SISSA
Scuola
Internazionale
Superiore di
Studi Avanzati

PhD course in Functional and Structural Genomics - Neuroscience Area

Transcriptional dynamics characterising the Zygotic Genome Activation and the Transposable Element expression in Metazoan early embryo

PhD student
Federico Ansaloni

Supervisors
Remo Sanges
Stefano Gustincich

A thesis submitted for the degree of Doctor of Philosophy
in Functional and Structural Genomics

December 2020
Academic year 2019-2020

“Nothing in Biology Makes Sense…

…Except in the Light of Evolution”

[Teodosij Dobžanskij]

Table of contents

Abstract

Although having been considered as junk DNA for long time, nowadays transposable elements are known to play key functional roles in diverse physiological processes. Among these, their involvement in the Metazoan embryogenesis results impressive. Mounting evidence has indeed shown how TEs result remarkably transcribed during the initial phases of the embryonic development of several Metazoan species, including *Drosophila*, mouse and human. However, it is still uncertain whether, in this specific biological context, TE transcription is just the passive consequence of the overall loss of heterochromatic regions occurring at these stages or whether TEs play specific functional roles. Toward this end, after having developed a bioinformatics pipeline capable to quantify the TE expression from RNA-seq datasets, the transcriptional dynamics characterising the TE expression in the early embryos of three different Metazoan species like *C. elegans*, zebrafish and mouse, have been investigated. Importantly, besides defining the TE transcriptional landscapes in the aforementioned species, the functional roles possibly linked to the TE expression in the early embryo have been explored as well as their grade of conservation across Metazoans. Altogether, my results support the evidence that transposable elements actively shape the transcriptional dynamics underlying the embryonic development of all the three analysed species. Importantly, the functions transposable elements play within this context appear to be conserved across different Metazoan species thus suggesting the key role they play in such a crucial biological event as the embryogenesis is.

List of figures

List of tables

List of abbreviations

TE	Transposable Element
TSS	Transcription Start Site
lncRNA	Long Non-Coding RNA
C. elegans	Caenorhabditis elegans
TIR	Terminal Inverted Repeat
CDS	Coding Sequence
TSD	Target Site Duplication
LTR	Long Terminal Repeat
VLP	Virus-Like Particle
RT	Reverse Transcriptase
LINE	Long Interspersed Nuclear Elements
SINE	Short Interspersed Nuclear Elements
UTR	Untranslated Region
ORF	Open Reading Frame
RNP	Ribonucleoprotein Particle
CNS	Central Nervous System
DNMT	DNA methyltransferase
KRAB-ZFP	Krüppel-associated box domain containing zinc-finger proteins
ERV	Endogenous Retroviruses
RNAi	RNA interference
PIWI	P-element induced wimpy testes
piRNA	PIWI-interacting RNAs
MITE	Miniature Inverted Repeats
TAD	Topologically Associated Domains
NPC	Neuronal Precursor Cells
AD	Alzheimer's Disease
ALS	Amyotrophic Lateral Sclerosis
FTD	Frontotemporal Dementia
ZGA	Zygotic Genome Activation
MZT	Maternal to Zygotic Transition
microRNA	miRNA
hpf	Hours Post Fertilisation
hESC	Human Embryonic Stem Cells
EM	Expectation-Maximization

PCA	Principal Component Analysis
RC	Rolling Circle
scRNA-seq	Single-Cell RNA-seq
iPSC	Induced Pluripotent Stem Cells
dsRNA	Double-Stranded RNA
AMP	Anti-Microbial Peptides

Chapter 1
Introduction

1.1 Transposable elements (TEs)

Transposable elements (TEs) are mobile DNA sequences ubiquitously distributed among the eukaryotic genomes (Chuong et al., 2017; Wicker et al., 2007). Through a process called transposition, TEs are able to move from one chromosomal location to another within the same genome (Wicker et al., 2007). TEs were first described in maize, more than 60 years ago by Barbara McClintock (McClintock, 1956). In her pioneering work McClintock described TEs as "normal components of the chromosome responsible for controlling, differentially, the time and type of activity of individual genes". Despite McClintock findings, TEs have been considered as *junk DNA* for long time, having no roles other than replicating themselves. However, the technology advancements occurred in the last 20 years have permitted to extensively investigate TE impact and functions in myriads of eukaryotic genomes. Intriguingly, such large-scale studies have revealed the engagement of a surprisingly huge fraction of TE sequences in a wide range of regulatory processes and molecular interactions, thus confirming McClintock findings (Chuong et al., 2017; Conley et al., 2008; Faulkner et al., 2009; Feschotte, 2008; Kapusta et al., 2013; Sundaram & Wysocka, 2020). As evidence of these observations, approximately 18% and 30% of the murine and human gene transcription start sites (TSSs) have been described to localize within TE-derived sequences (Faulkner et al., 2009). Additionally, approximately 19%, 25% and 35% of zebrafish, murine and human long non-coding RNAs (lncRNAs) sequences are derived from TEs (Kapusta et al., 2013).

Together, these observations highlight the massive contribution TEs have given, and currently give, to the evolution and regulation of coding and non-coding portions of the eukaryotic genomes.

1.1.1 TE genomic occupancy and typologies

Due to their repetitive nature TEs occupy large fractions of eukaryotic genomes. For instance, more than 40% of the zebrafish (*Danio rerio*), mouse (*Mus musculus*) and human (*Homo sapiens*) genomes accounts for TE sequences. Smaller, yet substantial, portions of other model organism genomes are occupied by TEs with approximately 10% and 15% of the *Caenorhabditis elegans* (*C. elegans*) and *Drosophila* (*Drosophila melanogaster*) genomes accounting for TE sequences (**Figure-1.1A**). TEs are broadly classified either as DNA transposons or as retrotransposons depending on the DNA or RNA intermediate exploited to mobilise (Wicker et al., 2007). The two groups are diversely distributed among different eukaryotic genomes. DNA transposons represent the most expanded elements in the *C. elegans* and zebrafish genomes with more than 80% of the TEs being represented by DNA elements. On the contrary, retrotransposons are the most frequent TEs in *Drosophila* and especially in mammalian genomes, as mouse and human, where more than 90% of the TEs are classified as retrotransposons (**Figure-1.1B**).

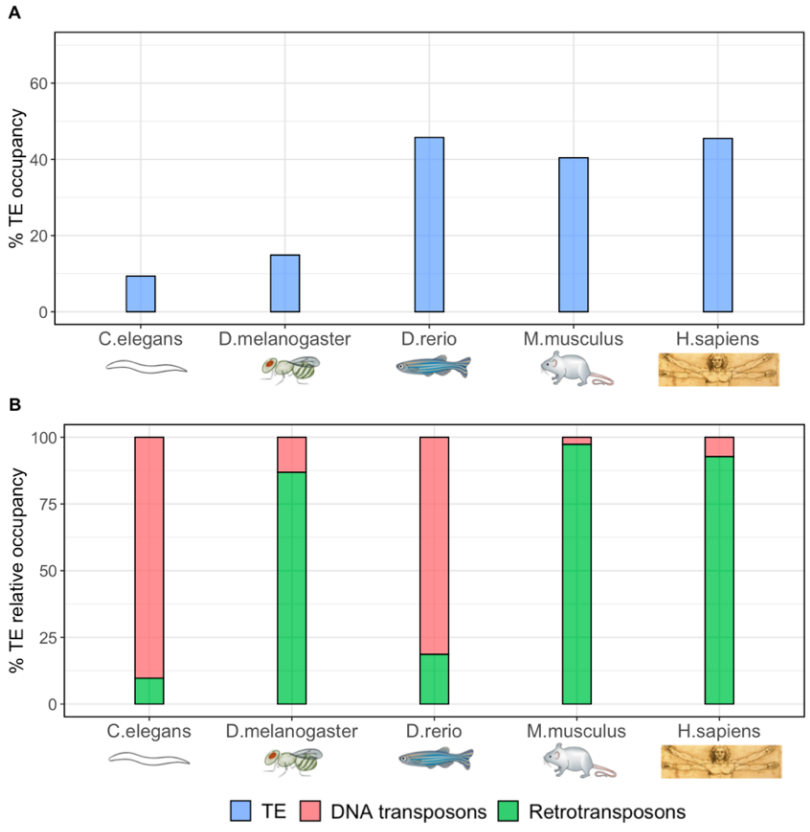


Figure-1.1: TE occupancy in *C. elegans*, *D. melanogaster*, *D. rerio*, *M. musculus*, *H. sapiens* genomes. (**A**) TE occupancy reported as ratio between of number of nucleotides occupied by TEs and the total number of nucleotides composing each genome. (**B**) Total number of TE-derived nucleotides have been classified as belonging to DNA transposons or to retrotransposons.

1.1.1.1 DNA transposons

DNA transposons are mobilised via a DNA intermediate and are subdivided in *cut-and-paste* transposons, *Helitrons* and *Mavericks*. *Cut-and-paste* transposons represent the largest and best-known subgroup of DNA transposons whereas little is known about *Helitrons* and *Mavericks*. *Helitron*s encode DNA helicase and nuclease proteins that mediate the transposition of the element. The mobilisation of the *Helitron*s occurs through a *peel-and-paste* replicative mechanism and exploit a circular DNA intermediate (Grabundzija et al., 2016). *Mavericks* transposons are ~20 kb long elements characterised by two 1 kb long terminal inverted repeats (TIR/ITR). Their transposition is a complex and still not completely understood event involving the synthesis of several self-encoded proteins as polymerases, integrases, proteases and ATPases (V. V. Kapitonov & Jurka, 2006).

Full-length *cut-and-paste* DNA transposons are 1-2 kb long elements composed by two external TIR sequences flanking a mono-cistronic coding sequence (CDS) encoding for a transposase protein (**Figure-1.2A**) (Kazazian, 2011; Muñoz-López & García-Pérez, 2010). The transcription of the element is mediated by the promoter region located at the 5' end. TE-derived mRNA is next moved to the cytoplasm where translation occurs. Once translated, the transposase proteins are shuffled to the nucleus where the transposition process begins with two transposase proteins binding the two TIR regions of the same TE (single-end complex) (**Figure-1.2B, left panel**). Next, the transposon ends are joined through the dimerization of the transposase proteins (paired-end complex) and consequently, the excision takes place (**Figure-1.2B, middle panel**). The excised transposon-transposases complex then binds a different genomic locus (target capture complex) (**Figure-1.2B, right panel**). Next, the transposases perform an overhang cut in the new genomic location inserting the excised transposon. Finally, the genomic gaps remaining as a consequence of the overhang cut are filled by the host genome DNA repair mechanism producing the so-called target site duplication (TSD) (Muñoz-López & García-Pérez, 2010).

Although approximately 1% and 3% of the murine and human genomes account for DNA transposons (Lander et al., 2001; Mouse Genome Sequencing Consortium et al., 2002),

none of these elements appear to be currently active. Indeed, active DNA transposons appear to not have survived a general extinction event that occurred about 40 million years ago in an anthropoid primate ancestor (Feschotte & Pritham, 2007). Anyhow, DNA transposons, like the other TE classes, have played crucial roles in prompting the host cell genome evolution altering gene functions, inducing chromosomal rearrangements and providing sources of coding and non-coding sequences (Feschotte & Pritham, 2007). On the contrary, DNA transposons have been reported to be active in organisms as *C. elegans*, zebrafish and *Drosophila* where numerous active elements have been identified (McCullers & Steiniger, 2017).

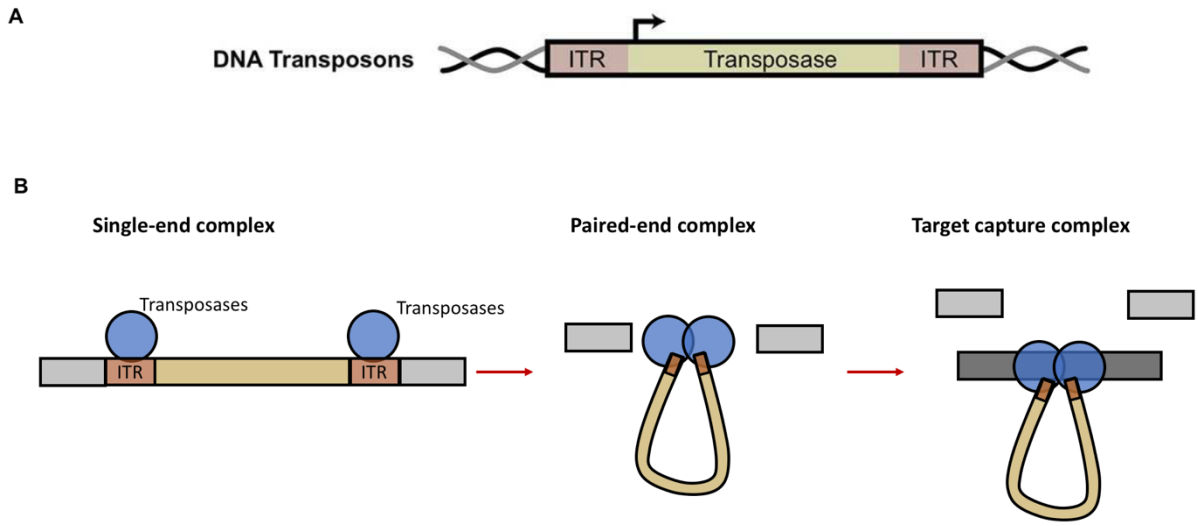


Figure-1.2: DNA transposons.
(**A**) DNA transposons characterised by two TIR/ITR at the 3' and 5' ends, an internal promoter and encoding for a mono-cistronic mRNA. (**B**) DNA transposition mechanism. Left: single-end complex is formed with two transposases binding the element far ends. Middle: transposases dimerization induces the paired-end complex formation and the excision occurs. Right: the excised complex recognises a new genomic location generating the target capture complex and inserting the DNA element in a new locus. ((**A**) adapted from Saleh et al., 2019).

1.1.1.2 Retrotransposons

Retrotransposons are the most expanded TE group in *Drosophila* and mammalian genomes (Haeussler et al., 2019; Lander et al., 2001). Their transposition (now on referred as retrotransposition) is characterised by a replicative *copy-and-paste* mechanism. As a consequence of the retrotransposition event, the original TE locus remains unaltered and a new copy of the element is inserted in a different genomic location. Through the evolution of such replicative mechanism, retrotransposons have massively increased the number of their copies within the host genomes to the point where at least 40% of the human genome is constituted by TE-derived sequences (Lander et al., 2001). According to the sequence features characterising the different elements, retrotransposons are subdivided into two major groups: long terminal repeats (LTR) retrotransposons and non-LTR retrotransposons (Kazazian, 2011).

