylation is associated with the transcriptional activation of LINE1 elements, providing me an excellent model that I will use for studying LINE1 activities.

1.2.1.2 ATRX: the enzyme for remodelling nucleosomes

In my thesis, in addition to the DNMT1 model, another cellular context composed of cells carrying the KO of the Alpha Thalassemia/mental Retardation syndrome X-linked (ATRX) gene will also be examined. ATRX is a protein which mediates the transcriptional regulation through remodelling of nucleosomes in many biological processes (Picketts *et al.*, 1996). It is an ubiquitously expressed nuclear protein, with high levels in the fetal brain suggesting an important role during the development of this organ (Gecz *et al.*, 1994). Indeed, ATRX mutations are at the basis of autism, intellectual disability (Gibbons *et al.*, 1995) and a variety of cancers, including neuroblastoma and glioma (Louis *et al.*, 2016).

From the structural point of view, important ATRX domains are the helicase/ATPase domain, the DAXX-binding motif and a nuclear localization signal (Hoelper *et al.*, 2017). In addition to these, two zinc-finger motifs suggest a role both in the binding of the DNA (Cardoso *et al.*, 2000) and in the chromatin-mediated transcriptional control (ADD domain) (Gibbons *et al.*, 1997). These functions are dependent by its interaction with the chaperone protein DAXX. Once the chromatin remodelling complex ATRX-DAXX is created, the ADD domain recognizes the H3K9me3-enriched chromatin regions and allows the deposition of the histone variant H3.3 (Goldberg *et al.*, 2010). The interested genomic regions are thus transcriptionally silenced (Voon *et al.*, 2015).

The lack of ATRX activity was observed to be associated to a higher chromatin accessibility both in repetitive and non-repetitive regions (Liang *et al.*, 2020). A possible role for ATRX into the organization and the maintenance of the repressive state in pericentromeric chromatin was also suggested (McDowell *et al.*, 1999). In this case, histone tails modifications seems to be crucial for the recruitment of the heterochromatin protein 1 (HP1) that in turn recruits the histone methyltransferases for the repressive trimethylation of H3K9 (Marano *et al.*, 2019). Similarly to pericentromeric regions, the telomeres are characterized by heterochromatin that is fundamental for the chromosomal stability (García-Cao *et al.*, 2004). In these regions, the lack of ATRX activity causes a decreased density of nucleosomes (Li *et al.*, 2019).

Regarding its involvement into the repression of transposable elements transcription at the DNA-level, in mouse ESCs the deletions of the ATRX and DAXX genes resulted in a lower level of histone variant H3.3 and H3K9me3 at ERV loci (Elsässer *et al.*, 2015). In addition, in human cancer cell lines, an increased genomic accessibility of retrotransposons was appreciated with ATAC-seq experiments upon the KO of ATRX (Liang *et al.*, 2020). Because of the importance of ATRX in chromatin remodelling to avoid aberrant activation of retransposons transcription, I decided to investigate RNA-seq data derived from the work of Denault *et al.* (Deneault *et al.*, 2018). In this study, the RNA was sequenced from cells characterized by the KO of *ATRX* gene first in reprogrammed human induced pluripotent stem cells (iPSC) and then in iPSC differentiated in neuronal cells.

1.2.2 Control of LINE1 transcript levels at RNA-level

In the event that LINE1s escape the repressive mechanisms at DNA-level, other pathways are ready to act at post-transcriptional level for regulating the RNA abundance of these TEs.

Antiviral factors represent a group of proteins capable to control nucleic acid amounts. It has been shown that they are able also to control levels of endogenous nucleic acids such as LINE1 transcripts, and their mutations are at the basis of autoimmune diseases like AGS (Crow *et al.*, 2015). In the category of factors acting on the LINE1s, there are proteins such as the ribonuclease L that is shown to cleave the LINE1 RNA (Zhang *et al.*, 2014) and the RNase H2 that degrades the RNA-DNA hybrids created during the LINE1 life-cycle (Choi, Hwang and Ahn, 2018). Among the proteins that are able to edit the RNA, a well studied group is represented by the APOBEC3 family. Their main activity is to catalyze the deamination of LINE1 cDNA transcripts, converting cytosine to uracil. As a result, the edited RNA copy undergoes the degradation through the activity of the repair mechanism which excises the uracils. The generated nicks induce the final LINE1 degradation (Feng *et al.*, 2017).

In embryonic germ cells, LINE1s restriction involves the Piwi-interacting RNA (piRNA)-signaling pathway (Pezic *et al.*, 2014). piRNAs are small non-coding RNAs that bind the PIWI proteins, a group of proteins that belong to the Argonaute family (AGO) (Ross, Weiner and Lin, 2014). Similarly to the RNA interfering pathways, the complex formed by piRNA and PIWI is shown to bind complementary LINE1 RNA copies, inducing the transcript degradation. In the “ping-pong” cycle (De Fazio *et al.*, 2011), a transposon-rich piRNA clusters is able to produce a variety of piRNAs. Once the piRNA-PIWI complex encounter the LINE1 target, the complex cleaves the mRNA, generating other piRNAs capable to target LINE1 transcripts in a self-amplifying loop.

If in embryonic germ cells piRNAs build the RNA interfering pathways acting at RNA-level to inhibit the LINE1 transcripts, in somatic cells this role is achieved by miRNAs. In the last few years, more knowledge became available on the miRNA-based mechanisms used to inhibit LINE1 elements.

1.2.2.1 MicroRNAs

Discovered in *C. elegans*, miRNAs are a class of small non-coding RNAs that have a fundamental role in gene expression control (Lee and Ambros, 2001). They are small (~22 nucleotides) RNA molecules transcribed from intragenic regions (especially introns) or independently by their own promoter (Kim and Kim, 2007; de Rie *et al.*, 2017). Frequently, they can be organized in a single long polycistronic transcript from which different miRNA are differently processed.

Following the transcription stage, the immature form of the miRNA is represented by the primary-miRNA (pri-miRNA), a molecule which can be longer than 1 kbp characterized by a stem-loop structure (Lee *et al.*, 2002). The pathways that result in the creation of the mature and functional form can be canonical or not, as depicted in Figure 7.

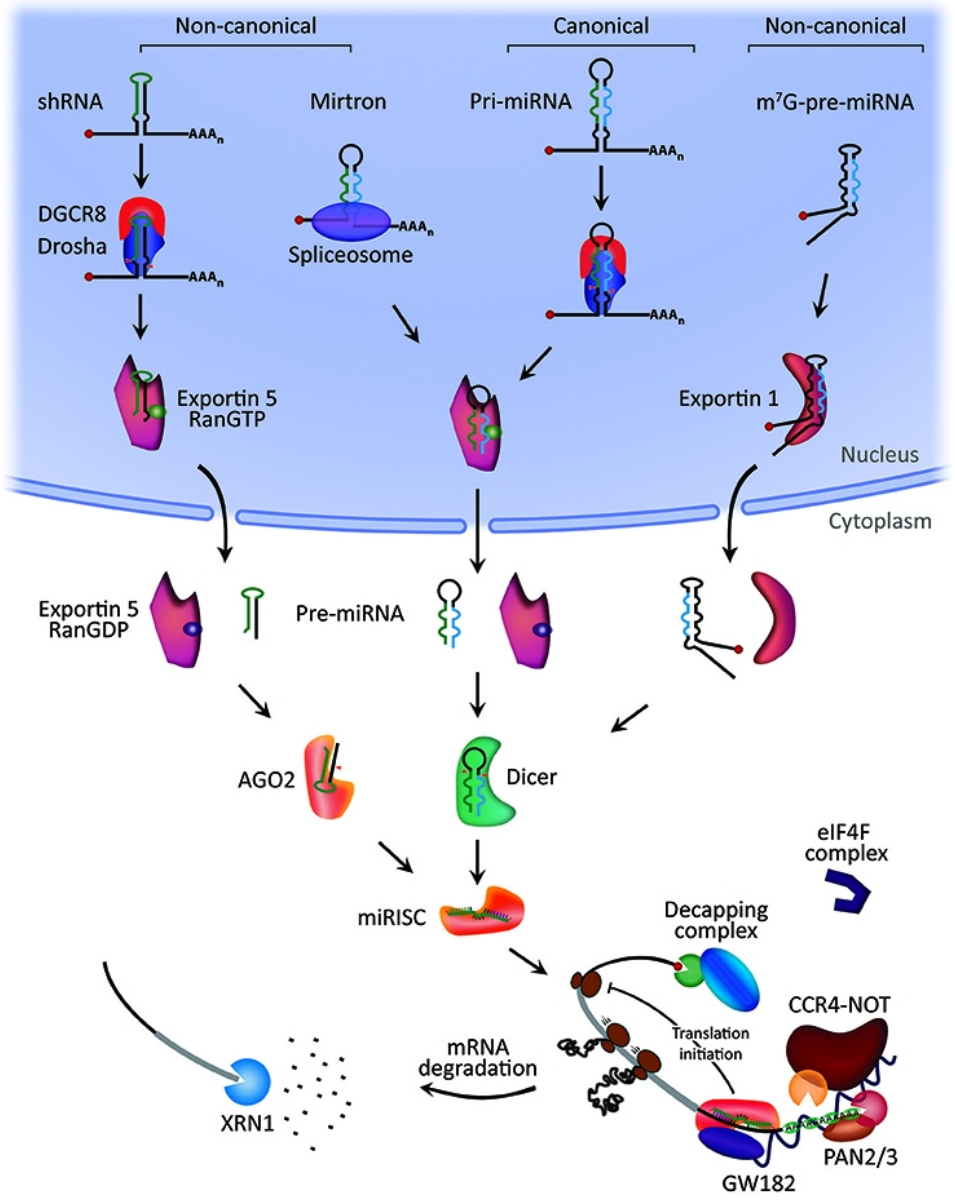

Figure 7: MicroRNA biogenesis. Canonical miRNA biogenesis pathway takes advantage of the Microprocessor complex in the nucleus. After exporting into the cytoplasm it is processed by Dicer to produce the mature miRNA duplex. In the non-canonical pathways, the miRNA generation can be Microprocessor- or Dicer-independent (O*'Brien et al., 2018)*.

The canonical pathway

In the canonical pathway, the Microprocessor complex, composed of the RNA-binding protein DiGeorge Syndrome Critical Region 8 (DGCR8) and the ribonuclease Drosha (Denli *et al.*, 2004), starts to process the pri-miRNA. While with the first component the complex is able to recognize RNA motifs, with the second one it cleaves the pri-

miRNA at the base. The formation of a 2-3 nt overhang in the RNA molecule is a feature of the precursor miRNA (pre-miRNA) form, a molecule still characterized by an hairpin secondary structure but with a strongly decreased length (~65 nucleotides) (Han *et al.*, 2004). After this step, the complex formed by the Exportin 5 and the RanGTP catalyzes the transport of the pre-miRNA into the cytoplasm (Okada *et al.*, 2009), exposing the RNA molecule to the activity of the endonuclease Dicer. Cleaving the terminal stem-loop, this complex enables the formation of the mature miRNA duplex (Zhang *et al.*, 2004). At this stage, the mature form has an average length of 22 nucleotides and it is composed of the 5p and 3p strands that respectively originate from the 5'/3' end of the RNA duplex.

Both strands are thus loaded into the AGO protein with a thermodynamic preference for one of them, the "guide" strand which is believed to be more biologically active with respect to the "passenger" strand, generally rapidly degraded by the cell creating a strong strand bias (Khvorova, Reynolds and Jayasena, 2003). Once the guide strand is loaded into the AGO protein, the mature miRNA-induced silencing complex (RISC) (Kawamata and Tomari, 2010) is ready to exploit its function of identifying miRNA target. Usually, the target mRNAs host a miRNA target-site located in the 3'UTR that is complementary to the "seed" region (nucleotides 2-8) of the miRNA guide (Ellwanger *et al.*, 2011). To recognize the target, the seed of the guide strand is pre-arranged in an A-form helix conformation that allows the scanning of the target-sites on the engaged transcript (Schirle, Sheu-Gruttadauria and MacRae, 2014). It is believed that, after the miRNA-mRNA interaction, the degree of complementarity determines the fate of the target transcript. A fully complementary interaction activates the AGO endonuclease function, leading to the degradation of the transcript (Jo *et al.*, 2015). On the other side, a not fully complementary interaction might lead to an initial translational inhibition driven by the RISC complex that interferes with the eukaryotic translation-initiation factor 4 (eIF4F) complex (Huntzinger and Izaurralde, 2011). The recruitment of effector proteins such as poly(A)-deadenylase complexes enables the deadenylation, then the RNA decapping and ultimately the RNA decay (Braun *et al.*, 2012).

The AGO genes represent a family virtually conserved in all multicellular organisms (Swarts *et al.*, 2014) because of their crucial role in gene silencing. The human genome encodes four types of AGO proteins (AGO1-4) that are highly conserved since they share ~85% of the sequence identity. They specifically consist of the following four domains:

- the N-terminal domain (N) which has two motifs required for the full catalytic activity (Hauptmann *et al.*, 2013);
- the PIWI/Argonaute/Zwille (PAZ) domain which anchors the 3' end of the guide strand (Lingel *et al.*, 2003);
- the MID domain that binds the 5' end of the guide strand (Boland *et al.*, 2010);
- the P-element-induced whimpy tested (PIWI) domain which is essential for the slicing activity (Parker, Roe and Barford, 2004);

While the PAZ and the MID domains are conserved among the four AGO proteins, the other two domains present differences that affect the catalytic efficiency. Regarding the N domain, AGO2 contains the two motifs fundamental for the catalysis, AGO1 has only one, whereas AGO3 and AGO4 have none. In addition to these differences, AGO2 and AGO3 have a fully functional PIWI domain while AGO1 and AGO4 lack key catalytic residues (Müller, Fazi and Ciaudo, 2020). As a result of these variations all AGO proteins may elicit the translational inhibition but only AGO2 is able to slice perfectly matched target transcripts. (Huntzinger and Izaurralde, 2011).

The non-canonical pathway

In the non-canonical miRNA biogenesis pathway, the miRNA generation processes can be divided into Microprocessor- and Dicer-independent.

In the Microprocessor-independent pathway, the pri-miRNA molecule is not recognized by the Microprocessor complex. For this reason, it skips this initial processing step and passes directly in the cytoplasm. Here, because of the similarities to a pre-miRNA, it is recognized and processed as a Dicer substrate, thus following the canonical processing. Instances of miRNAs that are not recognized by the Microprocessor complex are mirtrons and pre-miRNA 7-methylguanosine capped. Mirtrons are miRNAs generated from spliced introns which are reshaped into a short stem-loop form very close to the pre-miRNA structure (Ruby, Jan and Bartel, 2007).

On the other hand, pre-miRNA 7-methylguanosine capped are directly exported in the cytoplasm via Exportin 1 without the possibility to be processed by the Microprocessor complex (Xie *et al.*, 2013).

The non-canonical Dicer-independent pathway instead occurs on the endogenous short hairpin RNAs (shRNA), an artificial RNA molecule with a stem-loop structure that can be used to inhibit the gene expression. In this case, the miRNAs produced by these molecules are processed by the Microprocessor into a form that is insufficiently long to serve as a substrate for Dicer. Because of this, the maturation is completed by a trimming process mediated by the AGO2 protein (Yang *et al.*, 2010).

Once the mature form of the miRNA is generated, the mechanism of action of the RISC complex is the same between the canonical and the non-canonical pathways thus resulting in mRNA degradation or translational inhibition. Considering the number of known miRNAs encoded by the human genome (2656 catalogued in the latest release of the miRNA database miRBase (Griffiths-Jones *et al.*, 2006)) and that more than the 60% of human protein-coding genes contain at least one miRNA target-site, it is not surprising that probably miRNAs control the vast majority of protein-coding transcripts (Friedman *et al.*, 2009). miRNAs dysregulation have therefore been observed to be associated to many human diseases (Im and Kenny, 2012).

1.2.2.2 miRNAs controlling LINE1: the let-7 and miR128 cases

It is interesting how little is understood about the evolutionary history of miRNAs. In the study of Piriyapongsa *et al.* (Piriyapongsa, Mariño-Ramírez and Jordan, 2007) in 2007, the authors tested the possibility that TEs might have contributed to the creation of genes encoding miRNAs. Their idea was based on the evaluation of TE-specific features. Since they are ubiquitous, abundant and able to change their genomic location, each TE can potentially be a good candidate for spreading miRNA sequences within the genome during the evolution. For evaluating the capability of TEs to create miRNA genes, they compared the genomic location of TEs and 462 miRNAs experimentally annotated in human at that time. With this analysis they observed that 68 miRNAs share sequences with TEs and 47 of these have more than 95% of mature regulatory miRNA sequence covered by TEs. The TEs contributing to the miRNA sequences belonged to the four classes of human TEs: LINE, SINE, LTR and DNA transposons. In addition to this, the capability of these TE-derived miRNAs to regulate the gene expression was demonstrated to affect the metabolism and the transcriptional regulation. In the final part of the study, they postulated that miRNAs, like the counterpart siRNAs (Vastenhouw and Plasterk, 2004), might have evolved to silence TEs.

Despite these observations, the scientific community has discovered only in 2015 the first miRNA able to regulate the transcript levels of LINE1 in human as part of a repressive mechanism acting at RNA-level (Hamdorf *et al.*, 2015). In this work, the authors hypothesized that miRNAs might act to protect non-germ cells from the LINE1 activity. This function might take the place of the piRNA activities, which are mostly believed to be restricted to the germ cells. In order to demonstrate this miRNA role, they analyzed cell lines prone to overexpress LINE1 transcripts. The creation of libraries, in which short hairpin RNAs were used to neutralize specific endogenous miRNAs, allowed them to discover the miR-128 as a key player into the LINE1 regulation. In particular, they demonstrated that repressing the miR-128 there was an enhanced LINE1 retrotransposition within the genome. Moreover, this activity was dependent from the binding of the miRNA on a putative target-site located in the ORF2 of LINE1 transcripts. The final LINE1-miR128 interaction was tested by isolating the AGO complexes loaded with the miRNAs and their targets. Observing a

higher level of AGO-bound LINE1 RNA in cells overexpressing the miR-128, they presented a strong evidence that the miR-128 binds to LINE1 RNA. In their proposed model, miRNAs like the miR-128 have adopted a piRNA-like (Aravin *et al.*, 2007) role in somatic cells, working as guardians of genomic stability for the cells. Even though their results suggest the RNA degradation as the main mechanism for the LINE1 repression, they do not exclude the possibility that the miR-128 might exert its function inhibiting the translation processes.

Only recently Tristàn-Ramos and colleagues (Tristán-Ramos *et al.*, 2020) added support to the idea that miRNAs might control the activity of LINE1 also inhibiting their translation. In their work, they started hypothesizing that some miRNAs might control LINE1 retrotransposition and that their misregulation in tumors increases mobilization events. To investigate this, they mainly analyzed whole genome sequencing (WGS) and miRNA expression data of human lung tumor samples matched with the normal counterpart. The analyses revealed that samples carrying tumor-specific LINE1 insertions are characterized by a strong downregulation of miRNAs belonging to the let-7 family. The activity to repress the retrotransposition events was confirmed by retrotransposition assays with increased or inhibited let-7 expression. After the identification of a putative let-7 target-site in ORF2 region of LINE1, they demonstrated the physical interaction between AGO2, let-7 and LINE1 transcripts. Since no correlation between the expression levels of let-7 and L1HS was found, they tried to understand if the repression activity was at the translational level. Observing a negative correlation between let-7 levels and ORF2p protein levels, they finally demonstrated that this family of miRNAs exerts its functions impairing the translation of ORF2p, without affecting the mRNA stability (model in Figure 8).

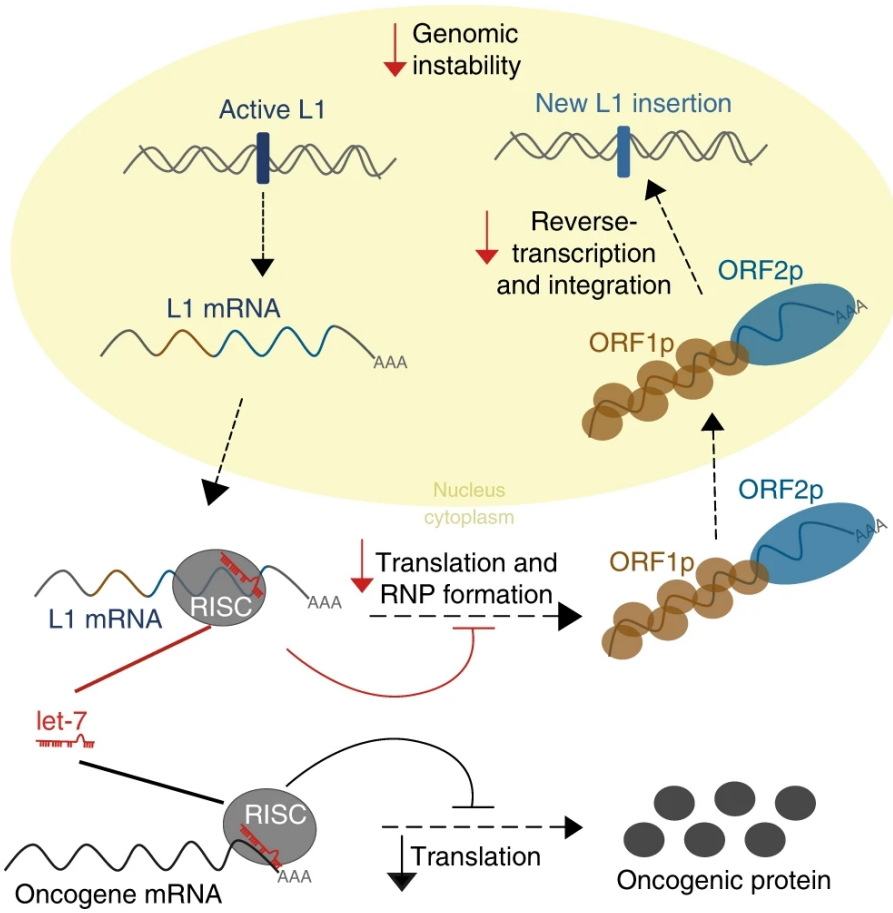

Figure 8: Model for controlling LINE-1 transcripts by let-7. Once LINE-1 RNA is transcribed and exported into the cytoplasm, let-7 binds and guides the RISC complex to the LINE1 mRNA. This binding leads to the translation inhibition of ORF2 impacting the formation of the RNP (Tristán-*Ramos et al., 2020)*.