1.1.1.2.1 LTR retrotransposons

Full-length LTR retrotransposons are 7-9 kb long elements composed by two external long terminal repeats (300-1,000 nucleotide long) flanking an internal protein coding region (**Figure-1.3A**) (Saleh et al., 2019). The two LTR sequences are identical and contain an internal promoter and a polyadenylation signal. The internal protein coding region instead encodes for viral-like proteins needed for the retrotransposition of the element (Chuong et al., 2017; Kazazian, 2011). Intriguingly, the genomic organisation of LTR retrotransposons resembles the exogenous retroviruses ones (Kazazian, 2011; Keegan et al., 2020; Lander et al., 2001). Similar to exogenous retroviruses, full-length LTR retrotransposons encode the *gag* and *pol* proteins but, unlike retroviruses, LTR elements carry no *env* gene, or produce a non-functional form. This makes LTR retrotransposons not capable to mediate inter-cellular spread and infection (Kazazian, 2011; Keegan et al., 2020).

The mobilisation of the LTR retrotransposons begins with the transcription of the element mediated by the promoter sequence located in the 5' LTR region. The 3' LTR portion instead accomplishes the polyadenylation of the transcribed mRNAs (Chuong et al., 2017). The LTR-derived mRNA is next shuttled in the cytoplasm where *gag* and *pol* proteins synthesis occurs. Once encoded, the *gag* protein forms a cytoplasmatic virus-

like particle (VLP). Within the VLP the LTR retrotransposon mRNA is reverse transcribed in a multi-step process driven by the reverse transcriptase (RT) and the RNAse-H enzymes, both encoded by the *pol* gene. The newly synthesized double-stranded DNA is next shuttled to the nucleus where the integrase protein, encoded by the *pol* gene, drives the insertion of the new copy of the element in the host genome (Kazazian, 2011).

LTR retrotransposons are the most expanded TE class in the *Drosophila* genome where 60% of the total TE content accounts for LTR elements. The *Gypsy* group, consisting of 27 families, is the largest group of LTR retrotransposons with few copies retaining retrotransposition competence (Kim et al., 1994; McCullers & Steiniger, 2017). LTR elements also occupy consistent fractions (8-12%) in mouse and human genomes (Lander et al., 2001). Evidence of LTR retrotransposition has been described in the mouse germline (Maksakova et al., 2006) whereas no active LTR elements have been identified in the human genome. Nevertheless, as furtherly discussed in the next paragraphs, a transcriptional bursts initiating within LTR sequences occurs in both mouse and human embryos during the early development (Hendrickson et al., 2017; Pontis et al., 2019).

1.1.1.2.2 non-LTR retrotransposons

Non-LTR retrotransposons are a heterogenous class of TEs particularly expanded within mammalian genomes (Lander et al., 2001; Mouse Genome Sequencing Consortium et al., 2002). Depending on their capability to encode the peptides needed for their own retrotransposition, non-LTR retrotransposons can be further classified as autonomous or non-autonomous elements. Autonomous elements are also defined as long interspersed nuclear elements (LINEs) whereas non-autonomous elements as short interspersed nuclear elements (SINEs) (Kazazian, 2011).

Long interspersed nuclear elements (LINEs)

LINEs are autonomous retrotransposons whose most representative element is the mammalian LINE L1. Full-length LINE L1 elements are ~6 kb long and are characterised by a 5' and 3' untranslated regions (UTR) flanking a protein coding region containing two open reading frames (ORF), ORF1 and ORF2 (**Figure-1.3B**) (Saleh et al., 2019). The 5' UTR contains a bidirectional promoter driving the sense transcription of the ORF1 and ORF2 and, at least in primates, the antisense transcription of the so-called ORF0. Although

the expression of the ORF0 has been linked with an increased activity of the LINE L1 element, the mechanisms by which this occurs are not completely understood (Denli et al., 2015). On the contrary, the role played by ORF1 and ORF2 proteins is much more defined with ORF1 encoding an RNA binding protein and ORF2 a peptide with endonuclease and reverse transcriptase activity. These proteins play a role in driving the retrotransposition of both autonomous and non-autonomous non-LTR elements (Richardson et al., 2014). Importantly, the LINE 3' UTR contains a polyA tail necessary to induce the element mobilisation promoting the interaction between the LINE L1 mRNA and the retrotransposition machinery (Dai et al., 2012).

With almost one million copies the LINE L1 elements cover approximately 17% of the human genome (Lander et al., 2001). LINE L1 is the only active autonomous element retaining the capability to retrotranspose within the human genome with ~100 full-length copies still showing retrotransposition potential (Richardson et al., 2014). Moreover, LINE L1 elements account for large genomic fractions also in the mouse genome (~20%) where LINE L1 represents the most expanded TE family with several copies retaining the capability to retrotranspose (Kazazian, 2011).

Short interspersed nuclear elements (SINEs)

SINE elements are 100-600 nucleotide long non-LTR retrotransposons (**Figure-1.3C**). Differently from LINEs, SINEs are non-autonomous elements thus not encoding the peptidic apparatus required to retrotranspose. Unlike all other TE classes that are transcribed by RNA-polymerase II, SINE elements are transcribed by RNA-polymerase III (Dewannieux et al., 2003). Additionally, SINEs are not ubiquitously distributed among eukaryotic genomes since no SINE elements have been described in *Drosophila* and in monocellular eukaryotes (Kramerov & Vassetzky, 2011).

SINEs are commonly composed by three modules: promoter region, middle body and 3' terminal tail. The SINE promoter region displays similarities with tRNA, 7SL and 5S rRNA. This suggests an evolutionary link between these elements with SINEs likely deriving from pseudogenes of such RNA classes. The middle body portion is a highly heterogenous region. Its origin is not completely understood with different SINE elements being characterised by different body regions. Finally, the 3' terminal tail is mainly composed by simple repeats and, importantly, it is polyadenylated. The SINE polyA plays a crucial

role in the element mobilisation as it competes with the LINE polyA for the retrotransposition machinery self-encoded by LINE elements (Kramerov & Vassetzky, 2011). Thus, SINE polyA is necessary to promote the non-autonomous retrotransposition of SINE elements (Kramerov & Vassetzky, 2011; Richardson et al., 2014; Saleh et al., 2019).

In humans, the most abundant SINE family is composed by primate-specific Alu elements. Alu elements cover almost 11% of the human genome and, with more than one million copies, are the most abundant TEs in humans (Lander et al., 2001). Intriguingly, Alu elements contain cryptic splice sites thus being prone to be captured as alternative exons (Bourque et al., 2018; Lev-Maor et al., 2008; Schmitz & Brosius, 2011). As a consequence, Alu elements are major contributors to new lineage-specific exons in primates (Sorek, 2007).

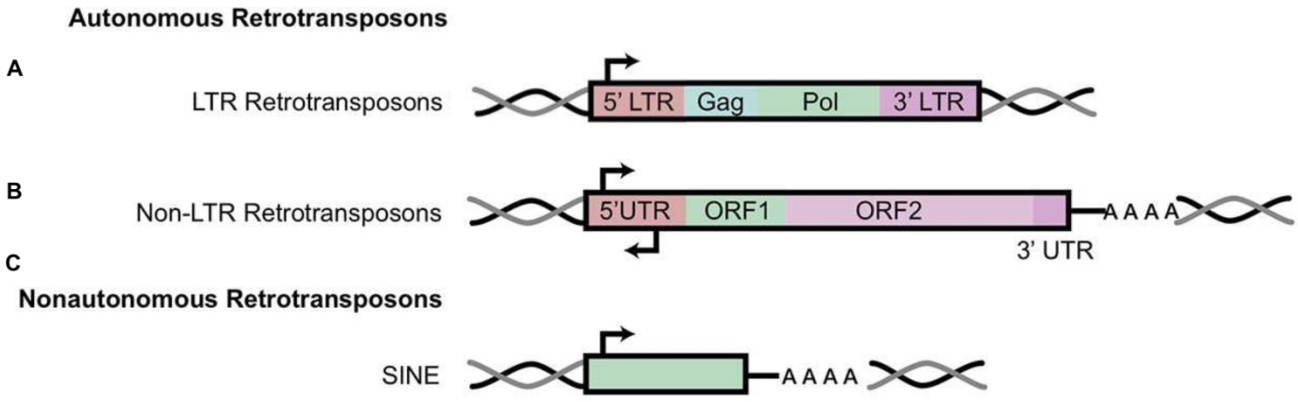


Figure-1.3: transposon classifications.
(**A**) Autonomous LTR retrotransposons characterised by two LTR regions at both the 3' and 5' ends flanking a protein coding region encoding for the *Gag* and *Pol* viral proteins. The transcription of the element is driven by the internal promoter located in 5' LTR. (**B**) Autonomous non-LTR retrotransposons display a 5' and 3' UTR regions flanking two a protein coding region encoding for two peptides. The first has RNA binding activity, the second displays both reverse transcriptase and endonuclease activity. Importantly, the autonomous non-LTR retrotransposon-derived mRNA is polyadenylated (**C**) Non-autonomous retrotransposons (SINEs), show an internal promoter at the 5' and a polyadenylation signal at the 3' end. Adapted from (Saleh et al., 2019).

Retrotransposition of LINE and SINE elements

Non-LTR autonomous (LINE) and non-autonomous (SINE) elements are believed to be the only two TE classes that have retained retrotransposition competence within the human genome (Mills et al., 2007). The mobilisation of both classes of elements relies on the peptidic machinery encoded by the LINEs (Erwin et al., 2014; Richardson et al., 2014). Retrotransposition begins with the nuclear transcription of a full-length LINE L1 element (**Figure-1.4A**). Next, the LINE-derived mRNA is translocated to the cytoplasm where the translation of ORF1 and ORF2 occurs (**Figure-1.4B**). Here, the LINE polyA tail mediates the recognition between the ORF1 and ORF2 proteins and the mRNA of the element itself. Namely, multiple ORF1 proteins and as few as one ORF2 protein bind the LINE mRNA generating the ribonucleoprotein particle (RNP) (**Figure-1.4C**) (Dai et al., 2012; Erwin et al., 2014; Richardson et al., 2014). Alternatively, LINE-encoded proteins can be hijacked by SINE polyA tail and bind this class of elements instead of LINE mRNA (**Figure-1.4D**). Next, the RNP containing the LINE or SINE mRNA and LINE-derived ORF1 and ORF2 proteins is shuttled into the nucleus (**Figure-1.4E**). Here, the ORF2 endonuclease domain generates a single strand nick in the genomic DNA. The single strand genomic DNA exposed at the newly formed nick acts as primer for the ORF2-driven reverse transcription of the LINE/SINE mRNA. Finally, a second genomic DNA cleavage occurs followed by the synthesis of the second filament of the retrotransposed element (**Figure-1.4F**). The retrotransposition process terminates with the insertion, in a new genomic locus, of a retrocopy of the LINE/SINE mRNA originally recognised by the ORF1 and ORF2 proteins (**Figure-1.4G**) (Erwin et al., 2014; Richardson et al., 2014). Nevertheless, as consequence of an incomplete reverse transcription reaction, the LINE retrotransposition process is far from being perfect with almost 30% of the inserted elements truncated at their 5' (Richardson et al., 2014).

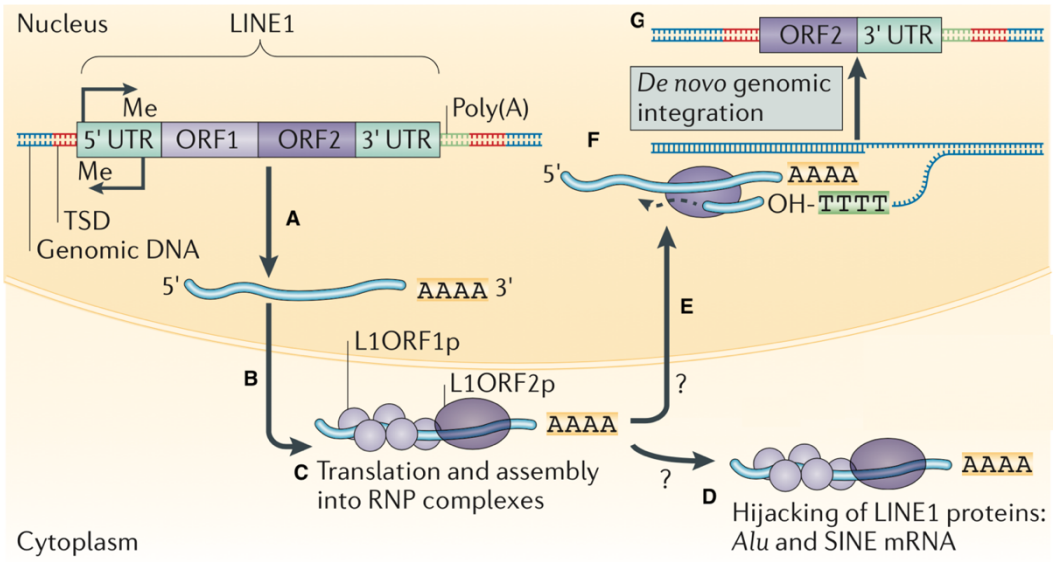


Figure-1.4: LINE/SINE retrotransposition event.
(**A**) Full-length LINE L1 element nuclear transcription driven by the sense promoter located within the element 5' UTR. (**B**) The LINE-derived mRNA is then translocated in the cytoplasm where translation of ORF1 and ORF2 occurs. (**C**) Ribonuclear particle (RNP) is formed through the recognition of ORF1 and ORF2 proteins by LINE polyA tail. (**D**) Alternatively, the LINE proteins may be hijacked by SINE polyA tail. (**E**) RNP is translocated in the nucleus. (**F**) ORF2 protein first generates a single strand nick, then reverse transcribes the LINE/SINE mRNA. Finally, a second genomic DNA cleavage occurs and the second filament of the retrotransposed element is synthetised. (**G**) A new copy of the LINE/SINE element is inserted in a new genomic locus. Adapted from (Erwin et al., 2014).

1.1.2 Host-transposon interaction

The TE expansion within the eukaryotic genomes has necessarily led to the interaction between TEs and the host. TEs have given multiple contributions to the eukaryotic genome evolution. Among them, one of the most remarkable is the impact that TEs, have had on the evolution of the genome size by replicating themselves within the host (Bourque et al., 2018; Petrov, 2002; Schubert & Vu, 2016). Several studies have indeed described how the differential expansion, accumulation and removal of TEs represent a major cause of genome size variation in plants and animals. In this context, species with larger genomes display large TE content and low TE removal rates whereas the opposite scenario is displayed by smaller genome species (Canapa et al., 2015; Gregory, 2005; Kapusta et al., 2017; Schubert & Vu, 2016). The reason why, over the evolution, TEs have massively expanded within eukaryotic genomes is still debated. However, one of the most intriguing explanation is the observation that TEs play crucial roles in the reorganization of the genomes and, through chromosomal rearrangements such as duplications, inversions, and translocations, generate stochastic genetic variability prompting genome evolution (Canapa et al., 2015).

On the other side of the coin, it is certainly true that an uncontrolled TE activation may also negatively interfere with critical physiological pathways of the host. However, in this context, the negative impact of TEs is smoothen by natural selection. TEs displaying excessive activity and having deleterious effects on the host are indeed eliminated by natural selection (Cosby et al., 2019). In this context, it should be considered how natural selection selects for the emergence of host-encoded mechanisms to suppress the activity of such TEs (described below). This will in turn place selective pressure on these TEs making them evolve to escape these mechanisms to avoid extinction (Cosby et al., 2019). This will, once again, pressure the host to evolve further mechanisms to compensate and thus creating a back-and-forth loop where the host and TEs evolve controlling and escaping mechanisms one after the other (Cosby et al., 2019). This process is usually referred as host-TE arm race.

The results of this arm race are various with remarkable examples of how the host has been able to domesticate the inserted TE sequences evolving new genes or gene

functions. One of the most studied examples is the generation of the *RAG1* and *RAG2* genes (Vladimir V. Kapitonov & Jurka, 2005; Vladimir V. Kapitonov & Koonin, 2015). In jawed vertebrates *RAG1* and *RAG2* proteins mediate the V(D)J (variability, diversity and joining) recombination that is required for the generation of highly diverse antigen receptors and thus for the proper functioning of the adaptive immune system (Flajnik, 2014; Litman et al., 2010; Martin et al., 2020). Importantly, both genes, and probably the DNA signals they recognize, were derived from an ancestral DNA transposon around 500 million years ago thus representing an excellent example of how the host has been able to domesticate TE-derived sequences (Bourque et al., 2018; Vladimir V. Kapitonov & Jurka, 2005; Vladimir V. Kapitonov & Koonin, 2015). The examples of TE domestication are nevertheless not limited to the domestication of DNA transposons. Diverse examples showed indeed how the LTR retrotransposon *gag* and *env* genes have been domesticated several times to perform functions in placental development. This process contributes to the host defence against exogenous retroviruses and creates new regulatory modules acting in brain development (Frank & Feschotte, 2017; Naville et al., 2016). Additionally, as extensively described in the next paragraphs, given the propensity of TEs in carrying *cis*-regulatory elements, the host-TE arm race has also led to the co-option of the TE-derived *cis*-regulatory sequences by the host coding and non-coding genes. This has thus generated large regulatory networks where the expression of multiple genes is coordinated by the same TE-derived sequence.

In the next paragraphs, I will discuss how TEs act as a genome-wide source of *cis*-regulatory sequences, the pathways evolved by the host to control the TE activity as well as the impact that TEs have on the host when escaping such pathways in both physiological and pathological conditions.

1.1.2.1 TEs as a genome-wide source of regulatory elements

Autonomous TEs are *selfish* genomic elements encoding exclusively the peptides needed for their own transposition. However, TE mobilisation cannot occur without contribution from the host cell as its transcription depends on the cellular RNA-polymerase II or III. Thus, TEs have evolved *cis* regulatory sequences mimicking the host promoter regions and binding factor sites in order to exploit the host cell transcriptional machinery (Chuong et al., 2017; Feschotte, 2008; Sundaram & Wysocka, 2020). As TEs evolved sequences that mimic the host regulatory elements, it naturally follows that a TE insertion landing nearby a host gene have the potential to strongly interfere with its expression (Chuong et al., 2017). This may have a negative, positive or neutral effect on the host organismal fitness consequently defining the negative or positive selection the TE insertion undergoes. For instance, a *de novo* TE insertion compromising the expression of a gene and decreasing the host fitness is negatively selected and therefore lost (**Figure-1.5A - left**). On the contrary, a *de novo* TE insertion conferring an adaptive function is positively selected and likely maintained throughout the generations (**Figure-1.5A - right**). Finally, a *de novo* TE insertion having a neutral effect generates an intermediate scenario in which the TE sequence accumulates mutations, remaining neutral or acquiring either a deleterious or an adaptive effect (**Figure-1.5A - middle**). Thus, by providing *cis* regulatory sequences, TE insertions occurred in the distant past have been engaged in a wide range of regulatory processes thus surviving through evolution. Moreover, this is of fundamental importance when considering that the majority of TEs occupy the eukaryotic genomes as fragmented and transpositionally inactive elements (**Figure-1.5B**). While a fragmented TE unlikely retains mobilisation competence, it may still maintain functional roles for the *cis* regulatory sequences they carry and therefore be actively involved in the gene regulatory networks of the host.

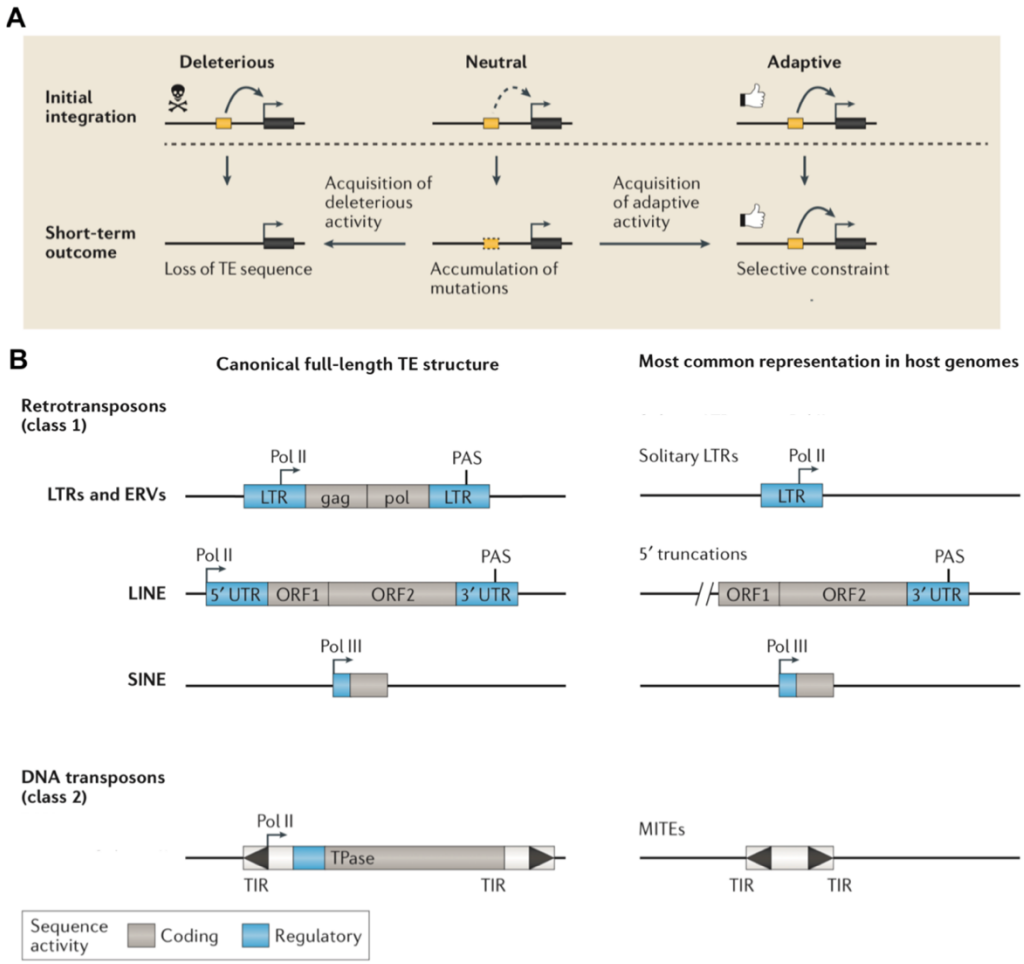


Figure-1.5: fate of *de novo* TE insertions and their most common representation in the host genomes.
(**A**) Evolutionary dynamics characterising the fate of a TE insertion providing deleterious, neutral or adaptive effects on the host fitness. (**B**) Canonical genomic structure of full-length TEs (left) and most common representation within the host genomes (right). LTR elements represented by solitary LTRs as consequence of ectopic recombination between 5' and 3' LTR regions. LINEs characterised by a 5' truncation due to an incomplete reverse transcription reaction and DNA elements represented as miniature inverted repeats (MITEs). Adapted from (Chuong et al., 2017).

Regardless of the full-length or fragmented structure an element is characterised by, TEs regulate the host gene expression through diversified mechanisms. The type of control TEs exert on nearby genes primarily depends on the nature of the *cis* regulatory element/s they carry. TE sequences can thus act as boundary elements, silencers, enhancers and promoters (**Figure-1.6A**). Intriguingly, there seems to be a correlation, at least in mammals, between the regulatory role a TE exerts and the class of transposon the element belongs to. For instance, rodent SINEs have been recently described to act as boundaries between topologically associated domains (TAD) (**Figure-1.6B**) (Kentepozidou et al., 2020). Intriguingly, such SINE elements carry CTCF domains that

mediate the formation of structural chromatin loops. Thus, it has been suggested that, in rodents, SINE elements contribute to the maintenance of clustered CTCF sites at TAD boundaries, promoting the maintenance of the genome organization (Kentepozidou et al., 2020; Sundaram & Wysocka, 2020). LINE L1s, instead, can regulate the expression of nearby genes through the bidirectional promoter contained in the 5' UTR of the element (**Figure-1.6C**) (Sundaram & Wysocka, 2020). LINE L1 sense promoter has been shown to activate the expression of neuronal protein-coding genes in human neuronal progenitor cells in absence of functional DNMT1 (Jönsson et al., 2019). On the contrary, LINE L1 antisense promoter has been described to regulate the expression of several human genes through the generation of chimeric transcripts (Mätlik et al., 2006; Nigumann et al., 2002). Additionally, recent evidence has shown how in mouse early embryo, LINE L1 acts as a scaffold binding transcriptional regulatory elements and acting as chromatin remodeller (**Figure-1.6D**) (Jachowicz et al., 2017; Percharde et al., 2018; Y. Wu et al., 2019). LTR elements, and especially ERVs, appear to be the class of TEs more deeply involved in the mammalian gene regulatory networks. LTRs act as enhancers and promoters in many biological contexts regulating the expression of diverse groups of genes (**Figure-1.6E**). For instance, LTR elements control the expression of interferon-stimulated genes (Chuong et al., 2016), of genes involved with pluripotency maintenance (Sundaram et al., 2017; J. Wang et al., 2014) and of early expressed genes during mouse and human early embryo development (Hendrickson et al., 2017; Macfarlan et al., 2012). Finally, it is worth noticing that providing *cis* regulatory regions appears to be a feature specifically evolved by retrotransposons with few, if any, examples for DNA transposons.

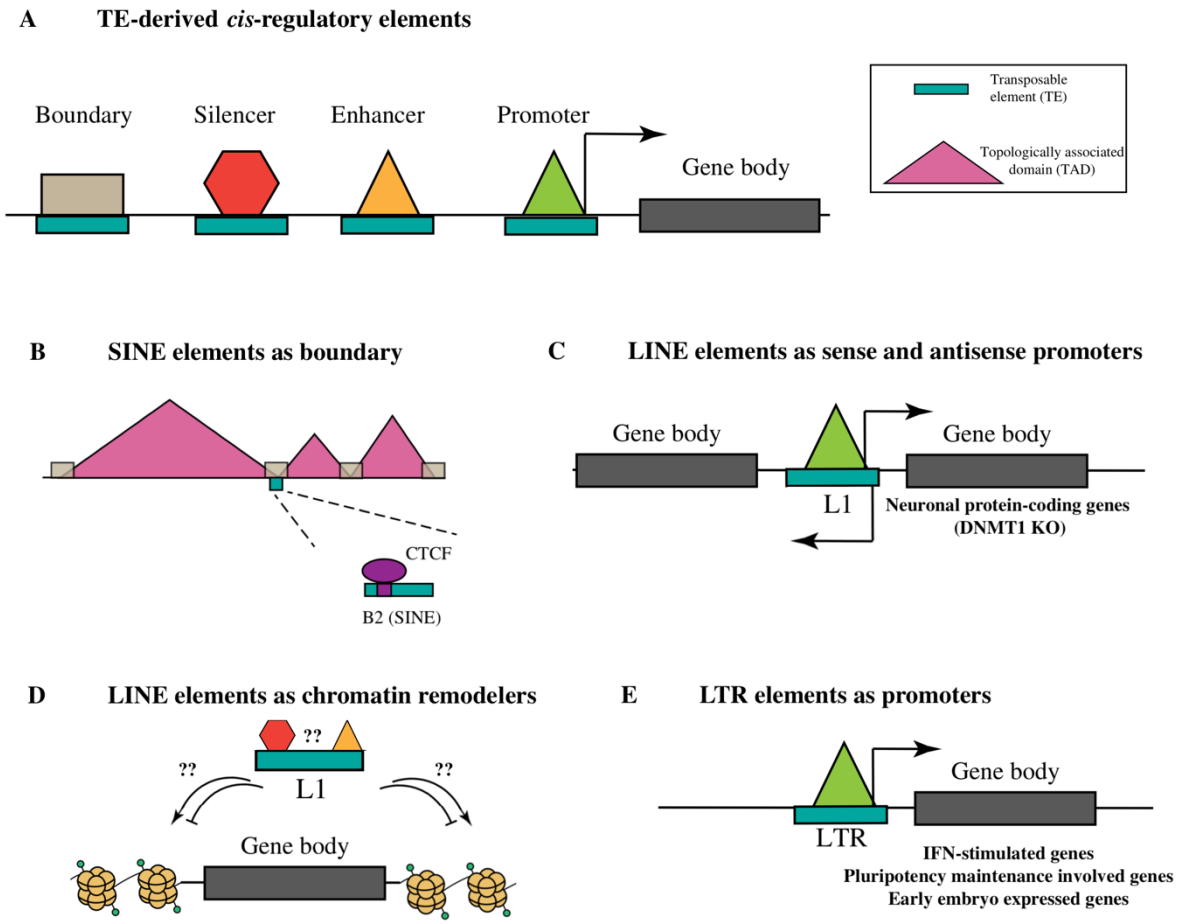


Figure-1.6: TEs as cis regulatory elements in the host genomes.
(**A**) Different type of regulation a TE can exert on the host genome gene expression. (**B**) SINE elements acting as boundaries between topological associated domains (TAD). (**C**) LINE elements driving sense and antisense transcription of nearby genes under specific conditions (e.g. driving expression of neuronal genes in absence of DNMT1). (**D**) LINE elements acting as chromatin remodellers in the mouse early embryo. (**E**) LTR elements driving expression of genes involved in many biological contexts (interferon-stimulated genes, genes involved with the maintenance of pluripotency, early expressed genes in mouse and human early embryos). Adapted from (Sundaram & Wysocka, 2020).