1.3 Competitive endogenous RNAs

In addition to the conventional miRNA-target pathway, several studies have shown the existence of a second layer of complexity in post-transcriptional gene regulatory networks (Subramanian, 2014). In this mechanism, both coding and long non-coding RNA transcripts are able to regulate the gene expression "in trans" acting as competitive endogenous RNA (ceRNA).

In the simplest scenario there are two transcripts (ceRNAs) which are targeted by the same miRNA because they both contain the same target-sites. If the expression levels of one transcript increase, the generated transcripts are able to bind and sequester more miRNAs from the second transcript. Since the second transcript is no longer controlled by the same miRNA amounts, the result is an upregulation, as shown in Figure 9.

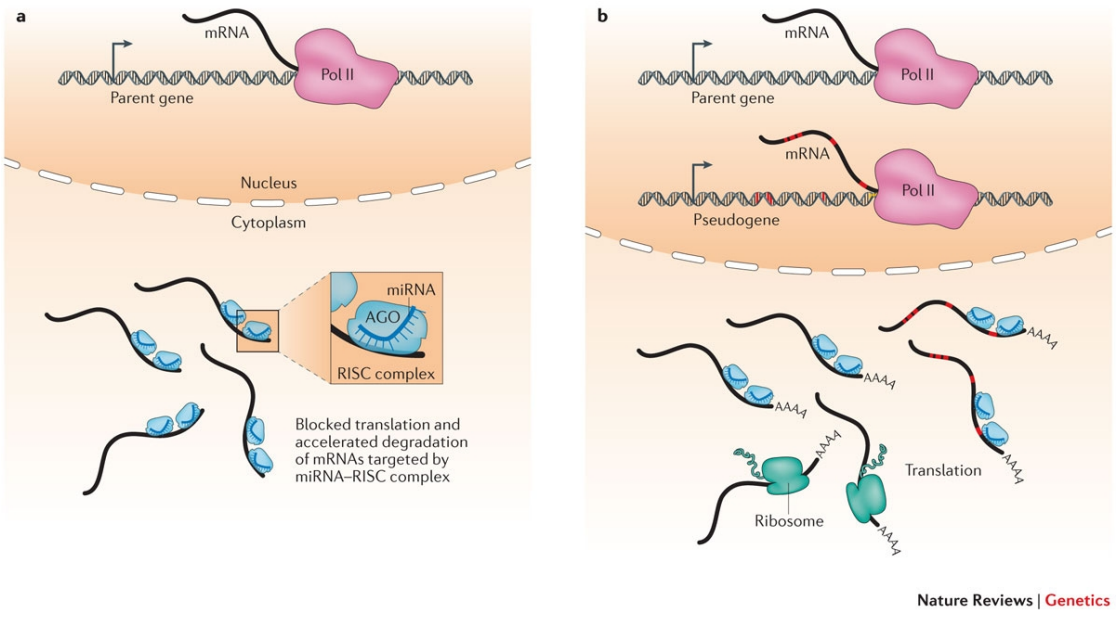

Figure 9: The Basis of the ceRNA mechanism. A. While the ceRNA like a pseudogene remains silenced, the parent mRNA is transcribed and targeted by miRNAs. B. When the pseudogene with competing miRNA target-sites (red) is transcribed, it competes for miRNA increasing the parent gene expression. Source: (Thomson and Dinger, 2016).

From the historical point of view, Ebert *et al.* (Ebert, Neilson and Sharp, 2007) were the firsts that used the capability of RNA to compete for the binding of miRNAs. In their technique, the synthetic microRNA "sponges" were transcripts containing multiple target-sites for a microRNA of interest. The transfection of these RNA molecules, was able to derepress the canonical miRNA targets genes that resulted upregulated. After this work, endogenous ceRNA acting as sponge were also observed. In plants, the miR-399 was the first miRNA observed to undergo the activity of a competitive endogenous RNA (Franco-Zorrilla *et al.*, 2007). In this study, the non-coding RNA Induced by Phosphate Starvation 1 (IPS1) sequesters the miRNA from the canonical target PHO2 (Phosphate Overaccumulator 2) directly causing its upregulation. The mechanism was later observed in human cells as well. In the hepatocellular carcinoma the non-coding RNA High Upregulated in Liver Cancer (HULC) sequesters the miR-372 leading to the PRKABC derepression, a protein fundamental for cancer progression (Wang *et al.*, 2010).

In spite of these findings, only in the 2011 were set the basis for the definition of the ceRNA hypothesis. With the work of Poliseno *et al.* (Poliseno *et al.*, 2010), the authors were able to obtain an experimental evidence of a competitive mechanism involving a protein-coding RNA and its pseudogene. Specifically, they demonstrated that the tumor suppressor gene PTEN and its pseudogene PTENP1 compete for the binding of the same miRNAs. The conservation of miRNA target-sites between the two genes allows the creation of a crosstalk between the two RNA molecules. As a consequence, the overexpression of PTENP1 3'UTR leads to miRNA sequestering from the functional PTEN gene resulting in its upregulation. Establishing the ceRNA hypothesis (Salmena *et al.*, 2011), they pinpointed the attention on pseudogenes since they are genomic loci that are similar to the functional genes that were believed to be completely non-functional. These "evolutionary relics" might be under a selective positive pressure to be maintained into the human genome specifically because of the sharing of miRNA target-sites with the functional counterpart showing a rather specific function (Pink *et al.*, 2011). Like the pseudogenes, also many long non-coding RNAs (lncRNA) have been observed to function as ceRNAs. The first evidence of a possible competition derived from this class of RNAs was found in the muscle (Cesana *et al.*, 2011). In this study, the authors discovered the muscle-specific linc-

MD1 that sequestering miR-133 and miR-135 enables the upregulation of the MAML1 and MEF2C genes controlling the muscle differentiation timing.

Considering the extensive complementarity of circular RNAs to their linear counterpart, it is not surprising that also this class of RNAs may be involved in miRNA competitition mechanisms (Tay, Rinn and Pandolfi, 2014). In addition, these molecules are considered to be more resistant to degradation with respect to the linear counterpart because of the circularizing covalent link at their ends. A ceRNA activity was described for this class of RNA molecules in the Hansen *et al.* (Hansen *et al.*, 2013) study. Here the authors observed the circular RNA ciRS-7 that, harboring target sites for the miR-7, activates the competition with the cerebellar degeneration-related protein 1 (CDR1) transcript in neuronal tissues. In zebrafish, the effect of this mechanism is analogous to miR-7 knockdown, causing an impaired midbrain development (Memczak *et al.*, 2013).

Adding support to the ceRNA hypothesis, also protein-coding RNAs were observed to regulate “in trans” the transcript levels one with each other within a competitive mechanism. In a large scale analysis of human glioblastoma samples (Sumazin *et al.*, 2011), about 7000 genes acting as ceRNAs were observed to create an intriguing cross-talk between different oncogenic pathways.

1.3.1 Factors regulating the ceRNA dynamics

As shown in Figure 10, different factors have been identified as possible key players in ceRNA dynamics.

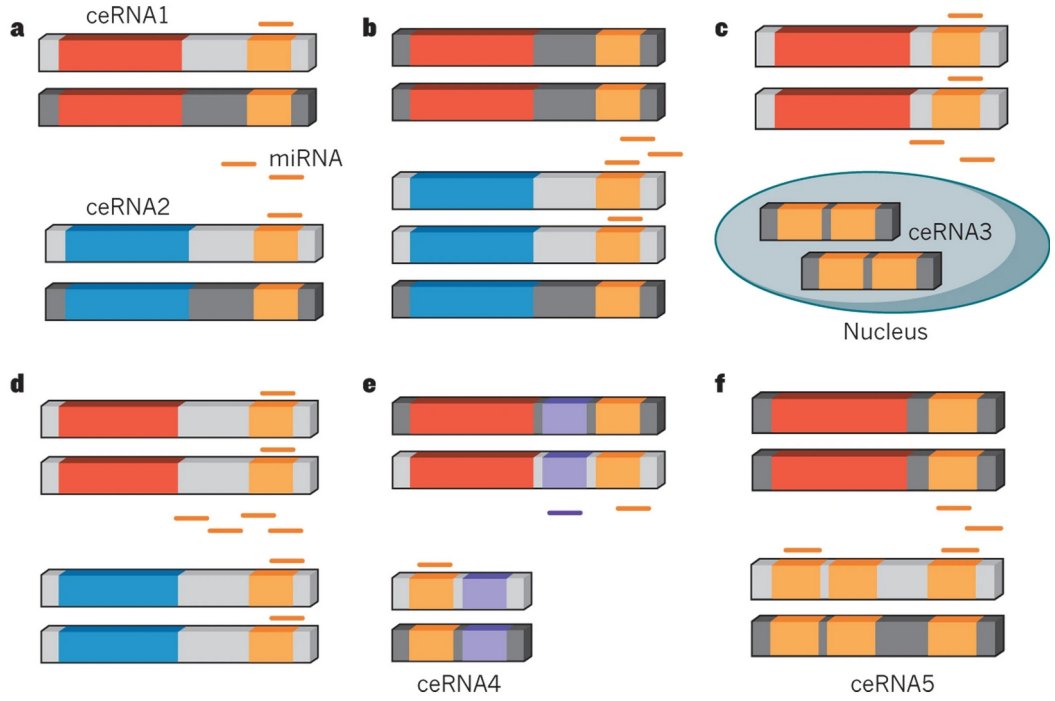

Figure 10: Factors influencing ceRNA dynamics. **A**. Steady state levels of ceRNA1 (red) and ceRNA2 (blue). Repression by miRNAs is shown in light grey. **B.** The ceRNA2 increase will induce the expression of ceRNA1. **C.** A different subcellular localization, such as the ceRNA3, may reduce the effectiveness. **D.** High miRNA levels will increase repression of both ceRNA1 and ceRNA2. **E.** ceRNA4 contains target-sites for multiple shared miRNAs, making it a more effective ceRNA. **F**. ceRNA5 contains more target-sites than ceRNA2 and making it a more effective ceRNA. Source (Tay, Rinn and Pandolfi, 2014).

Starting from a steady state (Fig. 9A), two exemplar ceRNA (shown in red and blue) transcripts are expressed at the same levels and they are competing for the binding of the miRNA. The miRNA, in turn, is able to bind both transcripts resulting in their degradation (shown with a light grey bar are copies that are degraded while in dark grey are the not degraded ones). When the expression levels of one ceRNA is increasing (Fig. 9B, blue ceRNA), there is an increased possibility for these transcripts to bind more miRNAs. The effect of the competition is the upregulation of the second ceRNA (red ceRNA) whose binding miRNAs were absorbed by the first upregulated ceRNA.

Nonetheless, factors that might potentially affect the ceRNA dynamics have been hypothesized to exist and they include:

- a different subcellular localization of one ceRNA that can reduce its ability to bind and therefore subtract miRNAs (Fig. 9C);
- the availability of multiple shared miRNA target sites (Fig. 9E) that can function concurrently for creating the competition;
- a different number of miRNA target-sites between the two ceRNAs that makes one of them a more effective competitor (Fig. 9F);

Considering what has been so far investigated, it was proposed that a competition between ceRNAs can occur only if the expression of the interested miRNA is within certain thresholds. High levels of the interested miRNA might in fact affect the total amount of transcripts derived from the two ceRNAs, abolishing in this way any competition (Fig. 9D). At the same time, low levels of the miRNA can be unlikely to contribute to the gene expression regulation of the ceRNA transcripts. In this case, even though a ceRNA level is increased, the low amount of the miRNA and therefore its weak regulatory action would have little impact on the transcript levels of the second ceRNA (Wee *et al.*, 2012). On the other side, also the ceRNA transcript levels play a significant role in these processes. Indeed, too high ceRNA transcript levels might potentially abolish the crosstalk. In this case, the miRNAs would be almost all bound to both abundant ceRNAs with a low amount of miRNAs available for the competition (Ala *et al.*, 2013).

The amounts of the AGO2 protein were finally observed to be a crucial rate-limiting factor for the ceRNA activity. In according to a mathematical model, it has been shown that the levels of the RISC complex must be kept within an intermediate range for the ceRNA mechanism to happen. Indeed given that, when transcribed, miRNAs are believed to be usually produced in excess, the limiting factor determining the existence of the ceRNA effect is the amount of AGO2 and therefore of the assembled RISC complexes. Low amount of Argonaute proteins promotes the competition, conversely, high protein amounts prevent any competition between ceRNA transcripts (Loinger *et al.*, 2012) because a limiting amount of complex is never reached.

1.3.2 The RIDL hypothesis: LincRNA-RoR and BACE1-AS cases

Despite sporadic findings, little is known about TEs acting as ceRNAs and their potential role as factors regulating ceRNA dynamics.

Strictly related to this, in the RIDL (Repeat insertion domains of lncRNAs) hypothesis the transposable elements were proposed to contribute to the functionalization of long non-coding RNAs (lncRNAs) (Johnson and Guigó, 2014). Indeed, despite the fact that ~13'000 lncRNAs in human are likely not able to encode for any protein (Ulitsky and Bartel, 2013), they have been shown to exert biological functions in cells. At the transcriptional level, those localizing in the nucleus may enable the recruitment of DNA-binding proteins that modulate both the genic transcription and the epigenetic state of the genome (Sun, Hao and Prasanth, 2018). Conversely, at the post-transcriptional level, the pairing of the lncRNA molecules with other RNAs can impact on their stability and splicing (X. Zhang *et al.*, 2019). According to the RIDL hypothesis, the capabilities of lncRNAs to mediate different functions are supposed to be addressed by specific TE sequences embedded within the transcripts. The different embedded TEs might therefore be considered as functional domains of lncRNAs. In particular (Figure 11), the Type I domains can provide motifs for interacting with proteins, whereas the Type II domains might permit the hybridization with nucleic acids. Various combinations of different domains might be at the basis of the functions of the lncRNAs.

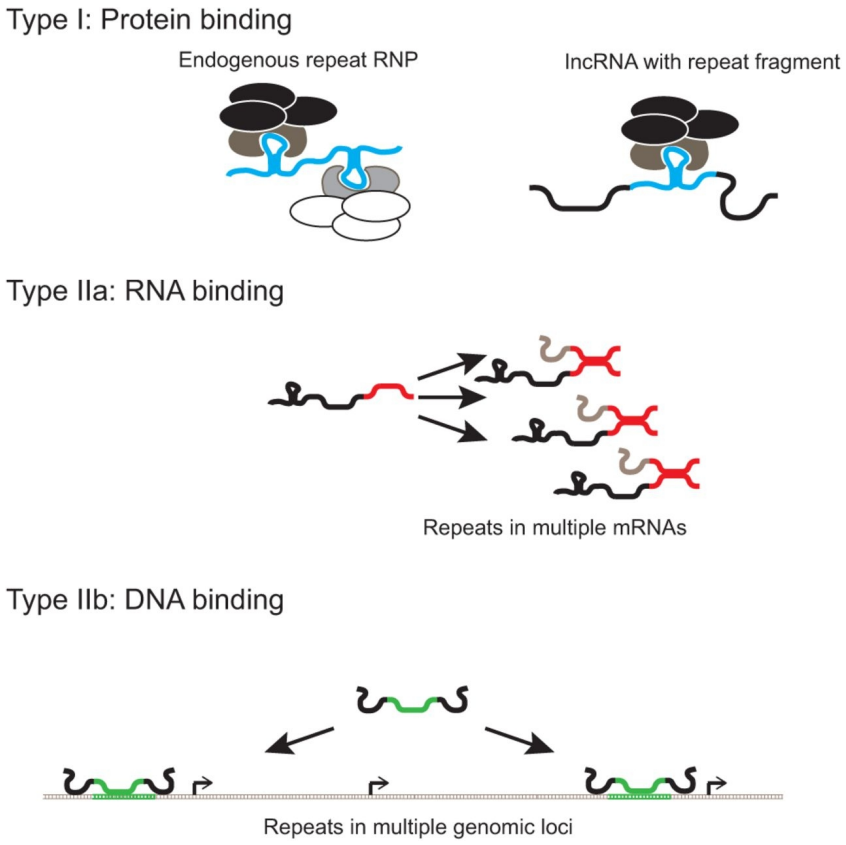

Figure 11: Functional classification of TE insertions. LncRNAs (black) may interact with proteins/mRNAs/genomic DNA (gray) because of TE insertions (Johnson and Guigó, 2014).

Intriguingly, lncRNAs have been observed to act as ceRNA and, accordingly to the high prevalence of TEs within the lncRNAs, there is a high probability that TE sequences may supply miRNA target-sites fundamental for the ceRNA activities. An evidence of this potential role is LINC-ROR (Long intergenic non-protein coding RNA), a human lncRNA whose ~73% of the sequence is derived from TEs (Wang *et al.*, 2013). Once expressed in human embryonic stem cells, this transcript is able to decoy miRNAs from transcription factors that are fundamental for the pluripotency status (OCT4, SOX2 and NANOG). The vast majority of the shared miRNA target-sites between this lincRNA and the transcription factors have been shown to overlap with TEs sequences within the lincRNA molecule, suggesting that these TEs are providing the functional sequences for the ceRNA activity. The putative final outcome of the ceRNA mechanism is the increased reprogramming efficiency of the cells.
In addition, also the BACE1-AS lncRNA has been observed to act as ceRNA for the sense transcript BACE1 (Zeng *et al.*, 2019). In this case, the antisense transcript

BACE1-AS shares the target-sites for several miRNAs with the sense transcript BACE1. Acting as ceRNA, BACE1-AS allow the overexpression of BACE1 which is a gene that encodes for an enzyme involved in the post-transcriptional processing of APP (amyloid Precursor Protein). Leading to the formation of beta-amyloid aggregates, the lack of BACE1 degradation is considered an hallmark of Alzheimer's disease. Given that the ~53% of the sequence composition of BACE1-AS is constituted by TE sequences, it is not surprising that also in this case the vast majority of shared miRNA target-sites are deriving from TEs. Although experimental evidences are not yet well established, it is reasonable to assume a contribution of TEs to the ceRNA activities of lncRNAs.

1.3.3 The quantification of TEs: challenges and solutions

If little is known about the TE contribution to the ceRNA activities of lncRNAs, even less is known about the autonomously transcribed TEs acting as ceRNAs. To address this question, it is necessary to start considering the difficulties of estimating the autonomous transcription of TEs in RNA-seq data (Lanciano and Cristofari, 2020). Let's discuss the case of LINE1 elements because it is the most important known autonomous TE in the majority of mammalian genomes excluding bats.

The first problem meet in the quantification from short RNAseq reads data is the repetitive nature of the LINE1 elements. This makes difficult the assignment of short sequence reads to the correct originating genomic loci. The existence of polymorphisms represents an additional problem. In fact, it is well known that the accumulation of mutations during evolution has caused the sequence of retrotransposons to diverge (Bourgeois and Boissinot, 2019). Additionally, insertion/deletion of a TE could result in private germline polymorphisms. Considering that on average 285 LINE1 sites can differ between two human individual genomes (Ewing and Kazazian, 2010), the difficulty in this case is represented by the human reference genome that is unable to represent all the different polymorphisms that might exist in a given population. Errors in assigning reads to the wrong genomic locations might lead to incorrect quantifications. Possible difficulties also depend on the formation of chimeric LINE1 transcripts. On one side, the weak polyadenilation

signal of LINE1 can cause the completion of the transcription in the downstream flanking sequence, resulting in a 3' readthrough mechanism (Holmes *et al.*, 1994). On the other side, LINE1s can be transcribed from a flanking 5' promoter, resulting in a 5' transduction process (Evrony *et al.*, 2012). Considering the possibility for a LINE1 to be involved in co-transcription phenomena, the complexity increases. In this case, a fragment from a TE can be included in a host gene and transcribed as a part of a canonical mRNA (Kapusta *et al.*, 2013) preventing a correct quantification of indipendently transcribed LINE1. The outcome of these issues is the difficulty to distinguish the autonomous transcription of LINE1 from transcription of transcripts containing LINE1 fragments. In this way, changes in LINE1 expression levels might simply reflect the variation of LINE1 host genes expression or the transcription of chimeric RNA molecules.

2 Aim of the project

Considering that little is known about autonomously transcribed TEs acting as ceRNA, my research intends to provide a preliminary assessment of a potential ceRNA mechanism driven by the autonomous transcription of LINE1s in human cells.

To achieve this, since no tool is currently able to discriminate active/passive TE transcription phenomena, I started developing a bioinformatics pipeline for the identification of samples showing autonomous transcription of LINE1 transcripts im situations in which disregulation of these elements is suggested. Then, I carried out transcriptomic analyses on different cellular conditions that are known to experience LINE1 overexpression. My PhD project drove me to explore the autonomous transcription of LINE1, their potential ceRNA activity and possible limiting factors that might abolish this phenomenon.

3 Materials and Methods

3.1 Data collection and pre-processing

To explore how regulatory mechanisms control LINE1 transcript levels in human cells, I took advantage of different publicly available dataset overexpressing LINE1s (Table 2). The dataset from Jonsson *et al.* (Jönsson *et al.*, 2019) is composed of poly-A RNA-seq data generated from embryo-derived human neural epithelial-like stem cell line Sai2. In this study, the authors produced 3 controls and 3 *DNMT1* samples in which *LacZ* and *DNMT1* genes were respectively knocked out with CRISPR-Cas9 technology. I retrieved paired-end raw FASTQ files from the ENA-EBI database (PRJNA420729 accession code). To validate the reliability of my method for the identification of autonomous/non-autonomous LINE1 transcription, I used the Marasca *et al.* (Marasca *et al.*, 2022) dataset. It is composed by total-RNA derived from quiescent naive $CD4^+$ T cells and naive $CD4^+$ T cells activated with anti-CD3-antiCD28 beads (3 individuals, 24 total sequencing). The download of paired-end raw FASTQ files was made from the ENA-EBI database (PRJEB41930 accession code). To test the miRNA-LINE1 expression levels association, I used RNA-seq and small RNA-seq data produced by the Geuvadis Project (Lappalainen *et al.*, 2013). In this project, polyA mRNAs were sequenced in lymphoblastoid cell lines derived from healthy individuals of the 1000 Genomes Project (Auton *et al.*, 2015). Quantification of miRNAs were retrieved from the ArrayExpress (Athar *et al.*, 2019) repository, while raw reads were retrieved for the TE quantification from ENA-EBI database (PRJEB3366 accession code). Aiming to investigate a cellular context in which LINE1 was artificially overexpressed, I took advantage of the publicly available RNA-seq data derived from the Ardeljan *et al.* study (Ardeljan *et al.*, 2020). In this work, the authors performed RNA sequencing of human Retinal Pigment Epithelial Cells (RPE) encoding a doxycycline-inducible (Tet-On) codon-optimized LINE1 (ORFeus) (An *et al.*, 2011) or luciferase as control. The two groups of cultures were sequenced in triplicate and the paired-end raw files were made publicly available in the ENA-EBI database (PRJNA491205 accession code). To explore other cellular contexts in which LINE1s should be deregulated, I analyzed the Deneault *et al.* (Deneault *et al.*, 2018)

polyA RNA-seq data in which, the *ATRX* gene was knocked-out in reprogrammed human induced pluripotent stem cells (iPSC) and iPSC differentiated in neuronal cells. The entire dataset composed by 12 controls and 8 treated samples was retrieved from the ENA-EBI database (PRJNA422099 accession code). The quality assessment of all the retrieved reads was performed with FastQC (*Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data*, no date).