1.1.2.2 TE silencing and controlling pathways

To limit and regulate TE activity, the eukaryote host genomes have evolved different defensive pathways. These regulatory mechanisms act at both transcriptional and post-transcriptional levels and employ DNA methylation, chromatin modification, small RNAs and RNA editing (Jönsson, Garza, Johansson, et al., 2020; Maupetit-Mehouas & Vaury, 2020; Orecchini et al., 2017). Recent findings have described how the lack of functionality of DNA methyltransferase 1 (DNMT1) in human neuronal precursor cells leads to the global loss of methylation in genomic loci occupied by TEs (Jönsson et al., 2019). This induces a transcriptional activation of several TE classes, and especially of LINE L1 elements. This highlights how, in physiological conditions, DNA methylation is a powerful tool to repress TE transcription (Jönsson et al., 2019). Additionally, in mouse and human early embryos and in neuronal precursor cells, TEs, and especially endogenous retroviruses (ERVs), have been shown to be transcriptionally regulated by the epigenetic corepressor protein TRIM28 (Jönsson, Garza, Sharma, et al., 2020; Pontis et al., 2019; Wolf et al., 2020). TRIM28 is recruited to genomic TE loci by the Krüppel-associated box domain containing zinc-finger proteins (KRAB-ZFPs) and attracts a multiprotein complex that establishes transcriptional silencing through the deposition of repressive histone marks (Sripathy et al., 2006). In the germline, TE mobilisation has been historically described to be repressed by small RNAs. The first line of evidence of this mechanism was drawn by the laboratory of Craig Mello in 1999 showing how *C. elegans* mutants defective in RNA interference (RNAi) process exhibited TE mobilisation in the germline (Tabara et al., 1999). Nowadays, small RNAs are well known germline TE repressors also in *Drosophila* and mammals acting through the association with the P-element induced wimpy testes (PIWI) proteins (Aravin et al., 2007). The PIWI-interacting RNAs (piRNA) are single-stranded small non-coding RNAs 21-35 nt-long that, once matured, complex with the PIWI proteins and recognise complementary TE target mRNAs inducing their cleavage and degradation (Aravin et al., 2007). Although the most common mechanism of TE repression mediated by the PIWI-piRNA complexes acts at the post-transcriptional level, PIWI can also function prior to transcription inducing the direct TE transcriptional silencing through DNA methylation and histone modifications (Carmell et al., 2007; Klenov et al., 2014; Russell et al., 2017). Finally, RNA editing can also contribute as an additional TE post-transcriptional control mechanism (Orecchini et al., 2017). The

ADAR1 protein, catalysing the adenosine to inosine (A-to-I) editing on double-stranded RNA, has indeed been described to interact with the human LINE L1 double-stranded RNA. Additionally, in 293T cell lines, the LINE L1 retrotransposition is decreased upon the overexpression of the ADAR1 protein. These two observations together suggest that the ADAR1-mediated A-to-I editing may interfere with the LINE-L1 retrotransposition and be an additional post-transcriptional mechanism regulating the TE activity (Orecchini et al., 2017).

1.1.2.3 Retrotransposons as source of somatic mosaicism in the brain

Until some years ago transposon mobilisation was believed to occur exclusively in the germ cells and in specific cell types such as pluripotent and cancer cells (Erwin et al., 2014). However, in 2005, retrotransposon mobilisation was first described in somatic, non-tumoral, tissues. Indeed, Muotri and colleagues observed human LINE L1 somatic retrotransposition events in rat neuronal precursor cells (NPC) (Muotri et al., 2005). Later on, somatic retrotransposon mobilisation was furtherly described in *Drosophila*, mouse and human neuronal cells (Coufal et al., 2009; Evrony et al., 2012; Perrat et al., 2013). As consequence of somatic retrotransposition, different neuronal cells within the same individual may harbour different TE insertions. Retrotransposition events can thus be a source of genomic variability changing the DNA sequence of single neuron subpopulations and being therefore responsible for the generation of somatic mosaicism within the brain (**Figure-1.7**). However, the magnitude of such events within the host cells is still unclear. Several studies attempted to estimate the frequency of somatic retrotransposition events highlighting a range of potential novel insertions per cell fluctuating from 0.2 to 16.3 (Evrony et al., 2016; Upton et al., 2015). Nevertheless, LINE L1 activity has been described to be involved in diverse physiological processes. In mammals, increased somatic LINE L1 retrotransposition has been observed in response to environmental stimuli such as voluntary exercise (Muotri et al., 2009), lack of maternal care (Bedrosian et al., 2018) and, most importantly, learning conditions (Bachiller et al., 2017). The common link among these processes is the observation that, at the individual-neuron level, retrotransposition may alter synaptic plasticity increasing the number of pathways a neuron can activate in response to a stimulus, whatever the origin of the

stimulus is (Erwin et al., 2014). Thus, retrotransposition can be considered as a stochastic generator of neuronal diversity broadening the variance of cellular phenotypes and thus increasing the types of downstream responses a neuron can activate (Erwin et al., 2014). Finally, it is intriguing to notice how the effects of somatic retrotransposition events cannot be inherited by the progeny as not affecting the germ cells. Thus, eventual benefits deriving from such events are restricted to the subpopulation of cells carrying the insertions. On the other hand, possibly damaging outcomes of retrotransposition events are not passed to the progeny as well and affect exclusively the fitness of the cell/s hosting the TE insertion.

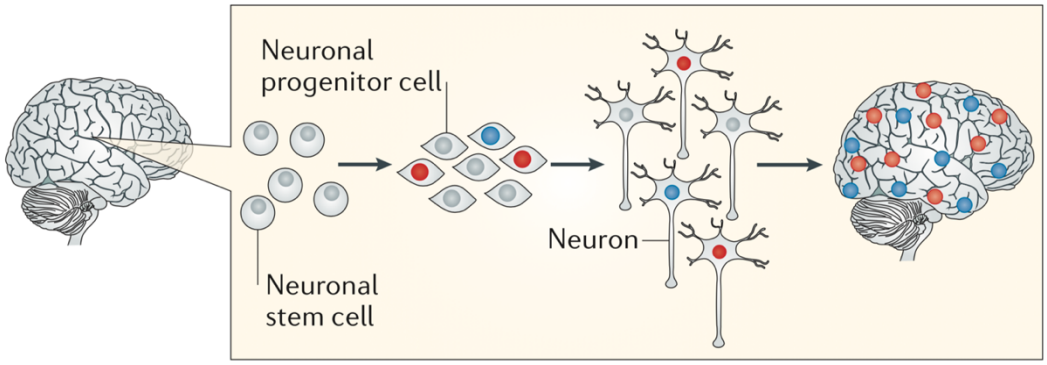


Figure-1.7: retrotransposition driven generation of somatic mosaicism in the brain.
Somatic LINE L1 retrotransposition has been described to occur in neuronal progenitor cells (NPC). Different NPCs may host different retrotransposon insertions (depicted in blue and red) leading to the generation of a genetic mosaicism. Adapted from (Erwin et al., 2014).

1.1.2.4 Retrotransposons in neurodegenerative diseases

Host cells have evolved diverse transcriptional and post transcriptional mechanisms to repress and control TE activation (Jönsson, Garza, Johansson, et al., 2020). However, in pathological conditions such mechanisms may result altered and, consequently, an un-controlled TE activation may occur having detrimental effects on the cell stability. In the last few years, several age-related disorders have been shown to be characterised by an extensive activation of TEs in *Drosophila*, mouse and human patients (Dembny et al., 2020; Guo et al., 2018; Krug et al., 2017; W. Li et al., 2012; Prudencio et al., 2017; Sun et al., 2018; Zhang et al., 2019). For instance, increased TE expression has been observed in Alzheimer's disease (AD) patients as well as in *Drosophila* models where Tau protein overexpression has been shown to induce a heterochromatin loss leading to an un-controlled TE activation (Dembny et al., 2020; Guo et al., 2018; Sun et al., 2018). Similarly, in amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD) patients carrying the C9orf72 expansion impaired heterochromatinization has been associated to an increased TE expression (Prudencio et al., 2017; Y.-J. Zhang et al., 2019). Additionally, mutant TDP-43, characterising both ALS and FTD patients, has been shown to directly interact with TE-derived transcripts thus additionally highlighting a post-transcriptional regulation of TEs in such pathologies (W. Li et al., 2012). Moreover, altered TE expression and mobilisation has also been described in *Drosophila* and murine Hungtington's disease models (Casale et al., 2020). Although different clinical and pathological features distinguish each of the aforementioned disorders, all of them are characterised by the generation of toxic protein aggregates and/or inclusions. Thus, a general mechanism linking toxic protein formation and TE activation has been recently proposed. According to this model, such toxic proteins may alter some of the repressive mechanisms the host cells have evolved to repress TE activation. This leads to a massive and un-regulated TE activation next inducing the generation of DNA damage then followed by a progressive neuronal cell death (Krug et al., 2017).

1.2 The Metazoan embryogenesis

The most fundamental property of evolving systems is their ability to replicate or reproduce (Berh et al., 2002). Reproduction, defined as the generation of viable offspring, is thus a defining process for species survival and evolution being therefore subject to strong selective pressure (Coward & Wells, 2013; Russell et al., 2017). Although the road to reproduction begins with gametogenesis, it reaches a crucial point during fertilisation and embryogenesis (Clift & Schuh, 2013). Fertilisation is defined as the fusion of haploid sperm and egg cells into a totipotent diploid zygote cell. Embryogenesis, instead, consists in the highly dynamic process by which the embryo develops from the totipotent zygote (Alberts et al., 2002). Although both paternal and maternal haploid genomes are necessary for a proper fertilisation and for the subsequent embryo development, the contribution of the two haploid cells is unbalanced (McGrath & Solter, 1984). Indeed, each gamete contributes its haploid genome, but the egg also provides a suitable environment for sperm–egg recognition and supplies all the transcripts and proteins necessary for the initial stages of the zygote development (L. Li et al., 2013; Stitzel & Seydoux, 2007; Zhou & Dean, 2015). Once formed, the totipotent zygote undergoes rapid cell divisions giving rise to the subsequent 2, 4, 8 and further cell stages. Importantly, during the first cell divisions the zygotic genome is transcriptionally quiescent. The development of the embryo is therefore driven by the maternally supplied transcripts and proteins originally deposited into the cytoplasm of the oocyte (Eckersley-Maslin et al., 2018; Schulz & Harrison, 2019). Nevertheless, the zygotic genome activation (ZGA) and the degradation of maternally deposited factors are required for the proper embryo development. ZGA and maternal transcript degradation are part of a broader process, called maternal to zygotic transition (MZT), by which the transcriptional control of the embryo is gradually passed from the maternal transcripts to the zygotic genome.

1.2.1 The Metazoan maternal to zygotic transition (MZT)

During the initial stages, the embryo development is driven by maternally provided factors and the newly formed zygotic genome is transcriptionally quiescent. However, the reasons behind this transcriptional quiescence are not completely understood. The most likely hypothesis suggests that, at least in mammals, a delayed transcriptional activation is necessary to permit the unification of the parental haploid genomes and the transition of the newly formed zygote to a totipotent state (Hamm & Harrison, 2018; Schulz & Harrison, 2019). In absence of transcription from the zygotic genome, maternally supplied transcripts and proteins are thought to play a crucial role in driving these events. Nevertheless, for a successful development, the zygotic genome has to be activated and the maternal products degraded. The transcriptional control of the embryo must thus transit from the mother to the zygote in a process called maternal-to-zygotic transition (MZT). MZT requires the synchronisation of several events as remodelling of the mitotic division cycle, morphological changes and, most importantly, widespread transcriptional activation of the zygotic genome and degradation of a subset of maternal transcripts and proteins (Hamm & Harrison, 2018; Tadros & Lipshitz, 2009). While the general molecular dynamics driving MZT are strikingly conserved across the Metazoans, the timing of these events differs among the species. In rapidly developing species such as *C. elegans*, *Drosophila* and zebrafish MZT takes hours whereas in lowly developing species like mammals it takes days to complete (Jukam et al., 2017; Pálfy et al., 2017). Although MZT is a multi-step process characterised by several events, maternal transcript degradation (clearance) and zygotic genome activation are the two most crucial ones. Importantly, there are no strict boundaries separating these two processes and they should be considered simultaneous, co-existing and strictly interconnected (**Figure-1.8**) (Walser & Lipshitz, 2011).

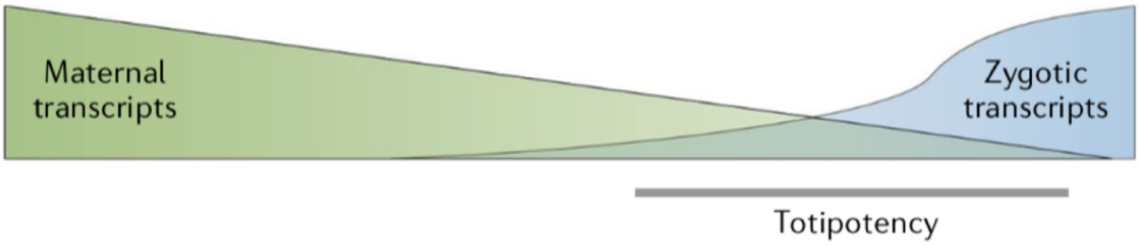


Figure-1.8: maternal to zygotic transition (MZT).
Embryo transcriptional control is passed from the maternally supplied transcripts to the zygotic genome. MZT mainly consists in two simultaneously, co-existing processes: maternal transcript clearance (green) and zygotic genome activation (blue). Adapted from (Schulz & Harrison, 2019).

1.2.1.1 Maternal transcript clearance

During the oogenesis, maternal transcripts and proteins are loaded into the oocyte cytoplasm. After fertilisation, such maternally provided factors represent the primary source of RNAs and proteins of the early embryo due to the transcriptionally quiescence of the newly formed zygotic genome. Their degradation is nevertheless necessary for the embryo to continue developing and it is mediated by the so-called maternal transcript clearance process (Schulz & Harrison, 2019; Zhou & Dean, 2015).