3.2 Analysis of locus-specific TE expression

To analyze the TE transcriptome, I used the *SQuIRE* (Yang *et al.*, 2019) software for quantifying the locus-specific TE expression starting from RNA-seq reads. The customized parameter *--build hg19* was used for allowing the download of the needed files referred to the human reference genome *hg19*. Once I obtained the quantifications, I performed the differential expression analysis of TEs using the *R* package *DESeq2 (Love, Huber and Anders, 2014),* classifying as differentially expressed TE the elements showing p-adjusted<0.05 and |fold change|>1.5. Information about the class/family/subfamily classification and the match on the consensus sequence are retrieved from the *SQuIRE* output.

3.3 Detection of LINE1 autonomous transcription

To understand if LINE1 up-regulation is caused by autonomous transcription of LINE1 loci, I developed a specific *R* pipeline that analyzes paired-end reads of an RNA-seq experiment. In my method, reads are aligned on the LINE1 consensus sequence (Dombroski, Scott and Kazazian, 1993) with *BWA (Li and Durbin, 2009)* (*mem* command in default parameters). Then, not primary and supplementary alignments are discarded using *Samtools* (Li *et al.*, 2009) (*-F 2304*). The remaining *fragments*, corresponding to a reads pair, are filtered out if at least one read of the pair has more than 20% of read nucleotides that are not perfectly matched on the LINE1 consensus sequence. At this point, I label the *fragments* that are completely aligned inside the LINE1 consensus as *Inside fragments* and the *fragments* with just one read aligned inside the LINE1 as *Outside fragments*. These sets of *fragments* are then aligned on the human reference genome *hg19* with *BWA (Li and Durbin, 2009)* (*mem* command in default parameters). On the resulting BAM file, I apply the same

previously described filters. In the final step, I use *Bedtools* (Quinlan and Hall, 2010) (*intersect* command) to intersect the mapping coordinates of the reads with a BED file containing the LINE1 elements annotated on the human reference genome (*Repeat Masker (Tarailo-Graovac and Chen, 2009)* track retrieved from UCSC Table Browser (Karolchik *et al.*, 2004)). *Fragments* with at least one read aligned on an annotated LINE1 element are kept into account for the final calculation. For each sample, the ratio between the number of *Inside* and *Outside fragments* is used as an indicator for the LINE1 autonomous transcription level in all the group comparisons except for the *ORFeus* model in which the same ratio was calculated before the genome mapping step.

3.4 Gene expression analysis

To investigate the canonical gene expression levels, I aligned the FASTQ RNA-seq reads to the human reference genome *hg19* (FASTA file of primary assembly retrieved from *Ensembl (Hubbard et al.*, 2002)*)* using *STAR* (Dobin *et al.*, 2013). The indexing step was performed giving as *--sjdbGTFfile* argument the GRCh37.87 GTF file retrieved from *Ensembl*. The –sjdbOverhang argument was instead set to *max(ReadLength)-1* for the different datasets. Then, the mapping was performed with the parameters *--quantMode GeneCounts* and *--twopassMode Basic*. After the generation of BAM files, the counting of mapped reads for each gene was obtained with *HTSeq (Anders, Pyl and Huber, 2015)* (*htseq-count* command with the arguments *-t exon -i gene_id)*. Finally, the differential gene expression analysis was performed using *R* package *DESeq2 (Love, Huber and Anders, 2014)* classifying, as differentially expressed, the genes showing p-adjusted < 0.05 and |fold change| > 1.5.

3.5 Functional enrichment analysis

To explore the transcriptomic changes in the *DNMT1* model, the differentially expressed genes were divided into upregulated and downregulated according to previous thresholds. The two separated groups were then used as input to perform the enrichment analysis with the *R* package *gProfiler2* (Kolberg *et al.*, 2020) (*exclude_iea = T, user_threshold = 0.1* and *correction_method = "fdr"*). The background (*custom_bg* argument) used for the analysis was composed of genes with at least 5

mapped reads in at least 50% of the samples. Gene sets composed of more than 1000 genes were filtered out since they were describing too many general cellular processes. For the final exploration, gene sets with at least 10 enriched query genes and FDR <= 0.1 were considered significant.

3.6 Overlap analysis

To study how LINE1 overexpression can impact regulatory gene networks, I performed an overlap analysis between genomic coordinates of LINE1s and protein-coding genes. For this analysis, I used the GRCh37.87 GTF annotation file retrieved from *Ensembl* (Hubbard *et al.*, 2002). From this, I selected all the exonic and UTR annotations of protein-coding genes. Then, I used *Bedtools merge (Quinlan and Hall, 2010)* for collapsing together the features belonging to different transcripts of the same gene. *Bedtools subtract* was used to exclude both the UTR annotations from the exons and to generate the intron coordinates for each gene. As a result, I created a BED file containing a single gene model for each protein-coding gene. The models were thus composed of the coordinates of genic structures 5'UTR, exons, introns, and 3'UTR. These coordinates were intersected with LINE1 coordinates (retrieved from *SQuIRE* (Yang *et al.*, 2019) output) by using *Bedtools intersect* with "*-s"* argument to force the strandedness. Each genic structure was then classified based on two information: the up/down-regulation of the belonging gene and the overlap with at least one up/down-regulated LINE1. At this point, for each type of genic structure, I calculated the frequency of each possible pair of classifications. For the final calculation of the Z-score, the genic classifications were randomized 1000 times, recomputing the calculation of the frequencies each time which represents the random distribution.

3.7 Analysis of miRNA target-sites sharing

To determine whether the set of upregulated extragenic LINE1 elements are sharing miRNA target-sites with the 3'UTR of protein-coding genes, I used the BED file of genic structures created in the overlap analysis reported above. From this BED, I selected the 3'UTR regions longer than 30 nucleotides belonging to expressed protein-coding genes. A gene was considered as expressed if it was quantified with at least one mapped read in at least one sample. The upregulated LINE1 elements were identified

with *SQuIRE (Yang et al.*, 2019) as reported above, while the extragenic positioning of these elements was determined by intersecting the genomic coordinates of LINE1 and features annotated in the GRCh37.87 GTF file. All LINE1s that did not overlap with any concordant/discordant feature (*Bedtools intersect* with *-s* and *-S* arguments (Quinlan and Hall, 2010)) were considered extragenic. Once selected the 3'UTRs and LINE1 genomic coordinates, I used *Bedtools getfasta* to retrieve the FASTA file of each element and a custom *Perl* script to identify target-sites by looking only for 8-nt seed-matched sites (Bartel, 2009) of the entire set of human miRNAs retrieved from *miRBase* database (Griffiths-Jones *et al.*, 2006). Hence, for each 3'UTR of protein-coding genes was calculated the number of miRNA target-sites that were found also in the pool of extragenic upregulated LINE1s. T-test were finally used for comparing the number of LINE1-shared miRNA target-sites between up and down-regulated genes.

3.8 Identification of miRNAs sequestered by LINE1s

To search for miRNAs that are possibly sequestered by LINE1s, I analyzed miRNAs that were observed to potentially target at least one upregulated extragenic LINE1. For each of the miRNAs, I identified the targeted protein-coding genes by searching for 8-nt seed-matched sites. The targeted genes were then classified as up or down-regulated based on the previous differential gene expression analysis. At this point, for each miRNA, I performed a proportion test (*prop.test R* function) to compare the proportion of up and down-regulated targeted genes to the proportion of all up and down-regulated genes. miRNAs in which the proportion of targeted upregulated genes was significantly (FDR < 0.1) higher than the proportion of downregulated ones were selected for further analyses.

3.9 miRNA-gene networks identification

To explore miRNA-genes interactions, I used the MIENTURNET (Licursi *et al.*, 2019) web-based tool. I provided the miRBase ID list of 117 miRNAs potentially decoyed by LINE1s in the DNMT1 experiment. The miRNA-target enrichment analysis was performed with the minimum number of miRNA-target interactions set to 2 and an adjusted p-value (FDR) threshold of 0.1. The interactions categorized by *miRTarBase* (Huang *et al.*, 2020) as strongly and weakly validated were taken into

account. The following enrichment analysis of 57 genes upregulated in the *DNMT1* model was performed with *EnrichR* (Chen *et al.*, 2013) considering significant the gene set enriched with an FDR below 0.1.

3.10 Association analysis of miRNA-LINE1 expression levels

To investigate the association between the expression levels of miRNAs and LINE1 elements in Geuvadis dataset (Lappalainen *et al.*, 2013), I used the quantifications retrieved from ArrayExpress (Athar *et al.*, 2019) or obtained with SQuIRE (Yang *et al.*, 2019) respectively. For each sample, the LINE1 expression level was obtained by summing the DESeq2 (Love, Huber and Anders, 2014) normalized counts of elements belonging to the L1HS/L1PA subfamilies and longer than 5000 bp; the sample "NA18861" was discarded since it appeared to be an outlier. The miRNA-LINE1 expression levels association was analyzed by performing Pearson's correlation tests.

3.11 Analysis of TE expression at consensus level

In order to investigate the TEs expression at the consensus level in the *ORFeus-OE* experiment, I used the *TEspeX* (Ansaloni *et al.*, 2022) software. For the analysis, I provided a modified version of TE consensus sequences from the *Dfam* database (Hubley *et al.*, 2016) that includes the LINE1-ORFeus sequence (https://www.addgene.org/browse/article/28204003/). Furthermore, the annotation files of coding and non-coding transcripts were retrieved from *Ensembl* (Hubbard *et al.*, 2002) and referred to the *hg19* version of the human genome. After obtaining the quantifications, I used the *R* package *DESeq2 (Love, Huber and Anders, 2014)* for the differential expression analysis, classifying as differentially expressed TE the subfamilies showing p-adjusted < 0.05 and fold change > |1.5|.

4 Results and Discussion

4.1 LINE1s deregulation upon DNMT1-KO

The first set of analyses was carried out to explore the transcriptomic changes of human neural progenitor cells (hNPCs) produced in the Jonsson's study (Jönsson *et al.*, 2019). This dataset (Table 2) is composed of RNA-seq data derived from 3 controls and 3 samples in which the DNMT1 gene was knocked-out (KO) with CRISPR-Cas9 technology. I decided to use this dataset since the authors were able to create a cellular model suitable for studying the effects of aberrant LINE1s transcription. The abrogated activity of the most important maintenance DNA methyltransferase (Hermann, Goyal and Jeltsch, 2004) erases a fundamental epigenetic repressive layer from the LINE1 defense mechanisms.

ID	Experiment name	Experiment	RNA	Source	Number of samples	Sample type
PRJNA420729	DNMT1 model	KO of DNMT1	Poly-A	embryo-derived human neural epithelial-like stem cell line Sai2	6 (3 vs 3)	CTR vs DNMT1-KO
PRJEB41930	$CD4^+$ model	Activation of T $CD4^+$ cells	Poly-A	$CD4^+$ T cells of healthy individuals	24 (12 vs 12)	Naive T $CD4^+$ cells vs Activated T $CD4^+$ cells
PRJEB3366	Geuvadis project	No treatment	Poly-A	EBV transformed lymphoblastoid cell line of healthy individuals	449	CTR
PRJNA491205	ORFeus model	Overexpression of LINE1	Poly-A	human Retinal Pigment Epithelial Cells (RPE)	6 (3 vs 3)	CTR vs LINE1-OE
PRJNA422099	ATRX model	KO of ATRX	Poly-A	reprogrammed human induced pluripotent stem cells (iPSC) and iPSC differentiated in neuronal cells	12 (6 vs 6) 8 (4 vs 4)	6 iPSC CTR vs 6 iPSC ATRX-KO 4 Neurons CTR vs 4 Neurons ATRX-KO

Table 2: Information about the analyzed dataset.

4.1.1 Evaluation of TE transcription

To understand if the LINE1 elements were autonomously transcribed (Lanciano and Cristofari, 2020; Gualandi *et al.*, 2022) in this model, I first started to explore and analyze the general expression of TEs. To obtain this, I used SQuIRE (Yang *et al.*, 2019) to analyze the locus-specific expression of TEs. The differential expression analysis highlighted that the KO of DNMT1 leads to a strong overexpression of TEs: 3015 are the up-regulated and 277 the down-regulated ones. Interestingly, 1660 LINE1s were upregulated making them the most upregulated TE family in this model, as shown in Figure 12A. Going deeper into the subfamily classification of the LINE1s that result upregulated, I can appreciate that ~70% of these elements belong to the young L1HS/L1PA subfamilies. Furthermore, it is also evident from the figure 12B that the elements belonging to these subfamilies are rather close to the canonical 6000 bp full-length size. Since this suggests that the CpG demethylation might activate the autonomous transcription of LINE1 elements, I examined how the LINE1 consensus sequence is represented in the set of upregulated LINE1s. As shown in Figure 12C, almost the 60% of the upregulated LINE1s are carrying the 5'UTR regions that putatively contains the internal promoter.

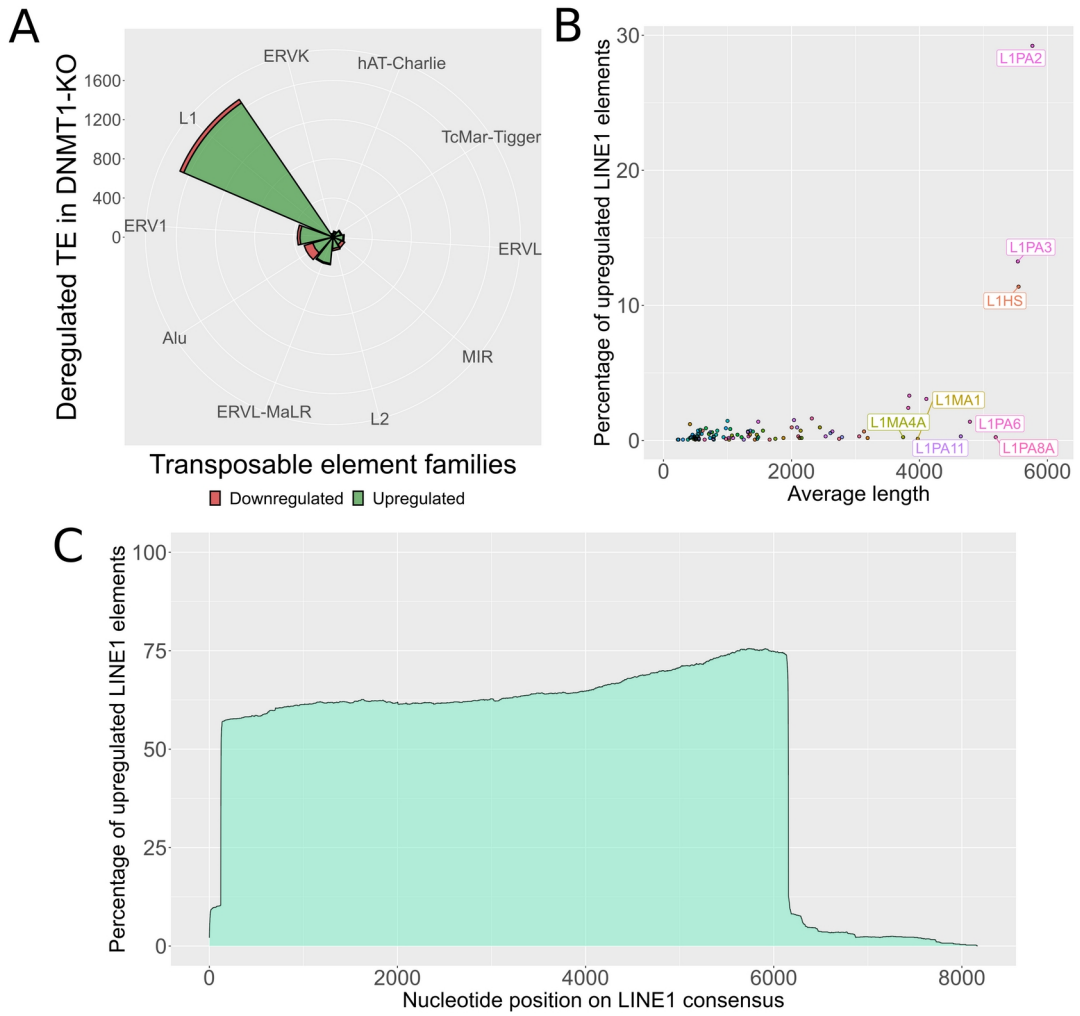

Figure 12: Evaluation of TE transcription upon the DNMT1-KO. **A.** Top-10 deregulated TE families in DNMT1 model. LINE1 is the most upregulated TE family with 1660 elements. **B.** For each upregulated LINE1 subfamily is represented the percentage with respect to the total number of upregulated LINE1 (Y-axis) and the average size of the elements (X-axis). **C.** For each nucleotide position of the LINE1 consensus sequence (X-axis) the percentage of the upregulated LINE1s that are covering the part is reported (Y-axis).

In agreement with the results provided by the authors (Jönsson *et al.*, 2019), upon the KO of the maintenance methyltransferase DNMT1, the cell cultures experience a strong CpG DNA demethylation that is strictly responsible for the transcriptional reactivation of a huge amount of LINE1 genomic loci. Moreover, analyzing the length and the subfamily classification of the LINE1s resulting upregulated, I observed that the majority of them belongs to the L1HS and L1PA subfamilies. As reported in literature (Brouha *et al.*, 2003), the LINE1 subfamilies HS (human specific) and PA (primates) are the youngest in the human genome, thus they have accumulated few

mutations. Since they might preserve a sequence very close to the functional one, I can hypothesize that they are elements carrying a possibly functional promoter that allows their transcription as autonomous transcriptional units. In support of this, I have also observed how the majority of upregulated LINE1s include the 5'UTR of the consensus sequence, thus containing the internal promoter region.

4.1.2 Tool for identifying autonomous LINE1 transcription

The promoters in the upregulated LINE1s might have accumulated mutations that inactivated the capability to start the transcriptional process. Indeed, it has been estimated that ~99% of LINE1 RNA sequences in human arise from LINE1s embedded in other transcripts rather than from LINE1 promoter (Deininger *et al.*, 2017). In order to confirm the autonomous transcription of LINE1 elements following DNMT1 KO, a new complementary approach was needed. For this reason I developed a specific bioinformatic pipeline.

With this tool, I took advantage of the paired-end reads produced by the RNA-seq experiment. In a paired-end experiment, the sequencing of an RNA fragment starts at one end, finishes at the specified read length and then starts another round of sequencing from the opposite end of the fragment. The result is a pair of sequence reads (referred to as *fragment*) that, when aligned on a reference genome, are divided by an inner distance as shown in Figure 13.

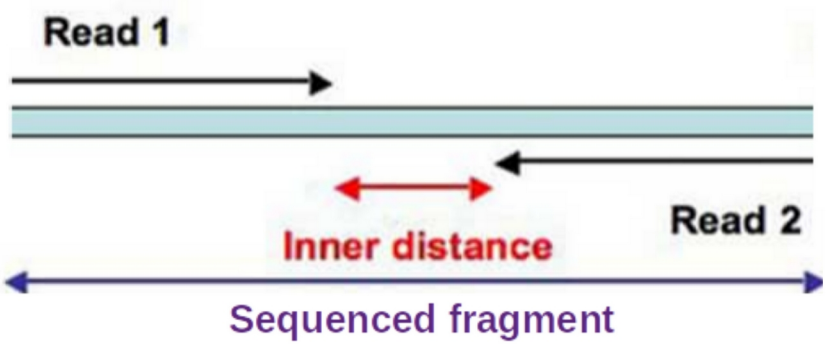

Figure 13: A schematic view of Illumina paired-end reads for a single fragment. Source: modified from https://thesequencingcenter.com/knowledge-base/what-are-paired-end-reads/.

In my pipeline (Figure 14A), paired-end reads are first aligned on the LINE1 consensus sequence and then on the human reference genome. The goal is to identify the number of *fragments* that are completely aligned inside the LINE1 element (*Inside fragments*) and the number of *fragments* with a read aligned inside a LINE1 and the other outside (*Outside fragments*). The sequencing of autonomous LINE1 transcripts should produce almost exclusively paired-end reads coming from the internal part of the element, hence mainly *Inside fragments*. Conversely, the transcription of LINE1 fragments as part of other transcriptional units, such as coding and non-coding canonical genes, should result in a much higher proportion of *Outside fragments*. According to this, by comparing the *Inside*/*Outside* ratio between two conditions in which LINE1s result disregulated, difference in the ratio levels will indicate whether LINE1 elements disregulation derives from autonomous transcription beginning in the LINE1 promoter.

4.1.3 Analysis of LINE1 autonomous transcription

Applying my methodology to the DNMT1 model, I was able to calculate a mean *Inside*/*Outside* ratio of 7.53 in the control group and a mean ratio of 9.33 in the KO group. A significantly (t-test p-value = 0.018) higher *Inside*/*Outside* ratio was observed in the DNMT1 KO samples with respect to controls (Figure 14B).

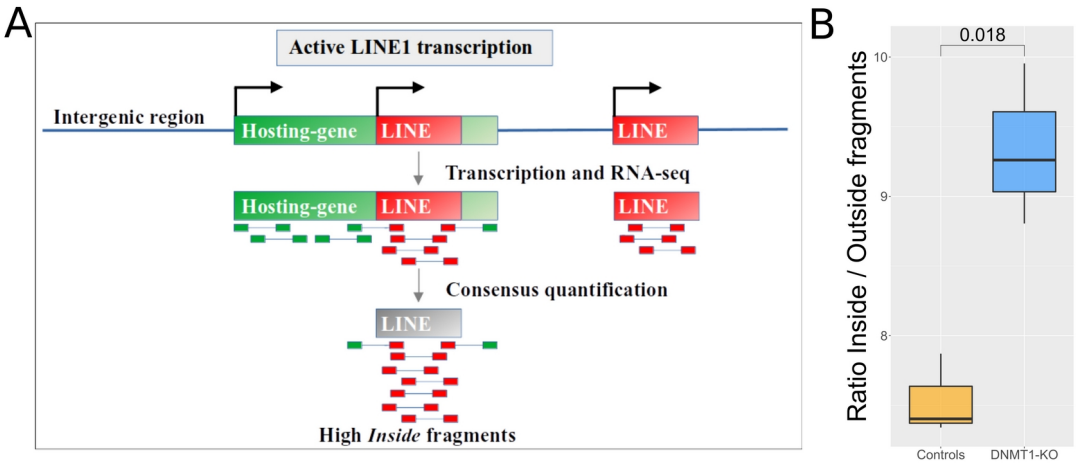

Figure 14: The KO of DNMT1 leads to autonomous LINE1 transcription. A. Rationale of my method for detecting autonomous transcription of LINE1 elements. Autonomous transcription of LINE1s will produce a higher amount of Inside fragments with both reads mapped inside the LINE1 consensus resulting in a higher Inside/Outside ratio. B. Analysis to detect autonomous transcription of LINE1 elements. Upon the KO of DNMT1, the upregulation of LINE1s derives from an autonomous transcription of elements.

This result thus suggests that the KO of DNMT1 leads to an autonomous and independent transcription of LINE1 elements. This sounds reasonable considering that the most upregulated retrotransposons belong to L1HS and L1PA subfamilies, they are not far from the canonical functional LINE1 element, and they might contain a functional LINE1 promoter. In the absence of the repressive CpG DNA methylation layer, it is realistic to conclude that these elements are likely to be recognized from the PolII for starting their autonomous transcription like the author demonstrated. These results support the validity of my approach.

4.1.4 LINE1s deregulation in a not-autonomously transcribed LINE1 dataset

In order to verify the capability of my method to identify situations in which LINE1 transcriptional overexpression is determined by not-autonomous and non-independent LINE1 transcription (i.e. LINE1s transcription mainly results from the transcription of their fragments as part of canonical genes), I took advantage of the Marasca *et al.* (Marasca *et al.*, 2022) dataset. This dataset is composed of total RNA-seq data derived from 3 individuals in which quiescent naive $CD4^+$ T cells were sequenced to create 12 samples. Once activated with anti-CD3-antiCD28 beads, the active $CD4^+$ T cells were sequenced to create another group of 12 samples. In this study, the authors observed that quiescent naive CD4+ T cells transcribe LINE1-containing transcripts as non-canonical splicing variants. Upon cell activation, modifications in the splicing pattern induce the downregulation of these alternative transcripts, promoting the transcription of the canonical ones.

4.1.4.1 Evaluation of TE transcription

To test my bioinformatic pipeline, I began quantifying the locus-specific expression of TEs with SQuIRE (Yang *et al.*, 2019). The differential expression analysis confirmed that quiescent naive cells, with respect to activated ones, are apparently characterized by the overexpression of LINE1 elements: 10,794 are the up-regulated and 5289 are the down-regulated (Figure 15A). Then, investigating the subfamily classification of the LINE1s upregulated in quiescent naive CD4+ T cells, I noted that the elements of the young L1HS and L1PA subfamilies are representing a little fraction (~16%) with respect to the total amount of upregulated LINE1s and they have an average length that is far from the canonical 6000 bp full-length size. In addition to this, the Figure 15B clearly show how the vast majority of the upregulated LINE1s belongs to the L1M subfamily. This more ancestral subfamily is present in different mammalian species and is characterized by a short length, ranging from 100 to 2000 nucleotides because it is formed principally by fragmented elements (Smit *et al.*, 1995).

To confirm the authors observation that apparent overexpression of LINE1 elements is not autonomous but is the results of transcription as part of hosting genes, I explored the coverage profile on the LINE1 consensus sequence. As shown in Figure 15C, just

about the 10% of the upregulated LINE1s carry the starting 5'UTR part of the LINE1 that contains the internal promoter. Moreover, with respect to the coverage profile of the DNMT1 model, the general low percentages observed here reflects the overexpression of elements that are relatively short, as also observed in Figure 15B. The pronounced peak at the 3'UTR region finally suggests how the upregulated LINE1 elements are mainly 5' truncated.

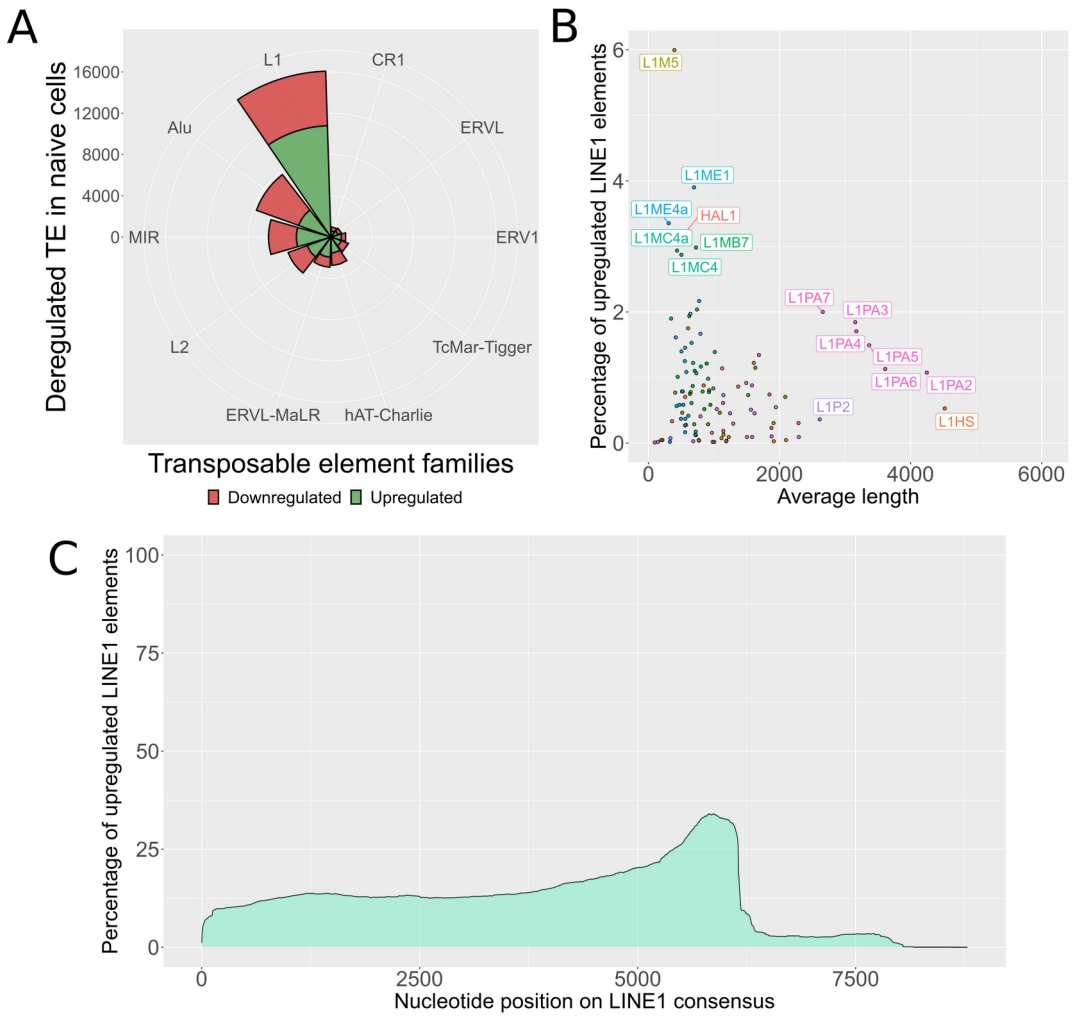

Figure 15: Evaluation of TE transcription in quiescent naive CD4⁺ T cells. **A.** Top-10 deregulated TE families in naive T CD4⁺ cells. LINE1 is the most upregulated TE family with 10,794 elements. **B.** For each upregulated LINE1 subfamily is represented the percentage with respect to the total number of upregulated LINE1 (Y-axis) and the average size of the elements (X-axis). **C.** For each nucleotide position of the LINE1 consensus sequence (X-axis) I report the percentage of the upregulated LINE1s that are covering the part (Y-axis).

In light of these analyses, the results suggest that the quiescent naive T *CD4*$^{+}$ cells transcribe a very large amount of LINE1 fragments belonging to the ancestral subfamilies. Since during the evolution these elements accumulated many inactivating mutations, it is very unlikely that they are transcribed as autonomous units. As described by the authors (Marasca *et al.*, 2022), their transcription depends on hosting genes that are transcribed as splicing isoforms embedding the LINE1 fragments. Due to the strong divergence of their sequences from the functional one and considering their short length depending on the 5' truncation, I hypothesize that they are elements not containing a possibly functional promoter. Since their transcription should be non-autonomous, this dataset is suitable for testing my pipeline for detecting autonomous/non-autonomous transcription of LINE1 elements.

4.1.4.2 Analysis of LINE1 non-autonomous transcription

To test my pipeline on a putative non-autonomous LINE1 transcription experiment, paired-end reads were firstly aligned on the LINE1 consensus sequence and then on the human reference genome to identify the *Inside* and the *Outside fragments*. In the case of non-autonomous transcription of LINE1 elements, I thought that the sequencing of hosting genes embedding LINE1s should not produce as many as *Inside fragments* like in cases of LINE1 produced during the autonomous transcription. In agreement to this, the *Inside*/*Outside* ratio should not result in differences between control and treated groups when LINE1 elements are transcribed as part of hosting transcripts (Figure 16A) and no alterations in the autonomous transcription of LINE1 happen.

To assess if the previously described LINE1 overexpression is the result of LINE1s transcribed as part of other transcriptional units, I used my pipeline to calculate the *Inside*/*Outside* ratio. Applying my methodology to the CD4+ model, I was able to calculate a mean *Inside*/*Outside* ratio of 87.43 in the naive group and a mean ratio of 87.12 in the active group, as reported in Figure 16B.

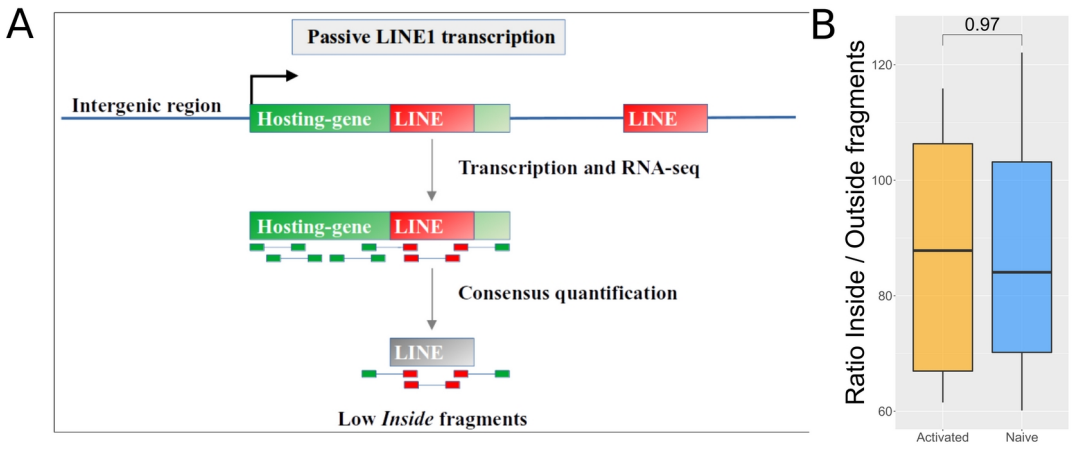

Figure 16: The quiescent CD4^{+} T cells overexpress LINE1 non-autonomously. **A.** Rationale of my method for detecting non-autonomous transcription of LINE1 elements. Transcription of LINE1s embedded in other transcriptional units will produce a low amount of Inside fragments. **B.** Analysis to detect autonomous transcription of LINE1 elements. In naive T CD4+ cells, the upregulation of LINE1s derives from a non-independent transcription of elements. Given that the two distributions are similar the overexpression of LINE1s is due to their transcription as part of other transcriptional units.

Since no statistical significant differences between the two groups were found (Figure 16B), the CD4+ model allowed me to confirm that the apperent overexpression of LINE1 in quiescent naive T cells is not due to anutonomous LINE1 transcription. Supporting this, also the previous qualitative analysis of upregulated LINE1 subfamilies clearly showed how the vast majority of them belonged to short-length ancient groups that do not contain functional promoters for the autonomous transcription.

Supported by the reliability of my method to discriminate between autonomous and non-autonomous transcription of LINE1 elements in case of observed overexpression, I can consider that the DNMT1 model represents an ideal opportunity for studying the influence that autonomous LINE1 transcription might have on the cells.

4.2 Effects of LINE1 dysregulation upon DNMT1-KO

Once the autonomous transcription of LINE1 elements was established in the DNMT1 model, I aimed at evaluating the impact of this deregulation on the expression of canonical genes. Indeed, it is well documented in literature how the overexpression of these elements can be detrimental for the cell and it is associated to a wide spectrum of human diseases (Zhang, Zhang and Yu, 2020).

4.2.1 Analysis of deregulated genes

To explore the effects of the deregulated LINE1 overexpression, I started carrying out the differential gene expression analysis for the identification of deregulated genes: 2188 genes resulted upregulated while 627 downregulated upon the KO of DNMT1. Functional enrichment analyses (Table 3) revealed that genes belonging to the piRNA pathway (GO:0034587) (De Fazio *et al.*, 2011) and to p53 transcriptional gene network (WP:WP4963) (Tiwari *et al.*, 2020) were enriched among the upregulated group while several proliferation-related genesets (e.g. GO:0042127) (Belgnaoui *et al.*, 2006) resulted enriched among the downregulated one.

Term ID	Term name	FDR
GO:0034587	piRNA metabolic process	2.37E-07
GO:0032504	multicellular organism reproduction	5.70E-05
GO:0022414	reproductive process	5.70E-05
GO:0048609	multicellular organismal reproductive process	5.70E-05
GO:0000003	reproduction	5.70E-05
WP:WP4963	p53 transcriptional gene network	1.05E-03
WP:WP4760	PKC-gamma calcium signaling pathway in ataxia	1.05E-03
WP:WP4673	Male infertility	1.88E-02
WP:WP2884	NRF2 pathway	2.31E-02
WP:WP2882	Nuclear receptors meta-pathway	3.73E-02
GO:0050680	negative regulation of epithelial cell proliferation	2.52E-03
GO:0050678	regulation of epithelial cell proliferation	5.63E-03
GO:0050673	epithelial cell proliferation	1.19E-02
GO:0032368	regulation of lipid transport	4.00E-02
GO:0001936	regulation of endothelial cell proliferation	4.00E-02

Table 3: Biological processes and WikiPathways enriched upon DNMT1-KO. In green and red are reported the enrichment coming from the analysis performed on upregulated and downregulated genes, respectively. For both biological processes and WikiPathways, the top-5 significant enriched gene-set were selected.

On one hand, these enrichments suggest that cells respond, in addition to a general demethylation, also to the transcriptional activation of LINE1s and possibly to a genotoxic stimuli (De Cecco *et al.*, 2019). On the other hand, they represent an important pool of genes to study for understanding how their deregulation can be directly or indirectly classified as a result of the LINE1 overexpression.

4.2.2 Overlap analysis

To further assess the impact of the active LINE1 transcription on regulatory gene networks, I carried out an overlap analysis of genomic coordinates between LINE1 and the genic structures of protein-coding genes. The idea behind this approach was to observe how the LINE1 hosted by genes were deregulated. The aim of this strategy was to understand if the LINE1 transcription was influencing transcriptional or post-transcriptional regulatory events of the hosting gene (Feschotte, 2008).

In this analysis, for each protein-coding gene I created a representative gene model composed of the coordinates of the genic structures: 5'UTR, exons, introns, and 3'UTR. These coordinates were intersected with LINE1 coordinates for classifying each genic structure based on the deregulation of both the gene and the overlapping LINE1s. The co-occurence frequencies of each possible couple of classifications was then statistically tested. With this procedure, the results revealed that upregulated genes were significantly enriched (Z-score > 3) to be in overlap with upregulated LINE1 elements in all the considered genic structures, as visible in Figure 17A.

While the enrichment in introns might be explained by phenomena like intron retention (Gualandi *et al.*, 2022) or the transcription of splicing variants that were including intronic LINE1s (Marasca *et al.*, 2022), the most intriguing finding was the observation that also the 3'UTRs of upregulated genes were significantly enriched to host upregulated LINE1 elements. In light of this observation, I wondered whether I could hypothesize a ceRNA activity for LINE1 transcripts and this enrichment could be an outcome of this mechanism. In my hypothesis, overexpressed LINE1 transcripts share miRNA target-sites with the 3'UTRs of a given gene set. As a result, LINE1s might be able to sequester miRNAs from the target transcripts. As a result of this

mechanism, the transcripts from which miRNAs are sequestered should result upregulated and enriched to contain upregulated LINE1s in their 3'UTR.

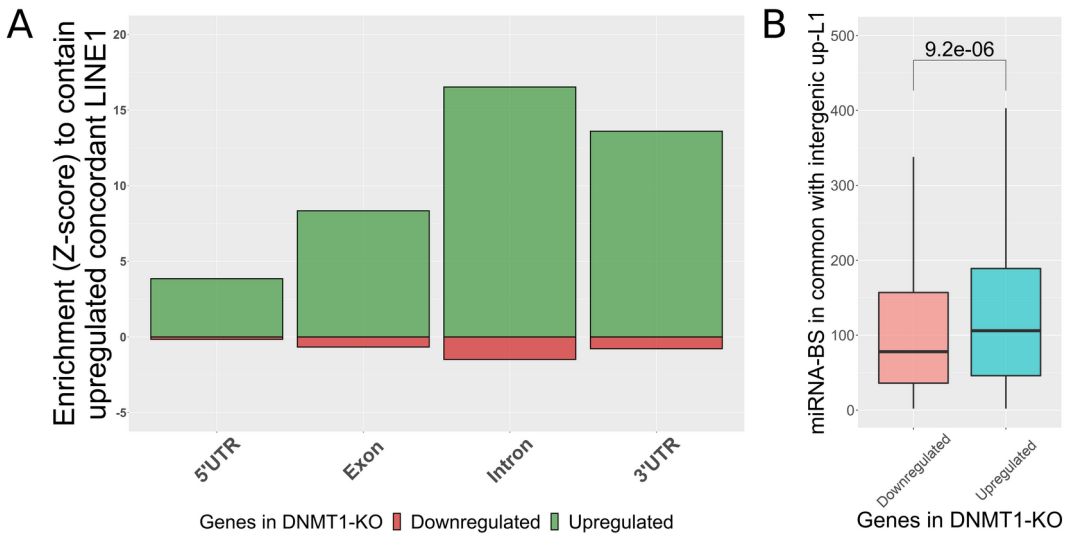

Figure 17: LINE1 transcripts might act as ceRNA. **A.** Overlap analysis of LINE1s and protein-coding genes. Regions of upregulated genes are significantly enriched to contain upregulated LINE1 fragments. I found particularly interesting the strong enrichment in the 3'UTRs. **B.** Analysis of miRNA target-sites sharing between autonomously transcribed LINE1s and the 3'UTRs of protein-coding genes in DNMT1 model. The upregulated genes share a significantly higher number of miRNA target-sites with active LINE1s with respet to the downregulated, adding support to a possible ceRNA activity of LINE1 transcripts.

In order to verify this hypothesis, I selected the 799 upregulated extragenic LINE1s to represent the group of autonomously transcribed LINE1 elements (referred to as *active* LINE1s) that might possibly act as ceRNA. This set was chosen since they should be less involved in passive transcription phenomena with respect to LINE1s embedded in genes (Hermann, Goyal and Jeltsch, 2004) and therefore I expect less background noise. At this point, I identified the miRNA target-sites both in the *active* LINE1s and in the 3'UTRs of protein-coding genes.

Upregulated genes shared on average 147 miRNA target-sites with *active* LINE1s while the downregulated genes shared about 114 miRNA target-sites. This difference is significant and indicates that upregulated genes, with respect to downregulated ones, have a significantly higher number of miRNA target-sites in common with the *active* LINE1s (Figure 17B, p-value = 9.2e-6).

Taken together, these results suggest that the group of upregulated genes, sharing more miRNA target-sites with autonomously transcribed LINE1s, might be undergoing a possible ceRNA activity of LINE1s. Since the DNMT1 model is characterized by a strong autonomous transcription of LINE1 elements, these transcripts might sequester the miRNAs from the protein-coding genes. The result of this mechanism would be the upregulation of this specific gene set.

4.2.3 Identification of decoyed miRNAs

Looking into the mechanism from the perspective of the miRNA, I expect that if a given miRNA is sequestered by the pool of *active* LINE1s, the effect must be reflected on the entire group of genes that are controlled by that miRNA. In particular, I thought that genes usually targeted by the decoyed miRNA should result mostly upregulated. This hypothesis was at the basis of the attempt to identify the pool of miRNAs that could be bound and decoyed by the autonomously transcribed LINE1s.

With this in mind, I analyzed 2563 miRNAs that targeted at least one *active* LINE1. For each miRNA, the targeted protein-coding genes were classified as up- or down-regulated based on the previous differential gene expression analysis. Then, I searched for miRNAs in which the proportion of targeted upregulated genes was significantly (FDR < 0.1) higher than the proportion of downregulated ones.

With this procedure, I identified 117 miRNAs for which the proportion of targeted genes resulting upregulated is significantly higher than the proportion of downregulated ones. These represent the most suitable pool of miRNAs undergoing the putative ceRNA activity of LINE1s (Figure 18A). Among them, worthy of note miRNAs are those belonging to the let-7 family. As previously described, this family of miRNAs was demonstrated to exert its functions by inhibiting the translation of the L1-ORF2p without affecting the overall mRNA stability (Tristán-Ramos *et al.*, 2020). Since the genes targeted by the let-7 family are mostly upregulated, I can believe that this method is reliable for detecting the decoyed miRNAs.