1.2.1.1.1 Molecular mechanisms and factors driving maternal transcript clearance

Maternal transcript clearance represents a cataclysmic event for the cell as it is a wide, systemic and quick process leading to the degradation of a huge number of transcripts in a small-time window. Indeed, maternal mRNAs represent large fractions of the Metazoan protein-coding genomes with ranges going from approximately 40% in *C. elegans* and mouse to 65% in *Drosophila* (Baugh, 2003; De Renzis et al., 2007; Lécuyer et al., 2007; Tadros et al., 2007; Tadros & Lipshitz, 2009; Q. T. Wang et al., 2004). During maternal clearance, 30-40% of these transcripts are eliminated and up to 60% are drastically reduced in abundance in a time window of 1.5-2.5 hours in fast developing species as *Drosophila* and zebrafish and 20 hours in slow developing species as mouse (Baugh, 2003; De Renzis et al., 2007; Hamatani et al., 2004; Pikó & Clegg, 1982; Tadros & Lipshitz, 2009; Thomsen et al., 2010; Walser & Lipshitz, 2011). In species as *C. elegans*, *Drosophila*, *Xenopus* and mouse the degradation of the maternally provided mRNAs is accomplished through a maternal and a zygotic activity. The former is mediated exclusively by maternal factors whereas the latter requires the activation of the zygotic genome as it is driven by zygotically transcribed factors (Sha et al., 2020; Walser & Lipshitz, 2011). The first historical evidence of the molecular mechanisms underlying such event come from zebrafish. In 2006, Giraldez and colleagues described how the zygotically encoded microRNA (miRNA) *miR-430* accelerates the deadenylation and subsequent degradation of several hundred maternal transcripts in the early embryo (Giraldez et al., 2006). miRNAs are 22-24 nucleotide-long RNA molecules targeting the 3' UTR of specific mRNAs and inducing degradation or translation inhibition of the targeted transcripts (Fire et al., 1998). Importantly, miRNAs are widely distributed and conserved among the tree of life.

Following Giraldez observations, the involvement of miRNAs in the maternal transcript clearance was then investigated in other fish and non-fish species. This resulted in the identification of a miRNA-driven maternal mRNA degradation also in non-fish species like *Xenopus* (*Xenopus laevis*) (Lund et al., 2009) and *Drosophila* (Bushati et al., 2008). In *Xenopus* the effector of the degradation of the maternal mRNAs is the ortholog of the zebrafish *miR-430*, the *miR-427* (Lund et al., 2009). In *Drosophila* this role is accomplished by 6 different miRNAs encoded by the miR-309 cluster that do not display any orthology with neither the zebrafish nor the *Xenopus* ones (Bushati et al., 2008). Importantly, the maternal transcript degradation is mediated by the same molecular mechanisms in *Xenopus* and *Drosophila*, as well as in zebrafish. To exert this role, the miRNAs induce the deadenylation of the targeted transcripts leading to their destabilisation and consequent degradation (Bushati et al., 2008; Giraldez et al., 2006; Lund et al., 2009). Moreover, the genomic loci encoding for such miRNAs in the three different species display the same genomic structure being all organised in genic clusters. However, differences exist as *Drosophila* miR-309 cluster encodes 6 different miRNAs targeting different transcripts whereas miRNAs encoded by zebrafish *miR-430* and *Xenopus miR-427* clusters are paralogues genes targeting the same transcripts (Bushati et al., 2008; Giraldez et al., 2006; Lund et al., 2009).

Overall, although no sequence orthology is displayed between *Drosophila* and zebrafish/*Xenopus* maternal clearance effector genes, the degradation mechanism, the structural organization of the involved genomic loci and the biological output are conserved among the three species thus suggesting a strong functional conservation of this mechanism.

1.2.1.1.2 Roles of maternal transcript clearance

The reasons behind the evolution of a mechanism exerting the active degradation of the maternal transcripts are not entirely known. Hypothetically, maternally supplied mRNAs could be passively degraded being diluted in the early embryo cells, cell division after cell division. It is also uncertain whether this process acts as general mechanism to prevent abnormal mRNA dosage in the embryo or whether it specifically eliminates particular maternal transcripts (Tadros & Lipshitz, 2009; Walser & Lipshitz, 2011). The strictly

regulated mechanisms underlying the process suggest the latter hypothesis is the most probable and, in this context, three roles of maternal clearance have been postulated (Walser & Lipshitz, 2011). The first hypothesis suggests that the degradation of maternal mRNAs ubiquitously distributed within the embryo cytoplasm permits the patterned transcription of their zygotic counterparts thus providing spatially and temporally restricted developmental control. In this context, the maternal clearance has a permissive role as its function is to allow the patterned zygotic transcripts to exert their influence (Walser & Lipshitz, 2011). The second hypothesis suggests instead an instructive rather than permissive role of the maternal mRNA degradation. In this scenario, the maternal clearance has been proposed to have a function in regulating the embryo cell cycle length. An evidence of this observation comes from *Drosophila* where increasing or decreasing the dosage of specific maternal transcripts has been associated to an increase or decrease, respectively, in the number of early embryonic mitoses occurring before MZT (Edgar & Datar, 1996). Finally, the third hypothesis proposes that the maternal clearance may have a role in removing transcripts with a function during oogenesis but no longer needed in the embryo (Tadros & Lipshitz, 2009; Walser & Lipshitz, 2011). Importantly, the three speculations assume an overall correlation between mRNA and protein levels that remains to be demonstrated. Globally, all the three speculations have limitations and no absolute conclusion on the roles maternal clearance has can be drawn.

1.2.1.2 Zygotic genome activation (ZGA)

The initial embryonic developmental phases are characterised by a transcriptionally quiescent zygotic genome with maternally supplied transcripts driving the embryo development. However, the activation of the zygotic genome is required for the embryo to continue developing beyond these stages (Schulz & Harrison, 2019). This process occurs through the so-called zygotic genome activation (ZGA). ZGA represents the process by which the embryo is gradually taken from a transcriptional quiescent to a transcriptional active state characterised by the expression of thousands of genes. ZGA is a highly conserved process and it is defined by the same molecular events among the Metazoans despite being driven by different players and taking different timings to be completed (Schulz & Harrison, 2019).

1.2.1.2.1 Molecular mechanisms and factors driving ZGA

The initial phases of the ZGA relies on maternally supplied transcription factors that by binding to specific DNA sequences direct the cell transcriptional machinery on particular genomic loci. Such genome activator factors are stored in the embryo cytoplasm and, prior to ZGA, their translation is repressed in order to prevent a premature activation of the zygotic genome. Once the embryo is mature to support its own transcription, the translation of these factors occurs and the synthesised peptides drive the awakening of the zygotic genome activating the transcription of specific loci (Jukam et al., 2017; Lee et al., 2014; Schulz & Harrison, 2019). The activation of the zygotic genome occurs through two transcriptional waves, a minor and a major one (**Figure-1.9A and B**) (Eckersley-Maslin et al., 2018; Schulz & Harrison, 2019). The first transcriptional wave, the minor one, leads to the activation of a specific subset of zygotically transcribed genes. These transcripts represent the first mRNAs to be actively expressed by the zygotic genome and are indeed involved in the accomplishment of primary functions within the developing embryo. For instance, in zebrafish, *Xenopus* and *Drosophila* the first genes to be expressed by the zygotic genome are miRNAs involved in the degradation of the maternal transcripts, process required for a proper embryonic development (Bushati et al., 2008; Giraldez et al., 2006; Lund et al., 2009). The minor wave is next followed by a second, major, broad transcriptional wave that leads to the complete activation of the zygotic genome. Once this second transcriptional event is completed, the zygotic genome is fully

activated and self-sufficient to support its own transcription. Importantly, the transcriptional bursts characterising the early embryo are not the only molecular changes occurring in this context. Indeed, the early embryo is also characterised by multiple levels of chromatin reorganisation as DNA methylation, histone modifications and changes in chromatin accessibility and nucleosome positioning (**Figure-1.9A and C**) (Schulz & Harrison, 2019). Nevertheless, it is unclear whether the changes in chromatin are required for the ZGA or whether the ZGA is instructive to the chromatin changes (Schulz & Harrison, 2019).

Although ZGA dynamics and mechanisms are functionally conserved among the Metazoan, the genome activator factors driving the activation of the zygotic genome appear to be species-specific. *Drosophila* genome activator, Zelda, was the first to be described in 2008 (Liang et al., 2008). Zelda encodes a zinc-finger protein that has been reported to promote the expression of approximately 120 genes, including the miR-309 transcripts involved with the maternal transcript degradation (Bushati et al., 2008; Liang et al., 2008). Zelda orthologs are restricted to the insect clade and thus, its sequence is not informative for the definition of genome activators in other species. However, in 2013, two independent studies identified *nanog*, *soxB1* family and *pou5f1* as zebrafish genome activators (Lee et al., 2013; Leichsenring et al., 2013). Intriguingly, the three factors activate several hundreds of zygotic genes including the *miR-430* genes involved in the maternal transcripts degradation (Lee et al., 2013). Thus, in both *Drosophila* and zebrafish the first genes to be transcribed by the zygotic genome appear to be functionally conserved as being factors involved in the degradation of the maternal transcripts. Only recently, mouse and human genome activators have been identified by Hendrickson and colleagues as the orthologs genes *Dux* and *DUX4* (Hendrickson et al., 2017). Intriguingly, *Dux* and *DUX4* target genes are mostly conserved among the two species with the zinc-finger transcription factor *Zscan4*, the *Kdm4* histone demethylases family and the *Pramef* gene family resulting among the earliest expressed *Dux*/*DUX4* targets (Hendrickson et al., 2017). In mouse and human no zygotically expressed factors involved with maternal transcript degradation have been described yet and thus whether *Dux*/*DUX4* activates the expression of proteins involved in this pathway, as shown in *Drosophila* and zebrafish, is still unclear.

Overall, although no sequence conservation is displayed among *Drosophila* Zelda, zebrafish *nanog*, *soxB1* and *pou5f1* and mammals *Dux*/*DUX4*, the molecular mechanisms and pathways leading to the activation of the zygotic genome seem to be functionally conserved among the Metazoans. Thus, as already observed for the maternal clearance process, the players involved within ZGA appear to be functionally, yet not evolutionary, conserved.

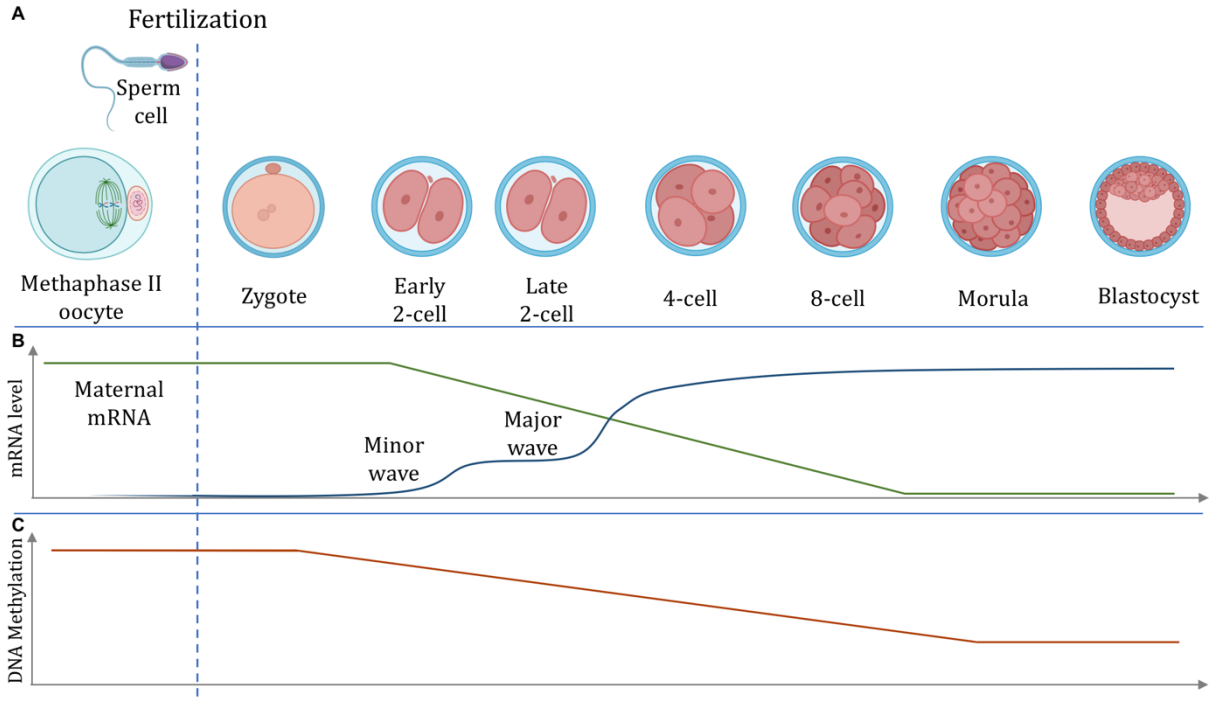


Figure-1.9: molecular events characterising the ZGA (timings referred to mouse embryo).
(**A**) Embryo cell divisions following fertilisation. Once the zygote is formed it undergoes quick cell divisions giving rise to 2-, 4-, 8- and subsequent cell stages. In mouse, an early 2-cell stage is distinguishable from a late 2-cell stage as characterised by the first transcriptional activation. (**B**) mRNA level within the embryo cytoplasm. Prior to ZGA only maternal mRNAs are present in the embryo cytoplasm. At ZGA onset zygotically transcribed mRNAs appear in the cytoplasm whereas maternal ones start to decrease. The ZGA occurs through two consecutive waves, a minor and a major one. (**C**) DNA methylation characterising the early embryo (exemplification not taking into account the different dynamics characterising paternal and maternal genome de-methylations). DNA methylation gradually decreases during the early embryo development. This allows the transcription of genomic loci whose expression is normally inhibited by such modifications as transposable elements.

1.2.1.2.2 Timings characterising ZGA

Although the timing by which the aforementioned factors activate the zygotic genome in the Metazoan species appears to be species specific, two main classifications can be made. In rapidly developing species as *C. elegans*, *Xenopus*, *Drosophila* and zebrafish the ZGA minor wave onset occurs few hours post fertilisation (hpf) (**Figure-1.10**). For instance, *Drosophila* and zebrafish embryos show zygotic transcription as soon as 1 and 2 hpf, respectively (**Figure-1.10**). On the contrary, in slowly developing animals such as mammals, the ZGA minor wave onset occurs later with mouse and human ZGA minor wave occurring 10 and 48 hpf (**Figure-1.10**) (Schulz & Harrison, 2019). Intriguingly, this difference between rapidly and slowly developing species appears to be related to the egg development strategy each group has evolved. Indeed, rapidly developing species are characterised by an external egg development strategy whereas the slowly developing species by an internal one. This observation could be biologically meaningful in the context of the more challenging environment the externally developing embryos should face compared to the internally ones as not being protected by the uterus. Thus, species characterised by an external egg development may have evolved a faster development to increase their survival likelihood. The quicker the embryonic development occurs, the sooner the embryo could deal with environmental challenges (Schulz & Harrison, 2019).