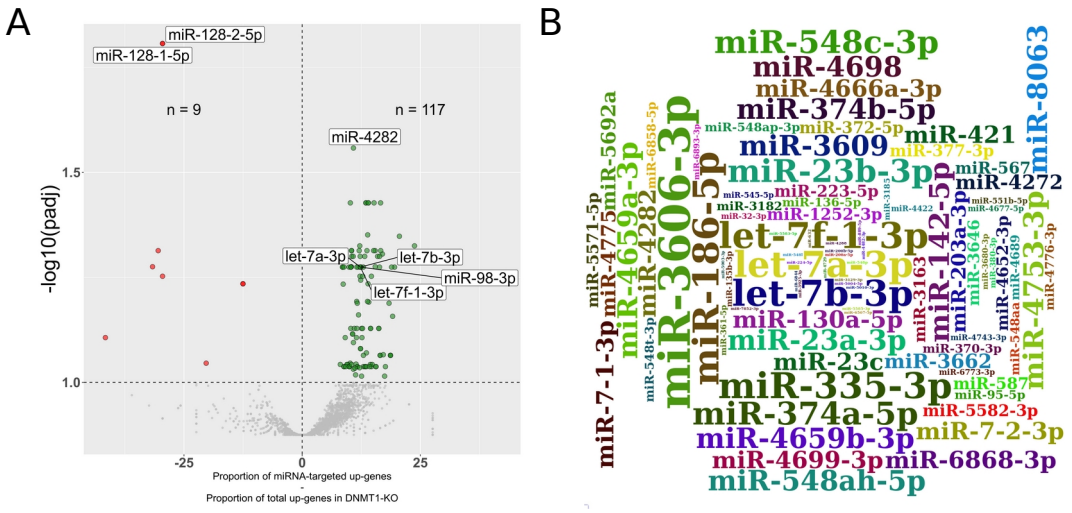

Figure 18: let-7 miRNA family might be decoyed by LINE1. **A.** Analysis to identify miRNAs sequestered by LINE1s. Each analyzed miRNA is represented by a point with the X-axis indicating the delta between the proportion of targeted upregulated genes and the total upregulated genes proportion. The Y-axis represents the -log10(FDR) of the proportional statistical test applied. The 117 miRNA in green are representing the most probable pool of miRNAs that are undergoing the ceRNA activity of LINE1s. **B.** In this wordcloud are shown the 117 miRNAs whose size font is proportional to the absolute number of experimentally validated target genes upregulated in DNMT1 model. The let-7 family is among the top-10 miRNAs with the higher number of connections to these upregulated genes.

Once obtained a set of miRNAs putatively decoyed from the *active* LINE1 transcripts, it was crucial to look into the potential direct effects of this phenomenon on the regulatory gene networks. For this reason, aiming to understand which are the genes that are upregulated upon the LINE1 ceRNA activity, I took advantage of the MIENTURNET (Licursi *et al.*, 2019) tool.

In this analysis, I provided the list of 117 identified miRNAs for discovering miRNA-genes interactions categorized by miRTarBase (Huang *et al.*, 2020) as experimentally validated. The tool found 3878 miRNA-genes interactions involving 108 out of 117 miRNAs and 543 genes. Interestingly, 57 of these genes result upregulated in the DNMT1 model. The enrichment analysis revealed that they are involved in processes that modulate the transcription (GO:0006355) and particularly the p53 transcriptional program (WP4963). Investigating the top-10 miRNAs with the most connections to the 57 upregulated genes, I found the group of four members (let-7a-3p, let-7b-3p, let-7f-1-3p and miR-98-3p) of the let-7 family (Roush and Slack, 2008) (Figure18B).

The overall results suggest that a putative LINE1-miRNA-gene network might involve miRNAs that are demonstrated to regulate the activity of LINE1s, like those belonging to the let-7 family. Moreover, the genes that result upregulated because of the ceRNA activity driven by the LINE1 transcripts might be required to build a part of defense cellular mechanisms that are specifically activated in cellular contexts overexpressing LINE1 elements, as the p53 transcriptional program (Wylie *et al.*, 2016). Due to its role as "guardian of the genome" the p53 activation might arrest the cell cycle progression and/or promote the establishment of repressive histone marks in LINE1 genomic loci. In this way, the ceRNA activity of LINE1s might function as sensor for triggering a transcriptional response preventing the possible DNA damage deriving from the LINE1 deregulation.

4.2.4 The miR-128 case

In addition to the group of 117 miRNAs, an unexpected set of 9 miRNAs showed an opposite enrichment (Figure 18A). Unlike the 117 putatively decoyed miRNAs, the putative targets of these 9 miRNAs resulted enriched among the downregulated genes. Top significant miRNAs of this set belonged to the miR-128 family, another family of miRNAs demonstrated to regulate the levels of LINE1 transcripts (Hamdorf *et al.*, 2015). In light of this observation, the strong enrichment for downregulated genes to share miR-128 targets with LINE1s is unexpected although it might hide an additional layer of regulatory complexity. If miR-128 expression would be the results of a cellular defense mechanism responding to an increase in LINE1s expression, then its expression levels could be related to the amount of LINE1 transcripts present in the cell. A cell would respond to an overexpression of LINE1s with the activation/increase of miR-128 expression. In this situation it is reasonable to assume that a portion of the *de-novo* produced miR-128 molecules would affect also its canonical targets resulting in their, *de-novo*, downregulation.

In an effort to find evidences for this scenario, I took advantage of the dataset generated in the Geuvadis Project (Lappalainen *et al.*, 2013). In this work, the RNA of lymphoblastoid cell lines derived from ~450 healthy individuals was sequenced and made publicly available to the scientific community. To understand if the active transcription of LINE1 elements might induce the miR-128 expression as part of a

putative cellular defense mechanism, I compared the expression of the miR-128-1 to the expression levels of LINE1s. Since the previous analyses pointed the attention towards the young full-length LINE1s, I used the expression information of the L1HS and L1PA subfamilies with a length higher than 5000 bp. As negative control for the analysis, I compared the expression of the let-7-a-1, whose expression is known not be induced by the expression of LINE1s (Tristán-Ramos *et al.*, 2020), to the expression levels of LINE1s.

As shown in Figure 19A, the expression of the miR-128-1-5p resulted positively correlated (p-value = 0.000032) to the LINE1 expression levels. The putative passenger 3p strand, instead, resulted negatively correlated (p-value = 0.0012) to the expression levels of LINE1, indicating the possible decay fate of the passenger strand (Ha and Kim, 2014). On the other hand, as already observed in literature (Tristán-Ramos *et al.*, 2020), the expression levels of let-7-a-1 do not correlate with the expression of LINE1 (Figure 19B).

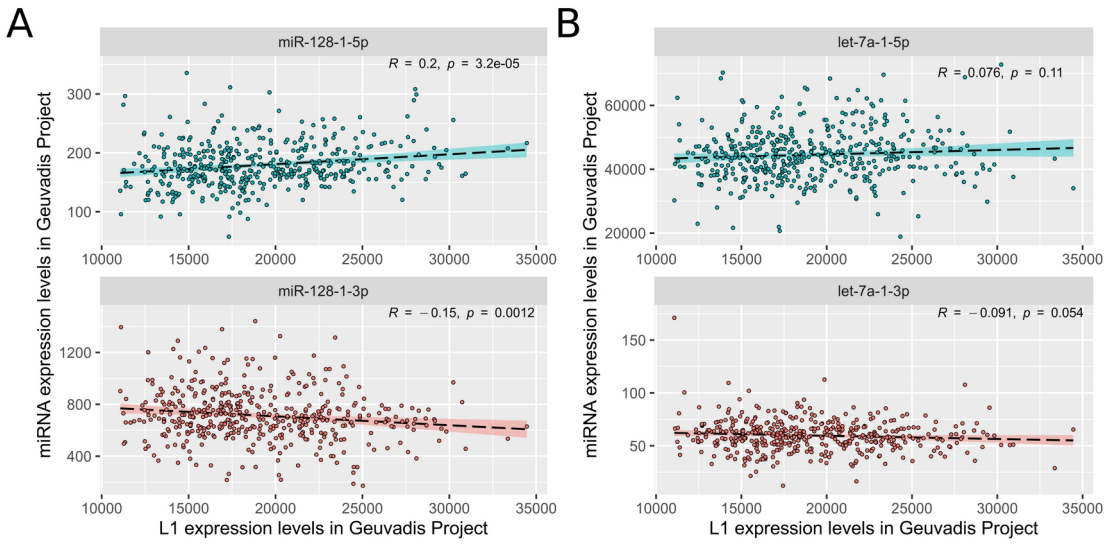

Figure 19: The miR-128 levels are associated to LINE1 transcripts. **A.** Correlation analysis between miR-128-1 (Y-axis) and LINE1 (X-axis) expression levels in Geuvadis dataset. Upon the LINE1 overexpression, the miR-128-1-5p levels concordantly increases probably as part of defense cellular mechanisms. **B.** Correlation analysis between let-7a-1 (Y-axis) and LINE1 (X-axis) expression levels in Geuvadis dataset. No significant associations were found.

These results suggest that, upon the massive activation of autonomous transcription of LINE1 elements, the cells increase the transcription of the miR-128 as part of a possible defense mechanism. So I assume that miR-128 is activated and therefore overexpressed in the DNMT1 KO model with respect to control cells, possibly abolishing the competition because it would be present in excess (Wee *et al.*, 2012). The result of this is shown by analyzing the miR-128 target genes that result mostly downregulated.

4.3 Support for a putative LINE1 ceRNA activity

Until now, I observed that autonomously overexpressed LINE1 might probably act as ceRNA sequestering miRNAs from genes that in the end result upregulated. To add support to this idea, I analyzed the Ardeljan *et al.* (Ardeljan *et al.*, 2020) dataset composed of RNA-seq data derived from Retinal Pigment Epithelial Cells (RPE). In this cellular model, 3 samples overexpressing the codon-optimized LINE1 ORFeus construct (An *et al.*, 2011) were compared to other 3 control samples. The aim of these analyses was to validate the activity of LINE1 as a ceRNA also evaluating a cellular context perturbed exclusively by the overexpression of an LINE1 without affecting DNA methylation.

4.3.1 TE quantification

For demonstrating that the ORFeus-OE model was representing an ideal, even if artificial, experimental setting for autonomous LINE1 transcription, I started to analyze the expression levels of TEs. To do this, I used the consensus-specific tool TEspeX (Ansaloni *et al.*, 2022) for quantifying TE transcription levels. I decided to use this tool since it is able to customize the analysis for allowing also the quantification of a transcript that, due to its artificial nature, is not annotated in the reference transcriptome. From the differential expression analysis, 7 TE subfamilies were upregulated and 2 downregulated. In addition, the strongest (log2FC = 10.92) upregulated TE was specifically the LINE1 ORFeus construct, as visible in Figure 20A.

Confirming the validity of the experiment conducted by the authors (Ardeljan *et al.*, 2020), this analysis demonstrates that these cells strongly overexpress the artificial LINE1 construct. This makes it a suitable condition for evaluating the hypothesis that the LINE1 transcripts might function as ceRNA.

4.3.2 Active LINE1 transcription analysis

For confirming the autonomous transcription of the LINE1 construct, I applied the previously developed methodology. In this case, I used the LINE1 ORFeus sequence as consensus and without considering the genome mapping step in the pipeline, since the artificial construct is not present in the genome.

With this slightly modified procedure I was able to calculate a mean *Inside*/*Outside* ratio of 2.94 in the Control group and a mean of 10.11 in the ORFeus-OE samples. Hence, I observed a significantly (t-test, p-value = 0.00036) higher *Inside*/*Outside* ratio for the ORFeus-OE samples with respect to controls (Figure 20B). These results are thus confirming the autonomous increase in the amount of LINE1 construct transcripts in this experimental setting, representing an ideal environment for testing the LINE1 ceRNA activity. Moreover, the results provide even more reliability to the methodology for discovering events of active LINE1 transcription, considering how the overall experiment was built from the authors (Ardeljan *et al.*, 2020).

4.3.3 miRNA target-sites sharing analysis

Once collected evidence that the LINE1 construct was overexpressed and autonomously transcribed, I tried to investigate if the produced transcripts were able to act as ceRNA. To do this, I started performing the differential gene expression analysis for identifying the deregulated genes. The analysis revealed a huge amount of deregulated genes: 3352 resulting upregulated and 2890 downregulated. Then, following the previous procedure, I identified the miRNA target-sites located both in the LINE1 construct and in the 3'UTRs of protein-coding genes. Upregulated genes shared on average 10.38 miRNA target-sites with the LINE1 construct while the downregulated ones 9.38. Upregulated genes, with respect to downregulated ones, resulted in sharing a significantly higher number of miRNA target-sites with the overexpressed LINE1 construct (Figure 20C, p-value = 0.014).

Also in this case, the findings support the idea that the group of genes that share more miRNA target-sites with the LINE1 construct is undergoing the ceRNA activity. Differently from the DNMT1 model, in this case the LINE1 artificial construct is the

driver of the phenomenon. Once the construct is overexpressed, it is able to sequester miRNA target-sites, causing the upregulation of the gene set.

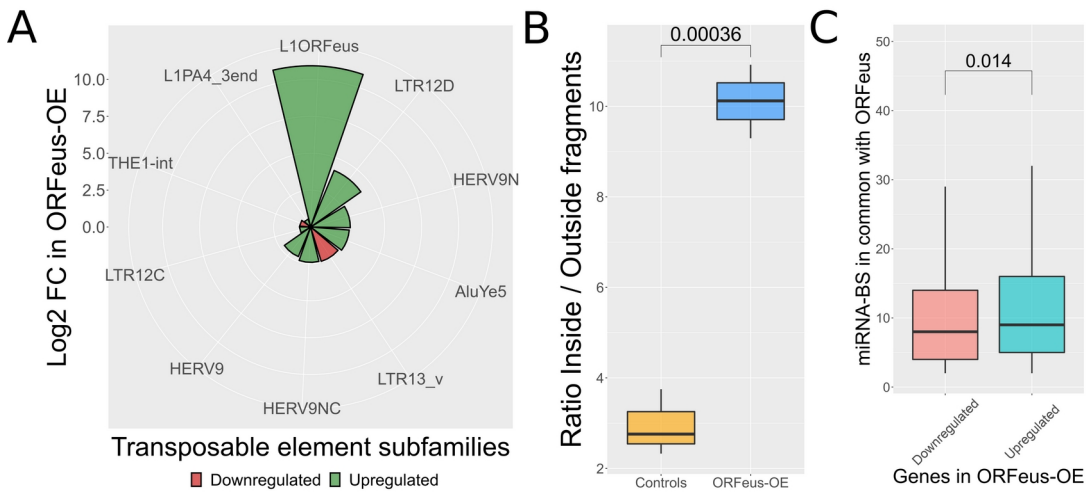

Figure 20: LINE1 could act as ceRNA when artificially overexpressed. **A.** TE subfamilies significantly deregulated in the ORFeus-OE model. In this experiment, the LINE1 construct is the most upregulated TE with a log2 fold-change of 10.92. **B.** Analysis to detect autonomous transcription of LINE1. In ORFeus-OE cells, the upregulation of LINE1 is deriving from an autonomous transcription of the ORFeus element. **C.** Analysis of miRNA target-sites sharing between artificial LINE1 construct and the 3'UTRs of protein-coding genes in ORFeus-OE model. The upregulated genes are sharing a significantly high number of miRNA target-sites with LINE1 construct, reflecting a possible ceRNA activity of the artificial construct.

4.4 ceRNA activity depend on autonomous LINE1 transcription and AGO2

After gathering evidence that LINE1 transcripts may work as ceRNA when overexpressed, I thought to investigate other conditions leading to LINE1 deregulation. To achieve this, I analyzed a cellular model in which Deneault *et al.* (Deneault *et al.*, 2018) sequenced RNA of cells characterized by the KO of ATRX gene first in reprogrammed human induced pluripotent stem cells (iPSC) and then in iPSC differentiated in neuronal cells. ATRX is a chromatin remodeler gene that causes epigenetic modifications in retrotransposons loci (Sadic *et al.*, 2015). Lacking a set of repressive modifications that prevent the aberrant transcription of retransposons, I expect that it might represent another cellular model for investigating the cellular effects of the LINE1 overexpression.

4.4.1 Evaluation of TE deregulation

Aiming to explore TE transcriptomic levels in the ATRX model, I used SQuIRE (Yang *et al.*, 2019) for quantifying the locus-specific expression of TEs. The differential expression analysis revealed a definite deregulation of TEs. Particularly, in iPSC cells the deregulation was more evident with 4490 up and 1217 downregulated TEs while the neuronal counterpart had 1820 up and 540 downregulated elements as shown in Figure 21A. A common feature of both cell lines was that ~50% of upregulated TEs was represented by LINE1 elements: 2438 were the upregulated in iPSC cells and 965 in neurons. As observed for the DNMT1 and ORFeus-OE models, also in this experimental setting the LINE1 is the most upregulated TE family making this experiment an additional model for dissecting the putative LINE1 ceRNA activity. Further investigating the subfamily classification of the upregulated LINE1s, I am able to appreciate that, similarly to the DNMT1 model, in the iPSC experiment ~70% of these elements belong to the L1HS or the L1PA subfamilies. On the other hand, the neurons are characterized by a less pronounced upregulation of elements belonging to these subfamilies (~43%). In addition to this, in both experiments (Figure 21B) the members belonging to these subfamilies have a length that is lower than the canonical 6000 bp full-length size. The results suggest a possible activation of non-autonomous LINE1 transcription in the ATRX model. In order to add further support to this

hypothesis I explored the coverage profile of the LINE1 consensus sequence of the upregulated LINE1s. As shown in Figure 21C, nearly 50% and 25% of the upregulated LINE1 contain the 5'UTR respectively in iPSC and neurons supporting the idea that the observed LINE1 transcription is, at least in part, determined by LINE1 fragments transcription as part of canonical genes. Combining these findings, the ATRX KO is a peculiar model representing a condition of LINE1 overexpression. In iPSC cells, the vast majority of the upregulated LINE1s belong to the young L1HS and L1PA subfamilies, possibly containing a functional LINE1 promoter. These features resemble to the active LINE1 transcription described in the DNMT1 model. A different behaviour is observed for the neuronal cells. Indeed, in this cell line I observed a lower activation of young LINE1 subfamilies as well as a lower representation of their 5'UTR in the expressed sequences. This leads me to speculate that the ATRX KO in neuronal cells is characterized, at least in part, also by the non-autonomous activation of the LINE1 transcription.

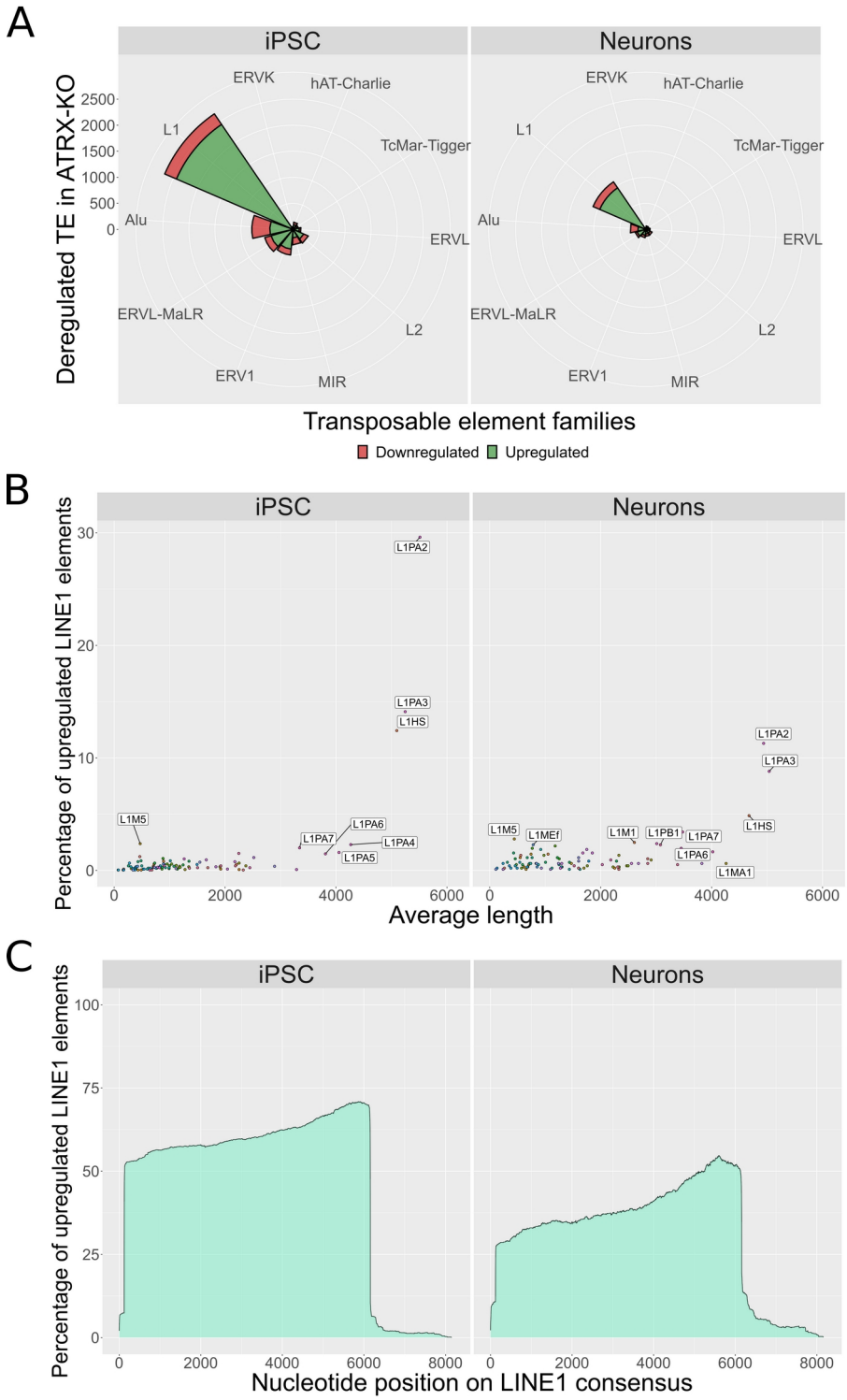

Figure 21: Evaluation of TE transcription upon the ATRX-KO. **A.** Top-10 deregulated TE families in ATRX model. Both in iPSC cells and in neurons, LINE1 family is the most upregulated TE family. **B.** For each upregulated LINE1 subfamily is represented the percentage with respect to the total number of upregulated LINE1 (Y-axis) and the average size of the elements (X-axis). **C.** For each nucleotide position of the LINE1 consensus sequence (X-axis) I report the percentage of the upregulated LINE1s that are covering the part (Y-axis).

4.4.2 Analysis of autonomous LINE1 transcription

In order to determine if the LINE1 upregulation was resulting from a general active or passive transcription of elements, I applied the previously described method (Figure 22A). For both cell lines the paired-end reads were aligned on the LINE1 consensus sequence and on the human reference genome for calculating the *Inside*/*Outside* ratio. In iPSC cells there was a clear difference in the *Inside*/*Outside* ratio between the Control and the ATRX-KO groups (mean Controls = 28.1 and mean ATRX-KO = 47.0). The neuronal cells, instead, were characterized to have a similar ratio between the two different groups (mean Controls = 24.3 and mean ATRX-KO = 27.0). Indeed, while the autonomous LINE1 transcription was confirmed for iPSC cells (t-test, p-value = 0.0013), this was not possible for the neuronal ones (t-test, p-value = 0.77).

These results suggest again that, in iPSC cells, the KO of ATRX causes the autonomous transcription of LINE1 elements. On the other hand, the neuronal cells carrying the ATRX KO might not autonomously overexpress LINE1 elements. As suggested by the previous analyses, it is reasonable that this condition leads to the transcriptional activation of genes that are hosting LINE1 elements, making this experiment quite similar to the CD4+ model.

4.4.3 miRNA target-sites sharing analysis

At this point, the ATRX model was representing two conditions characterized by the autonomous (iPSC) and non-autonomous (neurons) overexpression of LINE1 elements. This peculiarity was particularly intriguing since it gave me the opportunity to investigate the potential ceRNA activity of LINE1 elements in two potentially different deregulation contexts. For this reason, I started to explore the effects of LINE1 transcripts on the ceRNA dynamics by identifying the deregulated protein-coding genes. In iPSC, I found 818 up- and 1441 down-regulated genes while in neurons, they were respectively 319 and 369. Following this, I compared miRNA target-sites between the *active* LINE1s (1181 in iPSC and 479 in neurons) and the 3'UTRs of protein-coding genes. In contrast to DNMT1 and ORFeus-OE models, the downregulated genes were unexpectedly sharing a significantly (p-value < 0.5) higher number of miRNA target-sites with the active LINE1s in both experiments (Figure 22B). Even though at a first look these results seem to go against the hypothesized

model, a possible explanation can be found if I better consider which are the conditions that are needed for appreciating the ceRNA activity of LINE1s. In the neuronal cell line, this evidence might be easily explained since LINE1 transcripts are probably passively transcribed as part of hosting genes, decreasing their capability to sequester miRNAs. For the iPSC experiment, instead, I need a more complex explanation for the comprehension of this result.

4.4.4 AGO2 levels analysis

Current understanding of the ceRNA pathway suggests that the most limiting molecule for the effect is the availability of AGO2 protein with respect to the amount of miRNA molecules. Indeed, it has been shown that a low amount of AGO2 is crucial for the ceRNA effect, while an high amount of it does not give raise to any competition because there is no limitation of molecules for which to compete (Loinger *et al.*, 2012). I therefore decided to investigate the expression levels of AGO2 in a comparative manner between all the considered experiments. To make the expression level of AGO2 comparable among different experiments, I transformed the DESeq2 normalized counts in the corresponding percentile calculated with respect to the distribution of counts for all the transcripts in each analyzed sample. From this analysis, I observed that AGO2 is expressed at a much higher level in the ATRX experiments (~95 percentile) with respect to the DNMT1 and ORFeus (~73 percentile) ones (Figure 22C).

These observations suggest that the reason why the ATRX model does not support a general LINE1 ceRNA effect can be due to the lack of a limiting dose of AGO2. In this case, even if the iPSC cells are overexpressing LINE1s autonomously, the high amount of AGO2 abolishes the competition between the retrotransposons and the canonical protein-coding transcript targets.

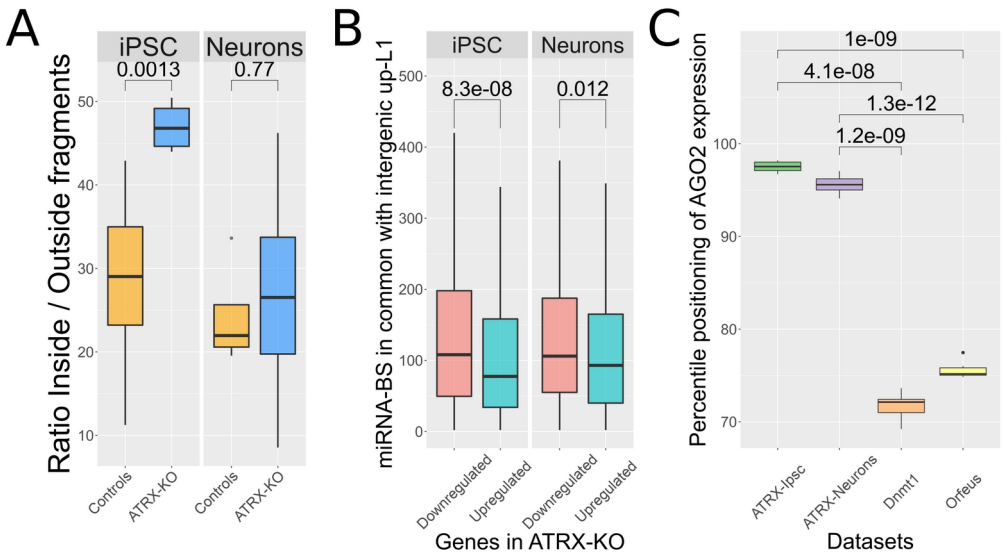

Figure 22: ceRNA activity might depend on autonomous transcription and AGO2. **A.** Analysis to detect autonomous transcription of LINE1 elements in ATRX model. Upon the KO of ATRX, the upregulation of LINE1 elements seems not to derive from an autonomous transcription of elements in neurons. **B.** Analysis of miRNA target-sites sharing between active LINE1s and the 3'UTRs of protein-coding genes in ATRX model. Downregulated genes show an higher number of miRNA target sites in common with overexpressed LINE1s. **C.** Comparative analysis of Ago2 expression levels in all analyzed datasets.

5 Conclusions

LINE1 is the unique TE family containing elements known to be still autonomously mobile in the human genome (Wicker *et al.*, 2007). Although some elements were co-opted for cellular beneficial functions (Kazazian, 2004), a deregulated LINE1 activity can be detrimental to the cells (Rangasamy *et al.*, 2015). As a result, the cells use several mechanisms to regulate LINE1 transcript levels such as the interfering activity of miRNAs (Hamdorf *et al.*, 2015). Beyond this, TEs embedded in lncRNA are supposed to provide miRNA target-sites competing with target genes for the same pool of miRNAs (Wang *et al.*, 2013; Zeng *et al.*, 2019). Despite the increasing interest in this field, it is still lacking a full comprehension of the regulatory layers in which LINE1 transcripts are involved or that regulate their levels and how their deregulation might impact transcriptional programs.

In my PhD project, I reanalyzed RNA-seq data of cells carrying mutations in genes affecting LINE1 transcript levels or overexpressing a specific LINE1 construct. Analyzing the DNMT1 KO model (Jönsson *et al.*, 2019), I confirmed a strong transcriptional activation of LINE1 loci. The upregulated protein-coding genes were enriched to contain LINE1 element fragments especially in their introns and 3'UTRs. This made me speculate that LINE1 transcripts might have a ceRNA activity competing with miRNAs normally targeting genes resulting upregulated. To support this hypothesis, I demonstrated that the upregulated genes, with respect to the downregulated ones, share a higher number of miRNA target-sites with the autonomously transcribed LINE1 elements. This feature is in support of a possible LINE1 ceRNA activity. In my model, overexpressed LINE1 transcripts sharing more miRNA target-sites with the 3'UTR of a given gene group are prone to sequester miRNAs from their canonical target transcripts. As a result, genes from which more miRNAs are sequestered result upregulated. Identifying the putatively sequestered miRNAs, I found a pool of 117 miRNAs including different miRNAs belonging to the let-7 family. This family is experimentally validated to bind the ORF2 of LINE1 transcripts (Tristán-Ramos *et al.*, 2020) supporting the reliability of the experimental procedure for identifying miRNAs sequestered by LINE1s. Analyzing the genes that

should be upregulated upon LINE1 ceRNA activity, I found an enrichment for genes involved in the p53 transcriptional gene network. p53 is a tumor suppressor gene that induces transcriptional programs for responding to a variety of stress signals. Among these, direct (Tiwari *et al.*, 2020) or indirect (Wylie *et al.*, 2016) functions are shown to control the LINE1 activity. The LINE1 ceRNA activity might be a mechanism that induces a cellular response, like p53 targets transcription, against LINE1 elements when these elements result overexpressed. The competition between LINE1 transcripts and canonical transcripts for the binding of a selected pool of miRNAs, might cause the transcriptional activation of genes that coherently work to repress the LINE1 activity, defending in this ways the cell. In this experiment, I also found an enrichment for downregulated genes to share miR-128 targets with LINE1s, which goes against my hypothesis since also this miRNA has been demonstrated to target LINE1s. Nevertheless, my model still hold if I postulate that the expression of miR-128 is increased/induced in DNMT1 KO cells, which is reasonable because the overexpression of miR-128 can justify the increase I have observed in the expression of p53 target-genes (Adlakha and Saini, 2011). In addition, the analysis of Geuvadis dataset clearly show a positive correlation between miR-128 expression and LINE1 transcript levels adding support to my hypothesis.

For testing my scenario in an experimental condition not affected by DNA methylation levels changes, I also analyzed an artificial condition overexpressing a LINE1 construct (Ardeljan *et al.*, 2020). Also in this case, I confirmed that upregulated genes are enriched for genes sharing an higher number of miRNA targets with the overexpressed LINE1 adding further support to a possible LINE1 ceRNA activity.

In an attempt to explore a third study having overexpression of LINE1s, I analyzed an experiment in which the ATRX gene had been knocked out in cultured neurons and iPSC cell lines to build a model of autism (Deneault *et al.*, 2018). In this analysis, my hypothesis did not get support since I found an enrichment for shared miRNA targets among genes downregulated upon ATRX KO. In order to understand whether this result completely invalidates my hypothesis or if there might be a different explanation, I took into account that AGO2 amount is considered to be the limiting factor for which a competition happens in silencing mechanism in the cell. When

AGO2 is in high amount, the competing effect does not manifest because other molecules of the mechanism such as miRNAs are present in high amount (Vickers *et al.*, 2007; Loinger *et al.*, 2012). I therefore compared the amount of AGO2 mRNAs in the experimental conditions under analyses and found that its level in ATRX KO cell model is much and significantly higher with respect to experiments supporting the LINE1 ceRNA hypothesis: the DNMT1 KO and Orfeus construct experiments. In addition to this, another possible limiting factor might reside in the non-autonomous transcription of LINE1 elements. The ceRNA activity might be a phenomenon more pronounced if the autonomous transcription of young LINE1 subfamilies occurs.

In conclusion, I have shown that cellular conditions with a strong autonomous LINE1 transcription are characterized by an higher sharing of miRNA target-sites between overexpressed LINE1s and upregulated genes with respect to downregulated ones. This sharing might be at the basis of a competition for the miRNA targeting. The sequesterion of miRNAs by LINE1 transcripts could then result in the upregulation of a given gene set. Thus, with my PhD project I provide initial evidence in support of the transposon acting as ceRNA (TAC) hypothesis. In this model, the ceRNA activity might even result a way for the cells to trigger defense mechanisms, as the p53 transcriptional program, when the LINE1 transcript levels overcome a certain threshold.

To deepen the TAC hypothesis, future analyses might be carried out on data derived from tumoral samples, notoriously characterized by a deregulated transcription of TEs (Anwar, Wulaningsih and Lehmann, 2017). Crucial might be the integration of miRNome data retrieved from huge databases like The Cancer Genome Atlas (Weinstein *et al.*, 2013), to finely define the group of miRNAs whose activity might be perturbed by the overexpression of LINE1. Once selected the most promising miRNA will be needed to experimentally validate the physical interactions with the LINE1 and the canonical target gene from which the miRNA is sequestered. The reporter gene assay coupled with miRNA mimics or inhibitors is the current gold standard to demonstrate the interactions (Nicolas, 2011). The final competition between the LINE1 and the canonical target gene might be demonstrated with high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP)

(Darnell, 2010) experiments performed in cells overexpressing LINE1. In these experiments, UV irradiations are used to generate crosslinks between RNA, miRNA and AGO. The following AGO2 immunoprecipitation followed by sequencing of the associated RNA might allow the quantification of the associated LINE1 and the canonical target gene highlighting changing in the competition.

Once experimentally validated, I believe that my hypothesis will help future studies for dissecting cellular responses both in developmental (Jachowicz *et al.*, 2017) and pathological (Rangasamy *et al.*, 2015) conditions characterized by transcriptional reactivation of LINE1 elements.

6 Bibliography

Adlakha, Y.K. and Saini, N. (2011) 'MicroRNA-128 downregulates Bax and induces apoptosis in human embryonic kidney cells', *Cellular and Molecular Life Sciences*, 68(8), pp. 1415–1428. Available at: https://doi.org/10.1007/s00018-010-0528-y.

Ala, U. *et al.* (2013) 'Integrated transcriptional and competitive endogenous RNA networks are cross-regulated in permissive molecular environments', *Proceedings of the National Academy of Sciences*, 110(18), pp. 7154–7159. Available at: https://doi.org/10.1073/pnas.1222509110.

Ali, A., Han, K. and Liang, P. (2021) 'Role of Transposable Elements in Gene Regulation in the Human Genome', *Life*, 11(2), p. 118. Available at: https://doi.org/10.3390/life11020118.

An, W. *et al.* (2011) 'Characterization of a synthetic human LINE-1 retrotransposon ORFeus-Hs', *Mobile DNA*, 2(1), p. 2. Available at: https://doi.org/10.1186/1759-8753-2-2.

Anders, S., Pyl, P.T. and Huber, W. (2015) 'HTSeq--a Python framework to work with high-throughput sequencing data', *Bioinformatics (Oxford, England)*, 31(2), pp. 166–169. Available at: https://doi.org/10.1093/bioinformatics/btu638.

Ansaloni, F. *et al.* (2022) 'TEspeX: consensus-specific quantification of transposable element expression preventing biases from exonized fragments', *Bioinformatics*, 38(18), pp. 4430–4433. Available at: https://doi.org/10.1093/bioinformatics/btac526.

Anwar, S.L., Wulaningsih, W. and Lehmann, U. (2017) 'Transposable Elements in Human Cancer: Causes and Consequences of Deregulation', *International Journal of Molecular Sciences*, 18(5), p. 974. Available at: https://doi.org/10.3390/ijms18050974.

Aravin, A.A. *et al.* (2007) 'Developmentally Regulated piRNA Clusters Implicate MILI in Transposon Control', *Science*, 316(5825), pp. 744–747. Available at: https://doi.org/10.1126/science.1142612.

Ardeljan, D. *et al.* (2017) 'The Human Long Interspersed Element-1 Retrotransposon: An Emerging Biomarker of Neoplasia', *Clinical Chemistry*, 63(4), pp. 816–822. Available at: https://doi.org/10.1373/clinchem.2016.257444.

Ardeljan, D. *et al.* (2019) 'LINE-1 ORF2p expression is nearly imperceptible in human cancers', *Mobile DNA*, 11(1), p. 1. Available at: https://doi.org/10.1186/s13100-019-0191-2.

Ardeljan, D. *et al.* (2020) 'Cell fitness screens reveal a conflict between LINE-1 retrotransposition and DNA replication', *Nature Structural & Molecular Biology*, 27(2), pp. 168–178. Available at: https://doi.org/10.1038/s41594-020-0372-1.

Athanikar, J.N., Badge, R.M. and Moran, J.V. (2004) 'A YY1-binding site is required for accurate human LINE-1 transcription initiation', *Nucleic Acids Research*, 32(13), pp. 3846–3855. Available at: https://doi.org/10.1093/nar/gkh698.

Athar, A. *et al.* (2019) 'ArrayExpress update - from bulk to single-cell expression data', *Nucleic acids research*, 47(D1), pp. D711–D715. Available at: https://doi.org/10.1093/nar/gky964.

Auton, A. *et al.* (2015) 'A global reference for human genetic variation', *Nature*, 526(7571), pp. 68–74. Available at: https://doi.org/10.1038/nature15393.

Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data (no date). Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed: 10 June 2022).

Bartel, D.P. (2009) 'MicroRNAs: Target Recognition and Regulatory Functions', *Cell*, 136(2), pp. 215–233. Available at: https://doi.org/10.1016/j.cell.2009.01.002.

Baylin, S.B. and Jones, P.A. (2011) 'A decade of exploring the cancer epigenome — biological and translational implications', *Nature Reviews Cancer*, 11(10), pp. 726–734. Available at: https://doi.org/10.1038/nrc3130.

Becker, K.G. *et al.* (1993) 'Binding of the ubiquitous nuclear transcription factor YY1 to a cis regulatory sequence in the human LINE-1 transposable element', *Human Molecular Genetics*, 2(10), pp. 1697–1702. Available at: https://doi.org/10.1093/hmg/2.10.1697.

Bejerano, G. *et al.* (2006) 'A distal enhancer and an ultraconserved exon are derived from a novel retroposon', *Nature*, 441(7089), pp. 87–90. Available at: https://doi.org/10.1038/nature04696.

Belgnaoui, S.M. *et al.* (2006) 'Human LINE-1 retrotransposon induces DNA damage and apoptosis in cancer cells', *Cancer Cell International*, 6, p. 13. Available at: https://doi.org/10.1186/1475-2867-6-13.

Boeke, J.D. (1997) 'LINEs and Alus — the polyA connection', *Nature Genetics*, 16(1), pp. 6–7. Available at: https://doi.org/10.1038/ng0597-6.

Boland, A. *et al.* (2010) 'Crystal structure and ligand binding of the MID domain of a eukaryotic Argonaute protein', *EMBO reports*, 11(7), pp. 522–527. Available at: https://doi.org/10.1038/embor.2010.81.

Bourgeois, Y. and Boissinot, S. (2019) 'On the Population Dynamics of Junk: A Review on the Population Genomics of Transposable Elements', *Genes*, 10(6), p. 419. Available at: https://doi.org/10.3390/genes10060419.

Braun, J.E. *et al.* (2012) 'A direct interaction between DCP1 and XRN1 couples mRNA decapping to 5′ exonucleolytic degradation', *Nature Structural & Molecular Biology*, 19(12), pp. 1324–1331. Available at: https://doi.org/10.1038/nsmb.2413.

Breiling, A. and Lyko, F. (2015) 'Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond', *Epigenetics & Chromatin*, 8(1), p. 24. Available at: https://doi.org/10.1186/s13072-015-0016-6.

Brouha, B. *et al.* (2003) 'Hot L1s account for the bulk of retrotransposition in the human population', *Proceedings of the National Academy of Sciences*, 100(9), pp. 5280–5285. Available at: https://doi.org/10.1073/pnas.0831042100.

Burns, K.H. (2017) 'Transposable elements in cancer', *Nature Reviews Cancer*, 17(7), pp. 415–424. Available at: https://doi.org/10.1038/nrc.2017.35.

Burns, K.H. (2020) 'Our Conflict with Transposable Elements and Its Implications for Human Disease', *Annual Review of Pathology*, 15, pp. 51–70. Available at: https://doi.org/10.1146/annurev-pathmechdis-012419-032633.

Cardoso, C. *et al.* (2000) 'ATR-X mutations cause impaired nuclear location and altered DNA binding properties of the XNP/ATR-X protein', *Journal of Medical Genetics*, 37(10), pp. 746–751. Available at: https://doi.org/10.1136/jmg.37.10.746.

Castro-Diaz, N. *et al.* (2014) 'Evolutionally dynamic L1 regulation in embryonic stem cells', *Genes & Development*, 28(13), pp. 1397–1409. Available at: https://doi.org/10.1101/gad.241661.114.

Cesana, M. *et al.* (2011) 'A Long Noncoding RNA Controls Muscle Differentiation by Functioning as a Competing Endogenous RNA', *Cell*, 147(2), pp. 358–369. Available at: https://doi.org/10.1016/j.cell.2011.09.028.

Chen, E.Y. *et al.* (2013) 'Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool', *BMC bioinformatics*, 14, p. 128. Available at: https://doi.org/10.1186/1471-2105-14-128.

Chen, T. *et al.* (2003) 'Establishment and Maintenance of Genomic Methylation Patterns in Mouse Embryonic Stem Cells by Dnmt3a and Dnmt3b', *Molecular and Cellular Biology*, 23(16), pp. 5594–5605. Available at: https://doi.org/10.1128/MCB.23.16.5594-5605.2003.

Choi, J., Hwang, S.-Y. and Ahn, K. (2018) 'Interplay between RNASEH2 and MOV10 controls LINE-1 retrotransposition', *Nucleic Acids Research*, 46(4), pp. 1912–1926. Available at: https://doi.org/10.1093/nar/gkx1312.

Chuong, E.B., Elde, N.C. and Feschotte, C. (2017) 'Regulatory activities of transposable elements: from conflicts to benefits', *Nature Reviews Genetics*, 18(2), pp. 71–86. Available at: https://doi.org/10.1038/nrg.2016.139.

Cost, G.J. *et al.* (2002) 'Human L1 element target-primed reverse transcription in vitro', *The EMBO journal*, 21(21), pp. 5899–5910. Available at: https://doi.org/10.1093/emboj/cdf592.

Cost, G.J. and Boeke, J.D. (1998) 'Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA

structure', *Biochemistry*, 37(51), pp. 18081–18093. Available at: https://doi.org/10.1021/bi981858s.

Coufal, N.G. *et al.* (2009) 'L1 Retrotransposition in Human Neural Progenitor Cells', *Nature*, 460(7259), pp. 1127–1131. Available at: https://doi.org/10.1038/nature08248.

Crow, Y.J. *et al.* (2015) 'Characterization of human disease phenotypes associated with mutations in TREX1, RNASEH2A, RNASEH2B, RNASEH2C, SAMHD1, ADAR, and IFIH1', *American Journal of Medical Genetics Part A*, 167(2), pp. 296–312. Available at: https://doi.org/10.1002/ajmg.a.36887.

Crow, Y.J. and Rehwinkel, J. (2009) 'Aicardi-Goutières syndrome and related phenotypes: linking nucleic acid metabolism with autoimmunity', *Human Molecular Genetics*, 18(R2), pp. R130–R136. Available at: https://doi.org/10.1093/hmg/ddp293.

Dai, L. *et al.* (2012) 'Poly(A) binding protein C1 is essential for efficient L1 retrotransposition and affects L1 RNP formation', *Molecular and Cellular Biology*, 32(21), pp. 4323–4336. Available at: https://doi.org/10.1128/MCB.06785-11.

Darnell, R.B. (2010) 'HITS-CLIP: panoramic views of protein-RNA regulation in living cells', *Wiley interdisciplinary reviews. RNA*, 1(2), pp. 266–286. Available at: https://doi.org/10.1002/wrna.31.

De Cecco, M. *et al.* (2013) 'Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues', *Aging*, 5(12), pp. 867–883. Available at: https://doi.org/10.18632/aging.100621.