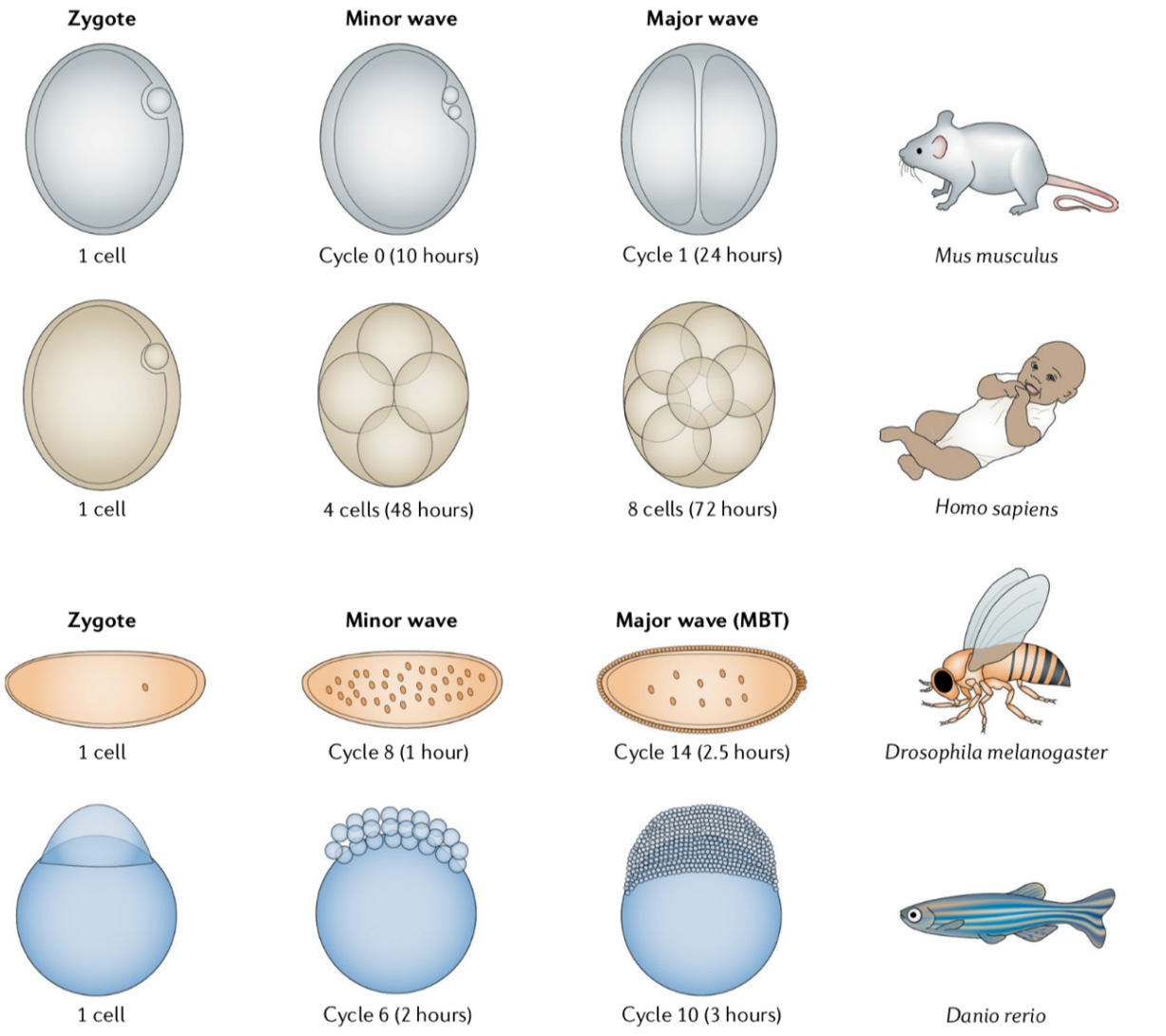


Figure-1.10: zygotic genome activation (ZGA) in different Metazoan model organisms.
ZGA occurs through a minor and a major way. Transcriptional waves dynamics are remarkably conserved across Metazoan while timings depend on the species. In species adopting internal egg development, such as mammals, ZGA minor wave occurs later than in species that have evolved an external egg development such as *Drosophila* and zebrafish (adapted from Schulz & Harrison, 2019).

1.3 Transposable element transcription during the initial phases of the embryo development

During the early stages of the embryonic development, the zygotic genome undergoes intense and dynamic epigenetic reorganisation. Among these events, global DNA demethylation and overall loss of heterochromatic regions are two of the most representative modifications occurring at these stages (Eckersley-Maslin et al., 2018). Thus, several genomic loci, whose transcription is normally silenced by such mechanisms in adult cells, result de-repressed and transcribed in the early embryo. Transposable element transcription, in particular, results remarkably enhanced during this specific time window. Importantly, TE expression during the earliest developmental phases appear to be evolutionary conserved among different species as *C. elegans*, *Drosophila* and mammals (Ansaloni et al., 2019; Eckersley-Maslin et al., 2018; Garcia-Perez et al., 2016; Hendrickson et al., 2017; Macfarlan et al., 2012; Parkhurst & Corces, 1987; Pontis et al., 2019). The reason why such a phenomenon has evolved in this specific context is likely related to the observation that a *de novo* TE insertion, in order to be transmitted to the progeny, has to occur before the germline cell development. Therefore, the earlier during the embryogenesis a *de novo* TE insertion occurs, the likelier it is for the insertion to be part of the germline. Thus, it is likely that TEs have evolved sequences and mechanisms to be transcribed during the early embryonic stages. TE transcription increases their likelihood to undergo transposition and to insert a new copy of the element in embryonic cells that will give rise to the germline being thus vertically transmitted (Chuong et al., 2017). From the host perspective, the TE activation during such a fragile stage may have severe detrimental effects impairing the proper embryonic development. The host has consequently evolved mechanisms to positively exploit this remarkable TE activation. The result of this evo-devo arm race is, at least in mammals, the adaptation of TEs as functional elements having crucial roles during mammalian zygotic genome activation and maternal to zygotic transition (Bourque et al., 2018; Chuong et al., 2017; Eckersley-Maslin et al., 2018; Pontis et al., 2019).

1.3.1 TEs regulating mammalian embryogenesis

During mouse and human zygotic genome activation, several hundred genes are transcriptionally activated. However, such transcriptional burst does not involve only protein-coding genes as it induces the transcription of a smaller, but still significant, number of TEs (Göke et al., 2015; Grow et al., 2015; Hendrickson et al., 2017; Jachowicz et al., 2017; Pontis et al., 2019; Torres-Padilla, 2020). In particular, *Dux* and *DUX4*, the mouse and human transcription factors primarily involved with the activation of the zygotic genome, have been shown to promote the transcription of specific ERV subfamilies: the murine MERVL and its human counterpart HERVL (Hendrickson et al., 2017). Despite the apparent coevolution in mouse and human of the *Dux*/*DUX4*-MERVL/HERVL pathway, the role played by such TE subfamilies within the early embryo context differs among the two species. In mouse, MERVL elements provide promoter sequences inducing the transcription of hundreds early expressed genes (Macfarlan et al., 2012; Peaston et al., 2004). On the contrary, in humans, HERVL, as wells as other HERV subfamilies like HERVK and HERVH, rarely function as promoters with HERV and HERK acting instead as long-distance enhancers (Göke et al., 2015; Grow et al., 2015; Pontis et al., 2019). Moreover, although its transcriptional activation does not appear to be regulated by *Dux*, LINE L1 element has been shown to play key roles during murine embryogenesis (Jachowicz et al., 2017; Percharde et al., 2018; Y. Wu et al., 2019). On the contrary, the functions LINE L1 plays in human embryos are still unclear.

1.3.1.1 MERVL promoting the transcription of early expressed genes in the mouse early embryo

In the mouse early embryo, at the onset of zygotic genome activation, TE transcription is intensely activated (Hendrickson et al., 2017). In particular, transcription of ERVs belonging to the MERVL subfamily is selectively promoted by *Dux*, the transcription factor primarily driving the murine ZGA (Hendrickson et al., 2017; Torres-Padilla, 2020). Importantly, MERVL elements regulate a network of early expressed genes, thus their activation plays a crucial role in promoting the establishment of the murine ZGA (Eckersley-Maslin et al., 2018; Hendrickson et al., 2017; Macfarlan et al., 2012; Rodriguez-Terrones & Torres-Padilla, 2018; Torres-Padilla, 2020). The first evidence of this observation was reported by Peaston and colleagues (Peaston et al., 2004) that in 2004

described how MERVL-derived sequences provide an alternative 5' exon to many genes expressed at the ZGA onset (2-cell stage) (**Figure-1.11**). As consequence of this phenomenon, MERVL-genes chimeric transcripts are formed with MERVL promoter regulating the transcription of the chimera. Importantly, MERVL transcription within the murine embryo is highly stage-specific and occurs exclusively at the 2-cell stage (ZGA onset). Consequently, such chimeric transcripts are found in the embryo cytoplasm only in a specifically restricted time window (Peaston et al., 2004). Peaston observations were furtherly confirmed on a larger scale by Macfarlan and collaborators (Macfarlan et al., 2012) that, in 2012, identified more than a hundred 2-cell specific genes that have co-opted regulatory sequences from MERVL elements to initiate their transcription (Macfarlan et al., 2012). Additionally, recent studies characterising the chromatin landscape of the murine early embryo, have defined how, at ZGA onset, MERVL sequences are characterised by broad ATAC peaks and marked by the histone modification H3K4me3, a histone hallmark for transcription initiation (**Figure-1.11**) (J. Wu et al., 2016; B. Zhang et al., 2016).

All together, these observations support the evidence that, only during a specific time window of the murine embryogenesis, i) the chromatin nearby MERVL element is in an open conformation, ii) MERVL loci are marked by transcription initiation modifications and iii) MERVL loci are selectively transcribed regulating the transcription of approximately one hundred early expressed genes (**Figure-1.11**).

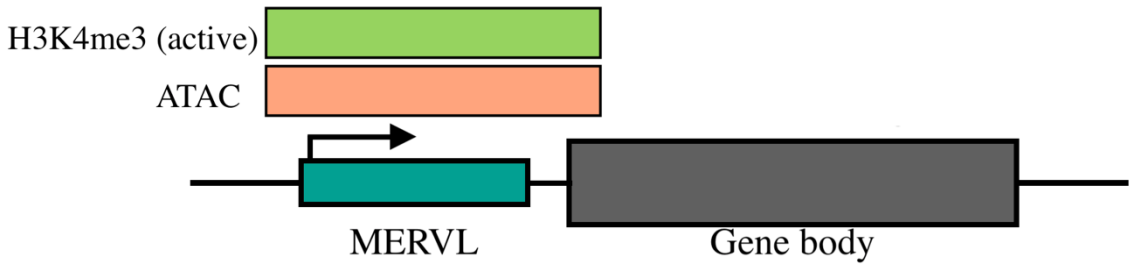


Figure-1.11: MERVL elements promoting expression of nearby genes at ZGA onset.
During mouse embryogenesis, at ZGA onset (2-cell stage), MERVL ERV elements are specifically transcribed by *Dux*. MERVL consequently promote the expression of nearby genes. Additionally, broad ATAC peaks (pink) and H3K4me3 histone modification (green) (hallmark of transcription initiation) mark MERVL loci at ZGA onset.

1.3.1.2 LINE L1 acting as chromatin remodeller in the mouse early embryo

At the onset of the murine ZGA, occurring at the 2-cell stage, TEs undergo a remarkable transcriptional activation (Hendrickson et al., 2017). As previously reported, ERVs are the main TEs being actively transcribed at these stages. However, additional TE subfamilies undergo transcriptional activation upon ZGA. Specific subfamilies of LINE L1 elements are indeed transcribed in the mouse early embryo and, importantly, their transcription is highly specific being confined at the 2-cell stage (Ancelin et al., 2016; Fadloun et al., 2013; Jachowicz et al., 2017; Peaston et al., 2004). Jachowicz and colleagues have indeed described that either the elongation of LINE L1 transcription beyond the 2-cell stage or its transcriptional repression immediately after fertilisation leads to severe phenotypes as the embryonic developmental arrest. In particular, prolonged activation of LINE L1 beyond the 2-cell stage leads to an increased DNaseI sensitivity, whereas premature silencing reduces it. These results indicate that LINE L1 expression modulates chromatin accessibility in the mouse 2-cell stage embryos (**Figure-1.12**). Interestingly, injection of the LINE L1 transcript within the cytoplasm of the 2-cell embryos has no effect on the embryonic development suggesting that LINE L1 transcription plays a role itself at the nuclear level independently from the coding nature of its transcripts (Eckersley-Maslin et al., 2018; Jachowicz et al., 2017). Additionally, LINE L1 mRNA has also been proposed to be required for the repression of the 2-cell-specific program activated by *Dux*. In this scenario, LINE L1 mRNA recruits the Nucleolin and Trim28 proteins repressing, through an unknown mechanism, the *Dux*/2-cell program and thus promoting the progression of the embryonic development (Percharde et al., 2018).

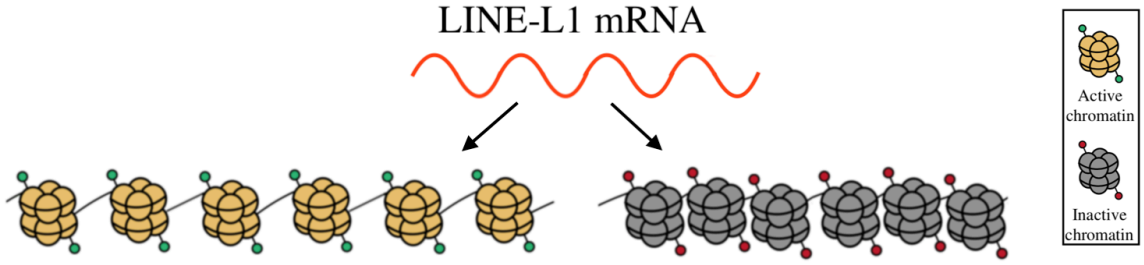


Figure-1.12: LINE L1 mRNAs acting as chromatin remodeller.
LINE L1 transcription has been shown to be confined at the 2-cell stage. Prolonged activation, as well as premature silencing, alters the chromatin structure of the embryonic genome thus suggesting that LINE L1 transcription plays a role in remodelling the embryo chromatin.

1.3.1.3 ERV elements having heterogeneous functions during the human embryogenesis

In human embryos, the transcription of several TE families is promoted upon zygotic genome activation (Göke et al., 2015; Hendrickson et al., 2017; Pontis et al., 2019). In this biological context, the transcription of ERV subfamilies HERVL, HERVK and HERVH, and their respective solitary LTR portions, appears finely regulated (Göke et al., 2015; Grow et al., 2015; Hendrickson et al., 2017; Pontis et al., 2019). Although parallelisms exist between human HERVL and its murine counterpart MERVL, as they are both transcriptionally activated by the ortholog factors *DUX4*/*Dux*, the two elements do not display the same function. While MERVL promotes the expression of approximately one hundred early expressed genes, HERVL appears not to play a similar role and its function remains unclear. On the contrary, the HERV subfamilies HERVK and HERVH appear to be functionally connected to the early human embryonic development (Gerdes et al., 2016; Göke et al., 2015; Grow et al., 2015). HERVK is the most recently evolved HERV subfamily (Gerdes et al., 2016; Subramanian et al., 2011). Importantly, it retains its protein-coding potential and, additionally, its LTR portion contains a binding domain for *OCT4*, transcription factor needed for the pluripotency maintenance (Grow et al., 2015). Consequently, when *OCT4* is expressed, like in the human early embryo, HERVK transcription is subsequently activated leading to the synthesis of viral-like proteins. Intriguingly, it has been hypothesized that such proteins may activate the antiviral response, protecting the human embryo against exogenous viral infections and associating to HERVK an immune-protective role (Grow et al., 2015). Moreover, HERVK, as well as HERVH, have been shown to act as long-distance enhancers controlling the transcription of coding and non-coding genes (Gerdes et al., 2016; Grow et al., 2015; Pontis et al., 2019). Additionally, especially in human embryonic stem cells (hESC), a model for early embryo development, HERVK and HERVH have been described as marker of pluripotency being involved in the pluripotency maintenance gene regulatory networks (Fort et al., 2014; Gerdes et al., 2016; Grow et al., 2015; Santoni et al., 2012).
Together, these observations suggest that HERVs have heterogenous functions within the human early embryo playing immune-protective roles, acting as long-distance enhancers and also participating in gene regulatory networks for the maintenance of pluripotency.

1.4 Research aims

During the initial embryonic developmental phases, the zygotic genome is transcriptionally quiescent. Then, it undergoes a remarkable transcriptional activation that involves, together with protein-coding gens, TEs. Intriguingly, this transcriptional burst is an extremely conserved feature among Metazoans. Inspired by the unique transcriptional dynamics characterising the zygotic genome activation (ZGA), by the extreme conservation of this process and by the contribution given by TE transcription to this context, my PhD project aims at the investigation of the TE transcriptional dynamics characterising the Metazoan early embryo development.

Toward this end, the implementation of a bioinformatics pipeline measuring the TE expression from RNA-seq data represents the first goal of my PhD project. Next, the second PhD project aim is represented by investigating the TE transcriptional landscape characterising the *C. elegans* early embryo. To extend the same biological question to other Metazoans, the definition of the TE transcriptional landscape in zebrafish and mouse early embryos represent two additional goals of my PhD project. In this scenario, the research questions are additionally aimed at the understanding of the possible relationships linking TE and gene transcriptional activations during ZGA. In zebrafish this has been inspired by the repetitive and non-coding nature of the micro-RNAs driving the zebrafish ZGA and by the lack of knowledge about the TE contribution to this event. In mouse, instead, the key role TE plays in the activation of the murine zygotic genome is largely known. However, it is still unclear whether such elements, and in particular the MERVL, actively contribute to the activation of the zygotic genome or whether their sequences have been passively co-opted by the zygotic genome since the role LINE L1 elements play in this context remains unclear.

In the following chapters, each of the aforementioned aims is extensively developed, described and discussed.

Chapter 2
TEspeX: a bioinformatics tool to quantify transposable element expression

2.1 Introduction

TEs are mobile and repetitive genomic sequences accounting for almost half of the murine and human genomes (Lander et al., 2001; Mouse Genome Sequencing Consortium et al., 2002). Despite their abundancy, the large majority of TEs have lost the ability to generate new copies and are represented as transpositionally inactive fragments (Chuong et al., 2017; Lander et al., 2001; Richardson et al., 2014). Additionally, TE activity is also limited by different repressing pathways the host genomes have evolved (Barau et al., 2016; Imbeault et al., 2017; Ozata et al., 2019; Walsh et al., 1998). Nevertheless, in specific biological contexts, mainly characterised by a low activity of such repressive pathways, TEs have maintained the ability to self-promote their own transcription, even when fragmented. TEs, indeed, contain internal promoter regions located in the 5' end of the element. When the promoter sequence is intact, it may drive the transcription of the TE itself regardless of the full-length or fragmented structure of the element. Importantly, TE transcriptional activation has been observed to play crucial roles in diverse physiological contexts as early embryogenesis (Ansaloni et al., 2019; Fadloun et al., 2013; Hendrickson et al., 2017; Rodriguez-Terrones & Torres-Padilla, 2018) and in neurons/neuron precursors (Erwin et al., 2014; Muotri et al., 2005; Perrat et al., 2013) as well as in pathological conditions like cancer (reviewed in Burns, 2017) and neurodegenerative diseases (Dembny et al., 2020; Guo et al., 2018; Krug et al., 2017; W. Li et al., 2012; Prudencio et al., 2017; Sun et al., 2018; Y.-J. Zhang et al., 2019).

Given both the recent technology advancements that have facilitated the generation of RNA-seq datasets and the involvement of TE transcription in such crucial physiological and pathological contexts, the development of bioinformatics tools enabling the large-scale study of TE expression is a current need of the biological and biomedical

communities. Toward this end, several bioinformatics pipelines quantifying the TE expression from sequencing data have been recently developed (Bendall et al., 2019; Jeong et al., 2018; Jin et al., 2015; Tokuyama et al., 2018; Yang et al., 2019). All the bioinformatics tools developed to quantify TE expression from RNA-seq data can be subdivided in two major groups depending on whether the RNA-seq reads are aligned to the genome or to the TE subfamily consensus sequences. While the first approach leads to the estimation of the expression of each single TE locus annotated in the reference genome, the second one provides the quantification of the expression of each TE subfamily that should summarise the general expression of the single loci. The use of one or of the other approach is not mutually exclusive and mainly depends on the research question and on the biological context of the analysis.

Nevertheless, regardless of the approach used and despite the discrete number of tools recently developed, the quantification of the TE expression is still a current challenge. The main issues are consequence of the intrinsic evolutionary processes that have led to the TE expansion within the eukaryotic genomes and are mainly due to i) the repetitive nature of TEs and ii) the large fractions of TE-derived sequences embedded in coding and non-coding transcripts.

The repetitive nature of TEs impairs the proper measurement of TE expression

Over the evolution, the mobile nature of TEs have led to the spreading of highly similar, if not identical, TE copies within the host genomes. Thus, each TE subfamily is represented by hundreds of copies dispersed in the genome that, especially for the evolutionary younger TE subfamilies, share a high degree of sequence conservation mostly lacking unique sequences. Due to the lack of unique sequences, a portion of the RNA-seq short reads transcribed from such highly similar genomic fragments can ambiguously align to many copies of the same TE fragments spread in different genomic locations (multimapping reads). As a consequence, the unequivocal assignment of these reads to unique genomic regions is often impossible. Nevertheless, this issue does not impact the quantification of the TE subfamily expression at the consensus level. Indeed, the reads deriving from identical genomic loci of the same TE subfamily are all assigned to the subfamily consensus sequence itself. On the contrary, the handling of the multimapping reads is a serious issue for the bioinformatics tools estimating the

expression of each single TE fragment annotated in the reference genome. As previously discussed, the lack of unique sequences among TE copies of the same subfamily makes it impossible to assign unequivocally multimapping reads to specific TE single loci. Trying to overcome this issue, single-locus TE expression quantification tools either discard multimapping reads (Deininger et al., 2017; Tokuyama et al., 2018) or assign them through the application of the expectation-maximization (EM) algorithm[1] (Bendall et al., 2019; Jin et al., 2015; Yang et al., 2019). Nevertheless, both approaches present limitations. The former does not provide a global picture of the TE expressional landscape as the majority of TE-derived reads are expected to result as multimapping reads. Thus, discarding them may lead to a remarkable underestimation of the TE expression (Lanciano & Cristofari, 2020). The latter, instead, provides a statistical solution not necessarily reflecting the real biological condition and, additionally, it may overestimate the expression of highly expressed TEs while underestimating the faintly expressed ones.

TE-derived sequences embedded in coding and non-coding transcripts may impair the proper measurement of TE expression

An additional challenge in the quantification of the TE expression is given by the large fractions of TE-derived sequences embedded in coding and non-coding transcripts (Lanciano & Cristofari, 2020). Approximately 9%, 10% and 16% of the zebrafish, murine and human protein-coding genes derive from TE sequences whereas the ratio increases to 19%, 25% and 35%, respectively, when considering the long non-coding RNAs (lncRNAs) (Kapusta et al., 2013). Consequently, upon transcription of canonical coding and non-coding genes the embedded TE-derived sequences are passively co-transcribed (Lanciano & Cristofari, 2020) and therefore a portion of the RNA-seq reads deriving from the expression of these genes result to align to TE consensus. As a consequence, an apparent change of TE expression may simply reflect the variation in the expression of transcripts containing TE-derived sequences (Lanciano & Cristofari, 2020). Quantifying the RNA-seq reads deriving from TE-derived sequences embedded in coding/non-coding transcripts can thus lead to an overestimation of the TE expression. Importantly, this issue affects both the tools quantifying the expression of TE single locus as well as the

[1] The expectation-maximization algorithm re-assigns the multimapping reads reiteratively. In each step the read counts of both unique-mapping and multimapping reads is used to reassign multimapping reads in the following step, until convergence is achieved.

ones measuring the expression of TE consensus sequences. To overcome such an issue, tools like ERVmap (Tokuyama et al., 2018), L1EM (McKerrow & Fenyö, 2019) and TeXP (Navarro et al., 2019), have developed specific computational approaches. Nevertheless, these tools appear to quantify only human specific TE subfamilies like ERVs or LINE L1s thus resulting not broadly usable.

Here, trying to overcome all the aforementioned issues, I present TEspeX a bioinformatics pipeline that discards the reads mapping on any annotated coding/non-coding transcript to limit the quantification of the RNA-seq reads deriving from embedded TE fragments. TEspeX is developed to measure the expression of any type of TEs, regardless of the TE classification and species of origin.

2.2 Methods and pipeline implementation

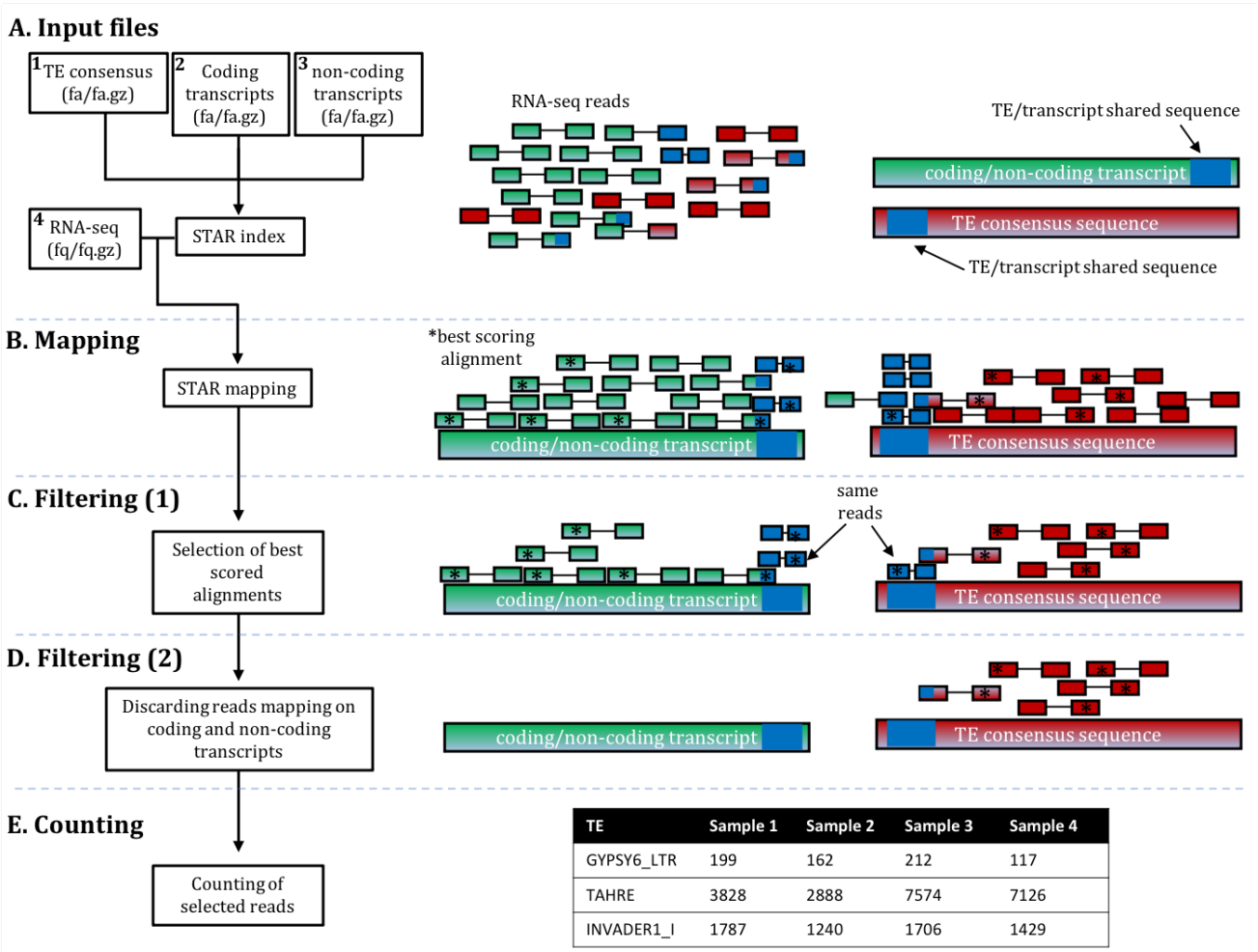
Pipeline implementation

TEspeX is developed in Python3 and takes advantage of STAR (Dobin et al., 2013), samtools (H. Li et al., 2009) and Picard (http://Broadinstitute.Github.Io/Picard) in order to map, filter and extract the final set of reads. The required input files consist in three FASTA files containing the sequences of the i) coding transcripts, ii) non-coding transcripts and iii) TE consensus sequences and in a plain text file reporting the full paths of the input FASTQ files. The input FASTA files can be in either compressed or un-compressed format as well as the FASTQ files. Additionally, TEspeX requires the user to specify i) whether the RNA-seq reads are paired or single-end, ii) the mean read length, iii) the strandness of the RNA-seq library and iv) the directory where to write the output files. Moreover, the number of threads, whether to remove or not the alignment BAM files and whether TEspeX has to build the genome index by its own or it is provided by the user can be optionally set. In case these optional parameters are not set by the user, 2 threads are used, the output BAM files are deleted and the genome index is automatically generated. Once the TEspeX run is successfully completed, the expression values of each TE consensus sequence are reported as read counts in the main output file. A file containing mapping statistics is also provided in output. Additionally, if required by the user, TEspeX can also output the BAM files containing the selected read alignments used

for the quantification of each TE consensus sequences as well as log files containing the list of executed operations on each analysed FASTQ input file. TEspeX can be run in either default mode (working on one sample at a time) or in 'wrapper' mode. When run in wrapper mode TEspeX splits the analysis in a number of parallel jobs defined by the user. The FASTQ input files are subdivided as homogenously as possible among the different jobs and each job is then run, in parallel, on a different computational node. When all the parallel jobs are successfully completed, TEspeX merges the results generated by each single job providing as output the file containing, for all the sample analysed, the expression values of each TE consensus sequence as well as a file containing the mapping statistics, as previously described. Of course, the 'wrapper' mode is available exclusively for the users having access to a multi-node HPC system. To date only the SLURM queue management system is supported by TEspeX.