De Cecco, M. *et al.* (2019) 'L1 drives IFN in senescent cells and promotes age-associated inflammation', *Nature*, 566(7742), pp. 73–78. Available at: https://doi.org/10.1038/s41586-018-0784-9.

De Fazio, S. *et al.* (2011) 'The endonuclease activity of Mili fuels piRNA amplification that silences LINE1 elements', *Nature*, 480(7376), pp. 259–263. Available at: https://doi.org/10.1038/nature10547.

Deininger, P. *et al.* (2017) 'A comprehensive approach to expression of L1 loci', *Nucleic Acids Research*, 45(5), p. e31. Available at: https://doi.org/10.1093/nar/gkw1067.

Deneault, E. *et al.* (2018) 'Complete Disruption of Autism-Susceptibility Genes by Gene Editing Predominantly Reduces Functional Connectivity of Isogenic Human Neurons', *Stem Cell Reports*, 11(5), pp. 1211–1225. Available at: https://doi.org/10.1016/j.stemcr.2018.10.003.

Denli, A.M. *et al.* (2004) 'Processing of primary microRNAs by the Microprocessor complex', *Nature*, 432(7014), pp. 231–235. Available at: https://doi.org/10.1038/nature03049.

Dobin, A. *et al.* (2013) 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics*, 29(1), pp. 15–21. Available at: https://doi.org/10.1093/bioinformatics/bts635.

Dombroski, B.A., Scott, A.F. and Kazazian, H.H. (1993) 'Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element', *Proceedings of the National Academy of Sciences of the United States of America*, 90(14), pp. 6513–6517. Available at: https://doi.org/10.1073/pnas.90.14.6513.

Doucet, A.J. *et al.* (2010) 'Characterization of LINE-1 Ribonucleoprotein Particles', *PLOS Genetics*, 6(10), p. e1001150. Available at: https://doi.org/10.1371/journal.pgen.1001150.

Doucet, A.J. *et al.* (2015) 'A 3′ Poly(A) Tract Is Required for LINE-1 Retrotransposition', *Molecular Cell*, 60(5), pp. 728–741. Available at: https://doi.org/10.1016/j.molcel.2015.10.012.

Ebert, M.S., Neilson, J.R. and Sharp, P.A. (2007) 'MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells', *Nature Methods*, 4(9), pp. 721–726. Available at: https://doi.org/10.1038/nmeth1079.

Ellwanger, D.C. *et al.* (2011) 'The sufficient minimal set of miRNA seed types', *Bioinformatics*, 27(10), pp. 1346–1350. Available at: https://doi.org/10.1093/bioinformatics/btr149.

Elsässer, S.J. *et al.* (2015) 'Histone H3.3 is required for endogenous retroviral element silencing in embryonic stem cells', *Nature*, 522(7555), pp. 240–244. Available at: https://doi.org/10.1038/nature14345.

Evrony, G.D. *et al.* (2012) 'Single-Neuron Sequencing Analysis of L1 Retrotransposition and Somatic Mutation in the Human Brain', *Cell*, 151(3), pp. 483–496. Available at: https://doi.org/10.1016/j.cell.2012.09.035.

Ewing, A.D. and Kazazian, H.H. (2010) 'High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes', *Genome Research*, 20(9), pp. 1262–1270. Available at: https://doi.org/10.1101/gr.106419.110.

Fatemi, M. *et al.* (2002) 'Dnmt3a and Dnmt1 functionally cooperate during de novo methylation of DNA', *European Journal of Biochemistry*, 269(20), pp. 4981–4984. Available at: https://doi.org/10.1046/j.1432-1033.2002.03198.x.

Feng, Q. *et al.* (1996) 'Human L1 Retrotransposon Encodes a Conserved Endonuclease Required for Retrotransposition', *Cell*, 87(5), pp. 905–916. Available at: https://doi.org/10.1016/S0092-8674(00)81997-2.

Feng, Y. *et al.* (2017) 'Deamination-independent restriction of LINE-1 retrotransposition by APOBEC3H', *Scientific Reports*, 7(1), p. 10881. Available at: https://doi.org/10.1038/s41598-017-11344-4.

Feschotte, C. (2008) 'Transposable elements and the evolution of regulatory networks', *Nature Reviews Genetics*, 9(5), pp. 397–405. Available at: https://doi.org/10.1038/nrg2337.

Flasch, D.A. *et al.* (2019) 'Genome-wide de novo L1 Retrotransposition Connects Endonuclease Activity with Replication', *Cell*, 177(4), pp. 837-851.e28. Available at: https://doi.org/10.1016/j.cell.2019.02.050.

Franco-Zorrilla, J.M. *et al.* (2007) 'Target mimicry provides a new mechanism for regulation of microRNA activity', *Nature Genetics*, 39(8), pp. 1033–1037. Available at: https://doi.org/10.1038/ng2079.

Friedman, R.C. *et al.* (2009) 'Most mammalian mRNAs are conserved targets of microRNAs', *Genome Research*, 19(1), pp. 92–105. Available at: https://doi.org/10.1101/gr.082701.108.

García-Cao, M. *et al.* (2004) 'Epigenetic regulation of telomere length in mammalian cells by the Suv39h1 and Suv39h2 histone methyltransferases', *Nature Genetics*, 36(1), pp. 94–99. Available at: https://doi.org/10.1038/ng1278.

Gecz, J. *et al.* (1994) 'Cloning and expression of the murine homologue of a putative human X-linked nuclear protein gene closely linked to PGK1 in Xq13.3', *Human Molecular Genetics*, 3(1), pp. 39–44. Available at: https://doi.org/10.1093/hmg/3.1.39.

Gibbons, R.J. *et al.* (1995) 'Mutations in a putative global transcriptional regulator cause X-linked mental retardation with α-thalassemia (ATR-X syndrome)', *Cell*, 80(6), pp. 837–845. Available at: https://doi.org/10.1016/0092-8674(95)90287-2.

Gibbons, R.J. *et al.* (1997) 'Mutations in transcriptional regulator ATRX establish the functional significance of a PHD-like domain', *Nature Genetics*, 17(2), pp. 146–148. Available at: https://doi.org/10.1038/ng1097-146.

Goldberg, A.D. *et al.* (2010) 'Distinct Factors Control Histone Variant H3.3 Localization at Specific Genomic Regions', *Cell*, 140(5), pp. 678–691. Available at: https://doi.org/10.1016/j.cell.2010.01.003.

Griffiths-Jones, S. *et al.* (2006) 'miRBase: microRNA sequences, targets and gene nomenclature', *Nucleic Acids Research*, 34(Database issue), pp. D140–D144. Available at: https://doi.org/10.1093/nar/gkj112.

Gualandi, N. *et al.* (2022) 'Meta-Analysis Suggests That Intron Retention Can Affect Quantification of Transposable Elements from RNA-Seq Data', *Biology*, 11(6), p. 826. Available at: https://doi.org/10.3390/biology11060826.

Gujar, H., Weisenberger, D.J. and Liang, G. (2019) 'The Roles of Human DNA Methyltransferases and Their Isoforms in Shaping the Epigenome', *Genes*, 10(2), p. 172. Available at: https://doi.org/10.3390/genes10020172.

Ha, M. and Kim, V.N. (2014) 'Regulation of microRNA biogenesis', *Nature Reviews Molecular Cell Biology*, 15(8), pp. 509–524. Available at: https://doi.org/10.1038/nrm3838.

Hafner, A. *et al.* (2019) 'The multiple mechanisms that regulate p53 activity and cell fate', *Nature Reviews Molecular Cell Biology*, 20(4), pp. 199–210. Available at: https://doi.org/10.1038/s41580-019-0110-x.

Hamdorf, M. *et al.* (2015) 'miR-128 represses L1 retrotransposition by binding directly to L1 RNA', *Nature Structural & Molecular Biology*, 22(10), pp. 824–831. Available at: https://doi.org/10.1038/nsmb.3090.

Han, J. *et al.* (2004) 'The Drosha-DGCR8 complex in primary microRNA processing', *Genes & Development*, 18(24), pp. 3016–3027. Available at: https://doi.org/10.1101/gad.1262504.

Hansen, T.B. *et al.* (2013) 'Natural RNA circles function as efficient microRNA sponges', *Nature*, 495(7441), pp. 384–388. Available at: https://doi.org/10.1038/nature11993.

Haoudi, A. *et al.* (NaN/NaN/NaN) 'Retrotransposition-Competent Human LINE-1 Induces Apoptosis in Cancer Cells With Intact *p53*', *BioMed Research International*, 2004, pp. 185–194. Available at: https://doi.org/10.1155/S1110724304403131.

Hauptmann, J. *et al.* (2013) 'Turning catalytically inactive human Argonaute proteins into active slicer enzymes', *Nature Structural & Molecular Biology*, 20(7), pp. 814–817. Available at: https://doi.org/10.1038/nsmb.2577.

Havecker, E.R., Gao, X. and Voytas, D.F. (2004) 'The diversity of LTR retrotransposons', *Genome Biology*, 5(6), p. 225. Available at: https://doi.org/10.1186/gb-2004-5-6-225.

Henckel, A. and Arnaud, P. (2010) 'Genome-wide identification of new imprinted genes', *Briefings in Functional Genomics*, 9(4), pp. 304–314. Available at: https://doi.org/10.1093/bfgp/elq016.

Herbst, H., Sauter, M. and Mueller-Lantzsch, N. (1996) 'Expression of human endogenous retrovirus K elements in germ cell and trophoblastic tumors.', *The American Journal of Pathology*, 149(5), pp. 1727–1735.

Hermann, A., Goyal, R. and Jeltsch, A. (2004) 'The Dnmt1 DNA-(cytosine-C5)-methyltransferase Methylates DNA Processively with High Preference for Hemimethylated Target Sites*', *Journal of Biological Chemistry*, 279(46), pp. 48350–48359. Available at: https://doi.org/10.1074/jbc.M403427200.

Hoelper, D. *et al.* (2017) 'Structural and mechanistic insights into ATRX-dependent and -independent functions of the histone chaperone DAXX', *Nature Communications*, 8(1), p. 1193. Available at: https://doi.org/10.1038/s41467-017-01206-y.

Holmes, S.E. *et al.* (1994) 'A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion', *Nature Genetics*, 7(2), pp. 143–148. Available at: https://doi.org/10.1038/ng0694-143.

Horn, A.V. *et al.* (2017) 'A conserved role for the ESCRT membrane budding complex in LINE retrotransposition', *PLOS Genetics*, 13(6), p. e1006837. Available at: https://doi.org/10.1371/journal.pgen.1006837.

Huang, H.-Y. *et al.* (2020) 'miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database', *Nucleic Acids Research*, 48(D1), pp. D148–D154. Available at: https://doi.org/10.1093/nar/gkz896.

Hubbard, T. *et al.* (2002) 'The Ensembl genome database project', *Nucleic Acids Research*, 30(1), pp. 38–41. Available at: https://doi.org/10.1093/nar/30.1.38.

Hubley, R. *et al.* (2016) 'The Dfam database of repetitive DNA families', *Nucleic Acids Research*, 44(D1), pp. D81-89. Available at: https://doi.org/10.1093/nar/gkv1272.

Huntzinger, E. and Izaurralde, E. (2011) 'Gene silencing by microRNAs: contributions of translational repression and mRNA decay', *Nature Reviews Genetics*, 12(2), pp. 99–110. Available at: https://doi.org/10.1038/nrg2936.

Im, H.-I. and Kenny, P.J. (2012) 'MicroRNAs in neuronal function and dysfunction', *Trends in Neurosciences*, 35(5), pp. 325–334. Available at: https://doi.org/10.1016/j.tins.2012.01.004.

Ivics, Z. *et al.* (1997) 'Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells', *Cell*, 91(4), pp. 501–510. Available at: https://doi.org/10.1016/s0092-8674(00)80436-5.

Jachowicz, J.W. *et al.* (2017) 'LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo', *Nature Genetics*, 49(10), pp. 1502–1510. Available at: https://doi.org/10.1038/ng.3945.

Jo, M.H. *et al.* (2015) 'Human Argonaute 2 Has Diverse Reaction Pathways on Target RNAs', *Molecular Cell*, 59(1), pp. 117–124. Available at: https://doi.org/10.1016/j.molcel.2015.04.027.

Johnson, R. and Guigó, R. (2014) 'The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs', *RNA*, 20(7), pp. 959–976. Available at: https://doi.org/10.1261/rna.044560.114.

Jönsson, M.E. *et al.* (2019) 'Activation of neuronal genes via LINE-1 elements upon global DNA demethylation in human neural progenitors', *Nature Communications*, 10(1), p. 3182. Available at: https://doi.org/10.1038/s41467-019-11150-8.

Jordan, I.K. *et al.* (2003) 'Origin of a substantial fraction of human regulatory sequences from transposable elements', *Trends in genetics: TIG*, 19(2), pp. 68–72. Available at: https://doi.org/10.1016/s0168-9525(02)00006-9.

Kaer, K. *et al.* (2011) 'Intronic L1 Retrotransposons and Nested Genes Cause Transcriptional Interference by Inducing Intron Retention, Exonization and Cryptic Polyadenylation', *PLOS ONE*, 6(10), p. e26099. Available at: https://doi.org/10.1371/journal.pone.0026099.

Kapitonov, V.V., Pavlicek, A. and Jurka, J. (2006) 'Anthology of Human Repetitive DNA', in *Reviews in Cell Biology and Molecular Medicine*. John Wiley & Sons, Ltd. Available at: https://doi.org/10.1002/3527600906.mcb.200300166.

Kapusta, A. *et al.* (2013) 'Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs', *PLOS Genetics*, 9(4), p. e1003470. Available at: https://doi.org/10.1371/journal.pgen.1003470.

Karolchik, D. *et al.* (2004) 'The UCSC Table Browser data retrieval tool', *Nucleic Acids Research*, 32(Database issue), pp. D493-496. Available at: https://doi.org/10.1093/nar/gkh103.

Kawamata, T. and Tomari, Y. (2010) 'Making RISC', *Trends in Biochemical Sciences*, 35(7), pp. 368–376. Available at: https://doi.org/10.1016/j.tibs.2010.03.009.

Kazazian, H.H. *et al.* (1988) 'Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man', *Nature*, 332(6160), pp. 164–166. Available at: https://doi.org/10.1038/332164a0.

Kazazian, H.H. (2004) 'Mobile Elements: Drivers of Genome Evolution', *Science*, 303(5664), pp. 1626–1632. Available at: https://doi.org/10.1126/science.1089670.

Kazazian, H.H. and Moran, J.V. (1998) 'The impact of L1 retrotransposons on the human genome', *Nature Genetics*, 19(1), pp. 19–24. Available at: https://doi.org/10.1038/ng0598-19.

Khvorova, A., Reynolds, A. and Jayasena, S.D. (2003) 'Functional siRNAs and miRNAs Exhibit Strand Bias', *Cell*, 115(2), pp. 209–216. Available at: https://doi.org/10.1016/S0092-8674(03)00801-8.

Kim, D.S. and Hahn, Y. (2011) 'Identification of human-specific transcript variants induced by DNA insertions in the human genome', *Bioinformatics*, 27(1), pp. 14–21. Available at: https://doi.org/10.1093/bioinformatics/btq612.

Kim, Y.-J., Lee, J. and Han, K. (2012) 'Transposable Elements: No More "Junk DNA"', *Genomics & Informatics*, 10(4), pp. 226–233. Available at: https://doi.org/10.5808/GI.2012.10.4.226.

Kim, Y.-K. and Kim, V.N. (2007) 'Processing of intronic microRNAs', *The EMBO Journal*, 26(3), pp. 775–783. Available at: https://doi.org/10.1038/sj.emboj.7601512.

Kines, K.J. *et al.* (2014) 'Potential for genomic instability associated with retrotranspositionally-incompetent L1 loci', *Nucleic Acids Research*, 42(16), pp. 10488–10502. Available at: https://doi.org/10.1093/nar/gku687.

Klein, C.J. *et al.* (2011) 'Mutations in DNMT1 cause hereditary sensory neuropathy with dementia and hearing loss', *Nature Genetics*, 43(6), pp. 595–600. Available at: https://doi.org/10.1038/ng.830.

Kolberg, L. *et al.* (2020) 'gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler', *F1000Research*, 9, p. ELIXIR-709. Available at: https://doi.org/10.12688/f1000research.24956.2.

Kolosha, V.O. and Martin, S.L. (1997) 'In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition', *Proceedings of the National Academy of Sciences*, 94(19), pp. 10155–10160. Available at: https://doi.org/10.1073/pnas.94.19.10155.

Lanciano, S. and Cristofari, G. (2020) 'Measuring and interpreting transposable element expression', *Nature Reviews Genetics*, 21(12), pp. 721–736. Available at: https://doi.org/10.1038/s41576-020-0251-y.

Lander, E.S. *et al.* (2001) 'Initial sequencing and analysis of the human genome', *Nature*, 409(6822), pp. 860–921. Available at: https://doi.org/10.1038/35057062.

Lappalainen, T. *et al.* (2013) 'Transcriptome and genome sequencing uncovers functional variation in humans', *Nature*, 501(7468), pp. 506–511. Available at: https://doi.org/10.1038/nature12531.

Lee, J.Y., Ji, Z. and Tian, B. (2008) 'Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3′-end of genes', *Nucleic Acids Research*, 36(17), pp. 5581–5590. Available at: https://doi.org/10.1093/nar/gkn540.

Lee, R.C. and Ambros, V. (2001) 'An Extensive Class of Small RNAs in Caenorhabditis elegans', *Science*, 294(5543), pp. 862–864. Available at: https://doi.org/10.1126/science.1065329.

Lee, Y. *et al.* (2002) 'MicroRNA maturation: stepwise processing and subcellular localization', *The EMBO Journal*, 21(17), pp. 4663–4670. Available at: https://doi.org/10.1093/emboj/cdf476.

Lei, H. *et al.* (1996) 'De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells', *Development*, 122(10), pp. 3195–3205. Available at: https://doi.org/10.1242/dev.122.10.3195.

Lerat, E. and Capy, P. (1999) 'Retrotransposons and retroviruses: analysis of the envelope gene', *Molecular Biology and Evolution*, 16(9), pp. 1198–1207. Available at: https://doi.org/10.1093/oxfordjournals.molbev.a026210.

Li, F. *et al.* (2019) 'ATRX loss induces telomere dysfunction and necessitates induction of alternative lengthening of telomeres during human cell immortalization', *The EMBO Journal*, 38(19), p. e96659. Available at: https://doi.org/10.15252/embj.201796659.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079. Available at: https://doi.org/10.1093/bioinformatics/btp352.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics (Oxford, England)*, 25(14), pp. 1754–1760. Available at: https://doi.org/10.1093/bioinformatics/btp324.

Li, S. *et al.* (2018) 'Hypomethylation of LINE-1 elements in schizophrenia and bipolar disorder', *Journal of Psychiatric Research*, 107, pp. 68–72. Available at: https://doi.org/10.1016/j.jpsychires.2018.10.009.

Li, W. *et al.* (2012) 'Transposable elements in TDP-43-mediated neurodegenerative disorders', *PloS One*, 7(9), p. e44099. Available at: https://doi.org/10.1371/journal.pone.0044099.

Liang, G. *et al.* (2002) 'Cooperativity between DNA Methyltransferases in the Maintenance Methylation of Repetitive Elements', *Molecular and Cellular Biology*, 22(2), pp. 480–491. Available at: https://doi.org/10.1128/MCB.22.2.480-491.2002.

Liang, J. *et al.* (2020) 'Global changes in chromatin accessibility and transcription following ATRX inactivation in human cancer cells', *FEBS Letters*, 594(1), pp. 67–78. Available at: https://doi.org/10.1002/1873-3468.13549.

Liberti, M.V. and Locasale, J.W. (2016) 'The Warburg Effect: How Does it Benefit Cancer Cells?', *Trends in biochemical sciences*, 41(3), pp. 211–218. Available at: https://doi.org/10.1016/j.tibs.2015.12.001.

Licursi, V. *et al.* (2019) 'MIENTURNET: an interactive web tool for microRNA-target enrichment and network-based analysis', *BMC Bioinformatics*, 20(1), p. 545. Available at: https://doi.org/10.1186/s12859-019-3105-x.

Lingel, A. *et al.* (2003) 'Structure and nucleic-acid binding of the Drosophila Argonaute 2 PAZ domain', *Nature*, 426(6965), pp. 465–469. Available at: https://doi.org/10.1038/nature02123.

Liu, N. *et al.* (2018) 'Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators', *Nature*, 553(7687), pp. 228–232. Available at: https://doi.org/10.1038/nature25179.

Loinger, A. *et al.* (2012) 'Competition between Small RNAs: A Quantitative View', *Biophysical Journal*, 102(8), pp. 1712–1721. Available at: https://doi.org/10.1016/j.bpj.2012.01.058.

Louis, D.N. *et al.* (2016) 'The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary', *Acta Neuropathologica*, 131(6), pp. 803–820. Available at: https://doi.org/10.1007/s00401-016-1545-1.

Love, M.I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15(12), p. 550. Available at: https://doi.org/10.1186/s13059-014-0550-8.

Luning Prak, E.T. and Kazazian, H.H. (2000) 'Mobile elements and the human genome', *Nature Reviews Genetics*, 1(2), pp. 134–144. Available at: https://doi.org/10.1038/35038572.

Lyko, F. (2018) 'The DNA methyltransferase family: a versatile toolkit for epigenetic regulation', *Nature Reviews Genetics*, 19(2), pp. 81–92. Available at: https://doi.org/10.1038/nrg.2017.80.

Mandal, P.K. and Kazazian, H.H. (2008) 'SnapShot: Vertebrate Transposons', *Cell*, 135(1), pp. 192-192.e1. Available at: https://doi.org/10.1016/j.cell.2008.09.028.

Marano, D. *et al.* (2019) 'ATRX Contributes to MeCP2-Mediated Pericentric Heterochromatin Organization during Neural Differentiation', *International Journal of Molecular Sciences*, 20(21), p. 5371. Available at: https://doi.org/10.3390/ijms20215371.

Marasca, F. *et al.* (2022) 'LINE1 are spliced in non-canonical transcript variants to regulate T cell quiescence and exhaustion', *Nature Genetics*, 54(2), pp. 180–193. Available at: https://doi.org/10.1038/s41588-021-00989-7.

María Martín-Núñez, G. *et al.* (2014) 'Type 2 diabetes mellitus in relation to global LINE-1 DNA methylation in peripheral blood: A cohort study', *Epigenetics*, 9(10), pp. 1322–1328. Available at: https://doi.org/10.4161/15592294.2014.969617.

Mariño-Ramírez, L. and Jordan, I.K. (2006) 'Transposable element derived DNaseI-hypersensitive sites in the human genome', *Biology Direct*, 1(1), p. 20. Available at: https://doi.org/10.1186/1745-6150-1-20.

Mathias, S.L. *et al.* (1991) 'Reverse Transcriptase Encoded by a Human Transposable Element', *Science*, 254(5039), pp. 1808–1810. Available at: https://doi.org/10.1126/science.1722352.

Mätlik, K., Redik, K. and Speek, M. (2006) 'L1 Antisense Promoter Drives Tissue-Specific Transcription of Human Genes', *Journal of Biomedicine and Biotechnology*, 2006, p. 71753. Available at: https://doi.org/10.1155/JBB/2006/71753.

Mavragani, C.P. *et al.* (2016) 'Expression of Long Interspersed Nuclear Element 1 Retroelements and Induction of Type I Interferon in Patients With Systemic Autoimmune Disease', *Arthritis & Rheumatology (Hoboken, N.J.)*, 68(11), pp. 2686–2696. Available at: https://doi.org/10.1002/art.39795.

McClintock, B. (1956) 'Controlling Elements and the Gene', *Cold Spring Harbor Symposia on Quantitative Biology*, 21, pp. 197–216. Available at: https://doi.org/10.1101/SQB.1956.021.01.017.

McDowell, T.L. *et al.* (1999) 'Localization of a putative transcriptional regulator (ATRX) at pericentromeric heterochromatin and the short arms of acrocentric chromosomes', *Proceedings of the National Academy of Sciences*, 96(24), pp. 13983–13988. Available at: https://doi.org/10.1073/pnas.96.24.13983.

Memczak, S. *et al.* (2013) 'Circular RNAs are a large class of animal RNAs with regulatory potency', *Nature*, 495(7441), pp. 333–338. Available at: https://doi.org/10.1038/nature11928.

Mills, R.E. *et al.* (2007) 'Which transposable elements are active in the human genome?', *Trends in genetics: TIG*, 23(4), pp. 183–191. Available at: https://doi.org/10.1016/j.tig.2007.02.006.

Mita, P. *et al.* (2018) 'LINE-1 protein localization and functional dynamics during the cell cycle', *eLife*. Edited by S.P. Goff, 7, p. e30058. Available at: https://doi.org/10.7554/eLife.30058.

Mourier, T. *et al.* (2014) 'Transposable elements in cancer as a by-product of stress-induced evolvability', *Frontiers in Genetics*, 5, p. 156. Available at: https://doi.org/10.3389/fgene.2014.00156.

Müller, M., Fazi, F. and Ciaudo, C. (2020) 'Argonaute Proteins: From Structure to Function in Development and Pathological Cell Fate Determination', *Frontiers in Cell and Developmental Biology*, 7. Available at: https://www.frontiersin.org/articles/10.3389/fcell.2019.00360 (Accessed: 1 December 2022).

Muotri, A.R. *et al.* (2010) 'L1 retrotransposition in neurons is modulated by MeCP2', *Nature*, 468(7322), pp. 443–446. Available at: https://doi.org/10.1038/nature09544.

Neri, F. *et al.* (2017) 'Intragenic DNA methylation prevents spurious transcription initiation', *Nature*, 543(7643), pp. 72–77. Available at: https://doi.org/10.1038/nature21373.

Ni, J.Z. *et al.* (2007) 'Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay', *Genes & Development*, 21(6), pp. 708–718. Available at: https://doi.org/10.1101/gad.1525507.

Nicolas, F.E. (2011) 'Experimental Validation of MicroRNA Targets Using a Luciferase Reporter System', in T. Dalmay (ed.) *MicroRNAs in Development: Methods and Protocols*. Totowa, NJ: Humana Press (Methods in Molecular Biology), pp. 139–152. Available at: https://doi.org/10.1007/978-1-61779-083-6_11.

O'Brien, J. *et al.* (2018) 'Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation', *Frontiers in Endocrinology*, 9. Available at: https://www.frontiersin.org/articles/10.3389/fendo.2018.00402 (Accessed: 5 September 2022).

Oettinger, M.A. *et al.* (1990) 'RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination', *Science (New York, N.Y.)*, 248(4962), pp. 1517–1523. Available at: https://doi.org/10.1126/science.2360047.

Okada, C. *et al.* (2009) 'A High-Resolution Structure of the Pre-microRNA Nuclear Export Machinery', *Science*, 326(5957), pp. 1275–1279. Available at: https://doi.org/10.1126/science.1178705.

Pace, J.K. and Feschotte, C. (2007) 'The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage', *Genome Research*, 17(4), pp. 422–432. Available at: https://doi.org/10.1101/gr.5826307.

Parker, J.S., Roe, S.M. and Barford, D. (2004) 'Crystal structure of a PIWI protein suggests mechanisms for siRNA recognition and slicer activity', *The EMBO Journal*, 23(24), pp. 4727–4737. Available at: https://doi.org/10.1038/sj.emboj.7600488.

Pearce, M.S. *et al.* (2012) 'Global LINE-1 DNA methylation is associated with blood glycaemic and lipid profiles', *International Journal of Epidemiology*, 41(1), pp. 210–217. Available at: https://doi.org/10.1093/ije/dys020.

Pezic, D. *et al.* (2014) 'piRNA pathway targets active LINE1 elements to establish the repressive H3K9me3 mark in germ cells', *Genes & Development*, 28(13), pp. 1410–1428. Available at: https://doi.org/10.1101/gad.240895.114.

Picketts, D.J. *et al.* (1996) 'ATRX Encodes a Novel Member of the SNF2 Family of Proteins: Mutations Point to a Common Mechanism Underlying the ATR-X Syndrome', *Human Molecular Genetics*, 5(12), pp. 1899–1907. Available at: https://doi.org/10.1093/hmg/5.12.1899.

Pink, R.C. *et al.* (2011) 'Pseudogenes: Pseudo-functional or key regulators in health and disease?', *RNA*, 17(5), pp. 792–798. Available at: https://doi.org/10.1261/rna.2658311.

Piriyapongsa, J., Mariño-Ramírez, L. and Jordan, I.K. (2007) 'Origin and Evolution of Human microRNAs From Transposable Elements', *Genetics*, 176(2), pp. 1323–1337. Available at: https://doi.org/10.1534/genetics.107.072553.

Poliseno, L. *et al.* (2010) 'A coding-independent function of gene and pseudogene mRNAs regulates tumour biology', *Nature*, 465(7301), pp. 1033–1038. Available at: https://doi.org/10.1038/nature09144.

Protasova, M.S., Andreeva, T.V. and Rogaev, E.I. (2021) 'Factors Regulating the Activity of LINE1 Retrotransposons', *Genes*, 12(10), p. 1562. Available at: https://doi.org/10.3390/genes12101562.

Quinlan, A.R. and Hall, I.M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26(6), pp. 841–842. Available at: https://doi.org/10.1093/bioinformatics/btq033.

Rangasamy, D. *et al.* (2015) 'Activation of LINE-1 Retrotransposon Increases the Risk of Epithelial-Mesenchymal Transition and Metastasis in Epithelial Cancer', *Current Molecular Medicine*, 15(7), pp. 588–597. Available at: https://doi.org/10.2174/1566524015666150831130827.

de Rie, D. *et al.* (2017) 'An integrated expression atlas of miRNAs and their promoters in human and mouse', *Nature Biotechnology*, 35(9), pp. 872–878. Available at: https://doi.org/10.1038/nbt.3947.

Riggs, A.D. (1975) 'X inactivation, differentiation, and DNA methylation', *Cytogenetic and Genome Research*, 14(1), pp. 9–25. Available at: https://doi.org/10.1159/000130315.

Robbez-Masson, L. *et al.* (2018) 'The HUSH complex cooperates with TRIM28 to repress young retrotransposons and new genes', *Genome Research*, 28(6), pp. 836–845. Available at: https://doi.org/10.1101/gr.228171.117.

Rodić, N. *et al.* (2014) 'Long Interspersed Element-1 Protein Expression Is a Hallmark of Many Human Cancers', *The American Journal of Pathology*, 184(5), pp. 1280–1286. Available at: https://doi.org/10.1016/j.ajpath.2014.01.007.

Rodriguez-Martin, B. *et al.* (2020) 'Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition', *Nature Genetics*, 52(3), pp. 306–319. Available at: https://doi.org/10.1038/s41588-019-0562-0.

Ross, R.J., Weiner, M.M. and Lin, H. (2014) 'PIWI proteins and PIWI-interacting RNAs in the soma', *Nature*, 505(7483), pp. 353–359. Available at: https://doi.org/10.1038/nature12987.

Roush, S. and Slack, F.J. (2008) 'The let-7 family of microRNAs', *Trends in Cell Biology*, 18(10), pp. 505–516. Available at: https://doi.org/10.1016/j.tcb.2008.07.007.

Ruby, J.G., Jan, C.H. and Bartel, D.P. (2007) 'Intronic microRNA precursors that bypass Drosha processing', *Nature*, 448(7149), pp. 83–86. Available at: https://doi.org/10.1038/nature05983.

Sadic, D. *et al.* (2015) 'Atrx promotes heterochromatin formation at retrotransposons', *EMBO reports*, 16(7), pp. 836–850. Available at: https://doi.org/10.15252/embr.201439937.

Salmena, L. *et al.* (2011) 'A ceRNA Hypothesis: The Rosetta Stone of a Hidden RNA Language?', *Cell*, 146(3), pp. 353–358. Available at: https://doi.org/10.1016/j.cell.2011.07.014.

Sanchez-Luque, F.J. *et al.* (2019) 'LINE-1 Evasion of Epigenetic Repression in Humans', *Molecular Cell*, 75(3), pp. 590-604.e12. Available at: https://doi.org/10.1016/j.molcel.2019.05.024.

Sassaman, D.M. *et al.* (1997) 'Many human L1 elements are capable of retrotransposition', *Nature Genetics*, 16(1), pp. 37–43. Available at: https://doi.org/10.1038/ng0597-37.

Schirle, N.T., Sheu-Gruttadauria, J. and MacRae, I.J. (2014) 'Structural Basis for microRNA Targeting', *Science (New York, N.Y.)*, 346(6209), pp. 608–613. Available at: https://doi.org/10.1126/science.1258040.

Scott, A.F. *et al.* (1987) 'Origin of the human L1 elements: Proposed progenitor genes deduced from a consensus DNA sequence', *Genomics*, 1(2), pp. 113–125. Available at: https://doi.org/10.1016/0888-7543(87)90003-6.

Smallwood, S.A. and Kelsey, G. (2012) 'De novo DNA methylation: a germ cell perspective', *Trends in Genetics*, 28(1), pp. 33–42. Available at: https://doi.org/10.1016/j.tig.2011.09.004.

Smit, A.F. (1999) 'Interspersed repeats and other mementos of transposable elements in mammalian genomes', *Current Opinion in Genetics & Development*, 9(6), pp. 657–663. Available at: https://doi.org/10.1016/s0959-437x(99)00031-3.

Smit, A.F.A. *et al.* (1995) 'Ancestral, Mammalian-wide Subfamilies of LINE-1 Repetitive Sequences', *Journal of Molecular Biology*, 246(3), pp. 401–417. Available at: https://doi.org/10.1006/jmbi.1994.0095.

Song, J. *et al.* (2011) 'Structure of DNMT1-DNA Complex Reveals a Role for Autoinhibition in Maintenance DNA Methylation', *Science*, 331(6020), pp. 1036–1040. Available at: https://doi.org/10.1126/science.1195380.

Speek, M. (2001) 'Antisense Promoter of Human L1 Retrotransposon Drives Transcription of Adjacent Cellular Genes', *Molecular and Cellular Biology*, 21(6), pp. 1973–1985. Available at: https://doi.org/10.1128/MCB.21.6.1973-1985.2001.

Spengler, R.M., Oakley, C.K. and Davidson, B.L. (2014) 'Functional microRNAs and target sites are created by lineage-specific transposition', *Human Molecular Genetics*, 23(7), pp. 1783–1793. Available at: https://doi.org/10.1093/hmg/ddt569.

Subramanian, S. (2014) 'Competing endogenous RNAs (ceRNAs): new entrants to the intricacies of gene regulation', *Frontiers in Genetics*, 5. Available at: https://www.frontiersin.org/articles/10.3389/fgene.2014.00008 (Accessed: 6 September 2022).

Sultana, T. *et al.* (2019) 'The Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-insertion Sequence Biases and Post-insertion Selection', *Molecular Cell*, 74(3), pp. 555-570.e7. Available at: https://doi.org/10.1016/j.molcel.2019.02.036.

Sumazin, P. *et al.* (2011) 'An Extensive MicroRNA-Mediated Network of RNA-RNA Interactions Regulates Established Oncogenic Pathways in Glioblastoma', *Cell*, 147(2), pp. 370–381. Available at: https://doi.org/10.1016/j.cell.2011.09.041.

Sun, Q., Hao, Q. and Prasanth, K.V. (2018) 'Nuclear Long Noncoding RNAs: Key Regulators of Gene Expression', *Trends in Genetics*, 34(2), pp. 142–157. Available at: https://doi.org/10.1016/j.tig.2017.11.005.

Sundaram, V. *et al.* (2014) 'Widespread contribution of transposable elements to the innovation of gene regulatory networks', *Genome Research*, 24(12), pp. 1963–1976. Available at: https://doi.org/10.1101/gr.168872.113.

Swarts, D.C. *et al.* (2014) 'The evolutionary journey of Argonaute proteins', *Nature Structural & Molecular Biology*, 21(9), pp. 743–753. Available at: https://doi.org/10.1038/nsmb.2879.

Swergold, G.D. (1990) 'Identification, characterization, and cell specificity of a human LINE-1 promoter', *Molecular and Cellular Biology*, 10(12), pp. 6718–6729. Available at: https://doi.org/10.1128/mcb.10.12.6718-6729.1990.

Tarailo-Graovac, M. and Chen, N. (2009) 'Using RepeatMasker to identify repetitive elements in genomic sequences', *Current Protocols in Bioinformatics*, Chapter 4, p. Unit 4.10. Available at: https://doi.org/10.1002/0471250953.bi0410s25.

Tate, P.H. and Bird, A.P. (1993) 'Effects of DNA methylation on DNA-binding proteins and gene expression', *Current Opinion in Genetics & Development*, 3(2), pp. 226–231. Available at: https://doi.org/10.1016/0959-437X(93)90027-M.

Tay, Y., Rinn, J. and Pandolfi, P.P. (2014) 'The multilayered complexity of ceRNA crosstalk and competition', *Nature*, 505(7483), pp. 344–352. Available at: https://doi.org/10.1038/nature12986.

Thomas, C.A. *et al.* (2017) 'Modeling of TREX1-Dependent Autoimmune Disease using Human Stem Cells Highlights L1 Accumulation as a Source of Neuroinflammation', *Cell Stem Cell*, 21(3), pp. 319-331.e8. Available at: https://doi.org/10.1016/j.stem.2017.07.009.

Thomson, D.W. and Dinger, M.E. (2016) 'Endogenous microRNA sponges: evidence and controversy', *Nature Reviews Genetics*, 17(5), pp. 272–283. Available at: https://doi.org/10.1038/nrg.2016.20.

Tiwari, B. *et al.* (2020) 'p53 directly represses human LINE1 transposons', *Genes & Development*, 34(21–22), pp. 1439–1451. Available at: https://doi.org/10.1101/gad.343186.120.

Tristán-Ramos, P. *et al.* (2020) 'The tumor suppressor microRNA let-7 inhibits human LINE-1 retrotransposition', *Nature Communications*, 11(1), p. 5712. Available at: https://doi.org/10.1038/s41467-020-19430-4.

Ulitsky, I. and Bartel, D.P. (2013) 'lincRNAs: Genomics, Evolution, and Mechanisms', *Cell*, 154(1), pp. 26–46. Available at: https://doi.org/10.1016/j.cell.2013.06.020.

Van Meter, M. *et al.* (2014) 'SIRT6 represses LINE1 retrotransposons by ribosylating KAP1 but this repression fails with stress and age', *Nature Communications*, 5(1), p. 5011. Available at: https://doi.org/10.1038/ncomms6011.

Vastenhouw, N.L. and Plasterk, R.H.A. (2004) 'RNAi protects the Caenorhabditis elegans germline against transposition', *Trends in Genetics*, 20(7), pp. 314–319. Available at: https://doi.org/10.1016/j.tig.2004.04.011.

Venkatesh, S. and Workman, J.L. (2015) 'Histone exchange, chromatin structure and the regulation of transcription', *Nature Reviews Molecular Cell Biology*, 16(3), pp. 178–189. Available at: https://doi.org/10.1038/nrm3941.

Verheul, T.C.J. *et al.* (2020) 'The Why of YY1: Mechanisms of Transcriptional Regulation by Yin Yang 1', *Frontiers in Cell and Developmental Biology*, 8. Available at: https://www.frontiersin.org/articles/10.3389/fcell.2020.592164 (Accessed: 2 December 2022).

Vickers, T.A. *et al.* (2007) 'Reduced levels of Ago2 expression result in increased siRNA competition in mammalian cells', *Nucleic Acids Research*, 35(19), pp. 6598–6610. Available at: https://doi.org/10.1093/nar/gkm663.

Voon, H.P.J. *et al.* (2015) 'ATRX Plays a Key Role in Maintaining Silencing at Interstitial Heterochromatic Loci and Imprinted Genes', *Cell Reports*, 11(3), pp. 405–418. Available at: https://doi.org/10.1016/j.celrep.2015.03.036.

Wang, J. *et al.* (2010) 'CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer', *Nucleic Acids Research*, 38(16), pp. 5366–5383. Available at: https://doi.org/10.1093/nar/gkq285.

Wang, Y. *et al.* (2013) 'Endogenous miRNA Sponge lincRNA-RoR Regulates Oct4, Nanog, and Sox2 in Human Embryonic Stem Cell Self-Renewal', *Developmental Cell*, 25(1), pp. 69–80. Available at: https://doi.org/10.1016/j.devcel.2013.03.002.

Wee, L.M. *et al.* (2012) 'Argonaute Divides Its RNA Guide into Domains with Distinct Functions and RNA-Binding Properties', *Cell*, 151(5), pp. 1055–1067. Available at: https://doi.org/10.1016/j.cell.2012.10.036.

Weinstein, J.N. *et al.* (2013) 'The Cancer Genome Atlas Pan-Cancer analysis project', *Nature Genetics*, 45(10), pp. 1113–1120. Available at: https://doi.org/10.1038/ng.2764.

Wicker, T. *et al.* (2007) 'A unified classification system for eukaryotic transposable elements', *Nature Reviews Genetics*, 8(12), pp. 973–982. Available at: https://doi.org/10.1038/nrg2165.

Wilson, A.S., Power, B.E. and Molloy, P.L. (2007) 'DNA hypomethylation and human diseases', *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1775(1), pp. 138–162. Available at: https://doi.org/10.1016/j.bbcan.2006.08.007.

Wylie, A. *et al.* (2016) 'p53 genes function to restrain mobile elements', *Genes & Development*, 30(1), pp. 64–77. Available at: https://doi.org/10.1101/gad.266098.115.

Xie, M. *et al.* (2013) 'Mammalian 5′-Capped MicroRNA Precursors that Generate a Single MicroRNA', *Cell*, 155(7), pp. 1568–1580. Available at: https://doi.org/10.1016/j.cell.2013.11.027.

Xie, S. *et al.* (1999) 'Cloning, expression and chromosome locations of the human DNMT3 gene family', *Gene*, 236(1), pp. 87–95. Available at: https://doi.org/10.1016/S0378-1119(99)00252-8.

Yang, J.-S. *et al.* (2010) 'Conserved vertebrate mir-451 provides a platform for Dicer-independent, Ago2-mediated microRNA biogenesis', *Proceedings of the National Academy of Sciences*, 107(34), pp. 15163–15168. Available at: https://doi.org/10.1073/pnas.1006432107.

Yang, W.R. *et al.* (2019) 'SQuIRE reveals locus-specific regulation of interspersed repeat expression', *Nucleic Acids Research*, 47(5), p. e27. Available at: https://doi.org/10.1093/nar/gky1301.

Yooyongsatit, S. *et al.* (2015) 'Patterns and functional roles of LINE-1 and Alu methylation in the keratinocyte from patients with psoriasis vulgaris', *Journal of Human Genetics*, 60(7), pp. 349–355. Available at: https://doi.org/10.1038/jhg.2015.33.

Zeng, T. *et al.* (2019) 'BACE1-AS prevents BACE1 mRNA degradation through the sequestration of BACE1-targeting miRNAs', *Journal of Chemical Neuroanatomy*, 98, pp. 87–96. Available at: https://doi.org/10.1016/j.jchemneu.2019.04.001.

Zhang, A. *et al.* (2014) 'RNase L restricts the mobility of engineered retrotransposons in cultured human cells', *Nucleic Acids Research*, 42(6), pp. 3803–3820. Available at: https://doi.org/10.1093/nar/gkt1308.

Zhang, H. *et al.* (2004) 'Single Processing Center Models for Human Dicer and Bacterial RNase III', *Cell*, 118(1), pp. 57–68. Available at: https://doi.org/10.1016/j.cell.2004.06.017.

Zhang, R. *et al.* (2019) 'LINE-1 Retrotransposition Promotes the Development and Progression of Lung Squamous Cell Carcinoma by Disrupting the Tumor-Suppressor Gene FGGY', *Cancer Research*, 79(17), pp. 4453–4465. Available at: https://doi.org/10.1158/0008-5472.CAN-19-0076.

Zhang, X. *et al.* (2019) 'Mechanisms and Functions of Long Non-Coding RNAs at Multiple Regulatory Levels', *International Journal of Molecular Sciences*, 20(22), p. 5573. Available at: https://doi.org/10.3390/ijms20225573.

Zhang, X., Zhang, R. and Yu, J. (2020) 'New Understanding of the Relevant Role of LINE-1 Retrotransposition in Human Disease and Immune Modulation', *Frontiers in Cell and Developmental Biology*, 8. Available at: https://www.frontiersin.org/article/10.3389/fcell.2020.00657 (Accessed: 7 June 2022).

Zhang, Z.-M. *et al.* (2015) 'Crystal Structure of Human DNA Methyltransferase 1', *Journal of Molecular Biology*, 427(15), pp. 2520–2531. Available at: https://doi.org/10.1016/j.jmb.2015.06.001.

Zingler, N. *et al.* (2005) 'Analysis of 5′ junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5′-end attachment requiring microhomology-mediated end-joining', *Genome Research*, 15(6), pp. 780–789. Available at: https://doi.org/10.1101/gr.3421505.