Pipeline workflow

As soon as all the required input files are provided by the user, a **reference transcriptome** is built merging the input FASTA files containing the coding transcripts, the non-coding transcripts and the TE consensus sequences (**Figure-2.1A**). This corresponds to the whole transcriptome, *i.e.* the collection of all the transcripts the species under investigation can produce. Next, the **reference transcriptome** index is generated by STAR (v2.6.0c) (Dobin et al., 2013). RNA-seq reads are then mapped on the generated **reference transcriptome** using STAR. During the RNA-seq read mapping process i) the primary alignment flag is assigned to all the alignments showing the best alignment score (STAR parameter: --outSAMprimaryFlag AllBestScore), ii) alignments deriving from singleton reads are discarded when the reads come from a paired sequencing and iii) reads mapping in more than 10 different genomic locations are discarded as well (**Figure-2.1B**). All the alignments flagged as primary (-F 0x100) are then selected using samtools (v1.3.1) (H. Li et al., 2009) (**Figure-2.1C**). To avoid the counting of reads aligning to TE sequences embedded in coding and/or long non-coding transcripts, reads mapping with best-scoring alignments on any annotated coding/non-coding transcript are discarded using custom Python scripts and Picard FilterSamReads (v2.18.4) (http://broadinstitute.github.io/picard) (**Figure-2.1C**). Selected reads, mapping exclusively on TEs and in the proper orientation (when the RNA-seq library is stranded), are finally counted (**Figure-2.1D**).



TE	Sample 1	Sample 2	Sample 3	Sample 4
GYPSY6_LTR	199	162	212	117
TAHRE	3828	2888	7574	7126
INVADER1_I	1787	1240	1706	1429

Figure-2.1: TEspeX pipeline workflow.
(**A**) Input files of the pipeline: RNA-seq reads, TE consensus sequences, coding and non-coding transcripts. The TE consensus and the coding/non-coding transcripts are merged in a unique reference file. Next, STAR index is generated. (**B**) RNA-seq reads are mapped by STAR on the reference transcriptome. Primary alignment flag is assigned to best scoring alignments. (**C**) Mapped reads are filtered, selecting reads deriving from best scoring alignments. (**D**) Reads deriving from alignments displaying best alignment score on both coding/non-coding transcripts are discarded. (**E**) Selected reads, mapping exclusively on TEs and in the proper orientation are finally counted. In the sketch on the right of the figure, coding/non-coding transcripts are depicted in green whereas TE consensus sequence in red. Both features share a common sequence depicted in blue. Reads are mapped on the TE consensus sequences and coding/non-coding transcripts and reads deriving from best scoring alignments are selected (indicated by an asterisk). Reads deriving from the transcription of the shared sequence between TE consensus sequence and coding/non-coding transcripts map on both (blue reads). TEspeX filters all the reads mapping with best alignment score on coding/non-coding transcripts. Reads aligning on the TE/transcript shared sequence (blue) are discarded as well as they are not unequivocally assignable to either TE consensus sequences or coding/non-coding transcripts.

RNA-seq reads simulation

To test the efficacy of TEspeX in discarding reads transcribed as part of coding/non-coding transcripts, four RNA-seq samples have been generated *in silico*. Briefly, coding and non-coding transcripts datasets have been downloaded from Ensembl (Zerbino et al., 2018) for *Drosophila melanogaster* (dm6), *Mus musculus* (mm10) and *Homo sapiens* (hg19). From such FASTA files 125 nt-long paired-ended RNA-seq reads have been simulated by using polyester (Frazee et al., 2015) and converted in FASTQ format using the reformat tool of the BBmap package (Bushnell, 2014). TE expression has been then quantified at the consensus sequence level using TEspeX and three of the most accurate and recent tools developed for the TE expression quantification, SalmonTE (Jeong et al., 2018), SQuIRE (Yang et al., 2019) and TEtranscripts (Jin et al., 2015). TE expression quantification has been measured following the best practice suggested by the manuals of each tool. Only TEs belonging to DNA transposons (also comprising rolling-circle transposons), LTR, LINE and SINE families have been considered in the calculations.

RNA-seq dataset

To test the efficiency of TEspeX in measuring the TE expression in non-simulated RNA-seq dataset, raw reads generated from a *Drosophila* ALS genetic model by Krug and colleagues (Krug et al., 2017) have been retrieved and analysed with TEspeX.

In the Krug study, wild-type human TDP-43 (hTDP-43) has been transgenically overexpressed in the fly brain under the control of either a pan-glial (*Repo*) or pan-neuronal (*ELAV*) promotor. The functional abnormality of TDP-43 protein characterizes human ALS and FTD pathologies. The transgenic overexpression of hTDP-43 induces the increase of the protein concentration above the endogenous levels, reproducing many of the signatures of the ALS disease in humans. Krug and colleagues highlighted how, in both neuronal and glial cells, the overexpression of hTDP-43 induces a remarkable transcriptional activation of TEs, and especially of retrotransposons. This is particularly remarkable in the glial cells where, out of 29 differentially expressed TEs, 23 resulted up rather than downregulated upon hTDP-43 overexpression. Importantly, the glial and neuronal cells share specific TE subfamilies that resulted upregulated in both models. On the contrary, upregulation of some specific TE subfamilies resulted private to one or the other model. Importantly, the *gypsy* LTR retrotransposon, one of the most active

Drosophila natural transposon, resulted upregulated upon the overexpression of hTDP-43 in the glial, but not neuronal, cells. Moreover, non-transcriptomic based experiments allowed Krug and colleagues to highlight how *gypsy* expression in glial cells remarkably contributes to the hTDP-43 mediated toxicity.

Here, TEspeX has been used to test its capability in replicating the detection of the TE upregulation upon the expression of hTDP-43 in both glial and neuronal cells. Krug dataset is composed by 3 biological samples, 2 replicates each, consisting in control *Drosophila* heads, *Drosophila* heads expressing human TDP-43 in either glial or neuronal cells. TEspeX has been run providing as input coding and non-coding transcripts downloaded from Ensembl (dm6 – BDGP6.28 version) and TE consensus sequences from RepBase database (v-24.07) (Bao et al., 2015). To identify TE consensus sequences resulting differentially expressed upon expression of hTDP43 in either glial or neuronal cells, edgeR (Robinson et al., 2010) has been used. The library size of each sample has been set providing the total number of reads mapped on the transcriptome (coding, non-coding and TE consensus sequences). Normalisation of raw read counts has been applied using the TMM method whereas the common, trended and tagwise dispersions have been estimated by maximizing the negative binomial likelihood (default). Next, differentially expressed TE consensus sequences in both conditions (glial expressing hTDP-43 vs controls and neuron expressing hTDP-43 vs controls) have been tested performing a quasi-likelihood F-tests (glmQLFit and glmQLTest). TE consensus sequences have been considered as differentially expressed when showing FDR < 0.05 and log2FC < -0.58 or > 0.58 (1.5-fold in linear scale).

Pipeline distribution

TEspeX is freely available at the dedicated GitHub repository: https://github.com/fansalon/TEspeX. In the repository, together with the code, the procedures to install the required software is described into details for both UNIX and Mac OS X systems. Additionally, information on how to run TEspeX from command line are provided as well as one testing dataset. Finally, slurm batch script to eventually run TEspeX on HPC systems is provided with detailed information and instructions to use it.

2.3 Results

TEspeX quantifies no TE expression when RNA-seq reads are generated from coding and non-coding transcripts

TEspeX has been implemented to discard RNA-seq reads mapping with best alignment score on any coding and non-coding transcript. To test the TEspeX efficiency in performing such a correction, the TE expression has been calculated using TEspeX as well as SalmonTE, SQuIRE and TEtranscripts in an *in silico* RNA-seq dataset generated from *Drosophila*, mouse and human annotated transcripts (see Methods). Importantly, no RNA-seq reads have been generated from any TE sequence. No TE expression is therefore expected to be measured.

The results of the TE expression quantification highlighted how TEspeX does not assign RNA-seq reads to any of the analysed TE consensus sequences (**Figure-2.2**). On the contrary, all the other tested tools report evidence of TE expression (**Figure-2.2**). This result was observed in all the three species. Given that no RNA-seq reads have been generated from TEs, the TE expression quantified by all the tested tools, except TEspeX, is a consequence of the counting of the RNA-seq reads deriving from TE sequences embedded in coding/non-coding transcripts. These reads are not generated by TE transcription itself as RNA-seq reads have been originally simulated exclusively from coding and non-coding transcripts. Nevertheless, when aligned back to the genome/transcriptome/TE subfamilies (depending on the tool), they map on the TE consensus sequences due to their sequence similarity with TEs. As a consequence of the correction applied, none of these reads is considered in the final TE expression calculation by TEspeX. On the contrary, lacking similar corrections, the other tested tools count such reads when measuring the TE expression (**Figure-2.2**).

While this result is a consequence of an *in silico* simulation, far from the real transcriptional scenario characterising the transcriptome of living beings, this data highlights how TEspeX specifically applies corrections to discard RNA-seq reads deriving from the transcription of TE fragments embedded in coding and non-coding transcripts. This correction may potentially limit the counting of RNA-seq reads possibly deriving from TE co-transcriptional events.

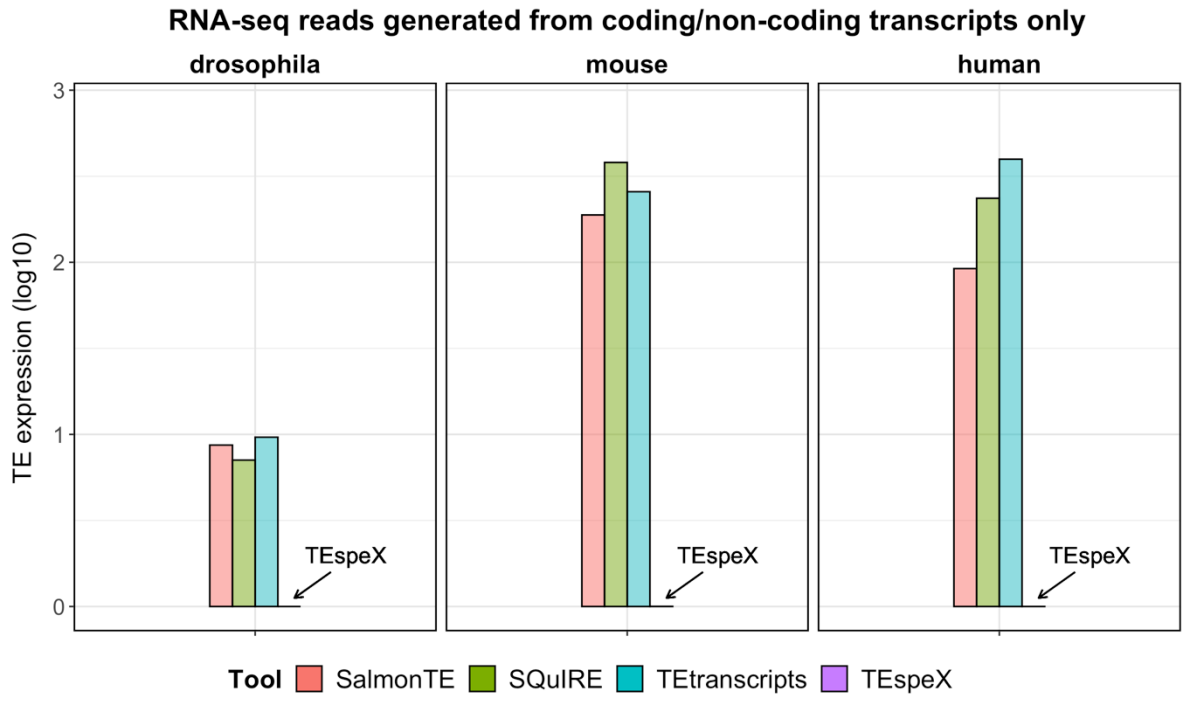


Figure-2.2: quantification of TE expression from RNA-seq reads generated from coding/non-coding transcripts only.
RNA-seq reads have been simulated *in silico* from coding and non-coding transcripts only for *Drosophila*, mouse and human. TE subfamily expression has then been calculated by SalmonTE (red), SQuIRE (green), TEtranscripts (cyan) and TEspeX (violet). Results show how TEspeX is the only tool to properly measuring no reads mapping on TE consensus sequences. TE expression is reported on y-axis as the mean (n=4) of raw number of reads mapped on the TE consensus sequences divided by the number of consensus sequences analysed by each tool. DNA, RC, LTR, LINE and SINE transposons have been considered in the calculation.

TEspeX recapitulates the TE upregulation upon expression of hTDP-43 in *Drosophila* glial and neuronal cells

In recent years, several studies have described how the expression of retrotransposons results remarkably increased in different neurodegenerative diseases such as ALS and Alzheimer's disease in diverse model organisms as well as human patients (Dembny et al., 2020; Guo et al., 2018; Krug et al., 2017; W. Li et al., 2012; Prudencio et al., 2017; Sun et al., 2018; Y.-J. Zhang et al., 2019). In particular, Krug and colleagues reported how the transgenic overexpression of the human TDP-43 in both glial and neuronal cells leads to the TE, and especially retrotransposon, upregulation in both cell types. In particular, according to Krug results, the upregulation of the *gypsy* element in glial cells remarkably contributes to the hTDP-43 mediated toxicity.

To test the ability of TEspeX in reproducing such results, RNA-seq reads from Krug and colleagues have been retrieved (Krug et al., 2017). The dataset consists in 3 biological conditions comprising control *Drosophila* heads as well as *Drosophila* heads expressing hTDP-43 in either glial or neuronal cells. First, TE expression has been quantified by using TEspeX. Then, hierarchical clustering and principal component analysis (PCA) have been performed to explore the similarity of the analysed samples based on the TEspeX-calculated TE expression values. The hierarchical clustering highlighted how, control and hTDP-43 expressing samples clustered in two different branches of the dendrogram (**Figure-2.3A**). Moreover, within the hTDP-43 expressing sample branch, the samples expressing hTDP-43 in either glial or neuronal cells cluster separately according to their cell type. (**Figure-2.3A**). These results were confirmed by the PCA showing how, according to the values of the first two principal components, the samples were differently grouped on the basis of their original biological condition (**Figure-2.3B**). Together, these results suggested that TEspeX is able to correctly identify the three different TE expression profiles of the three different biological conditions analysed.

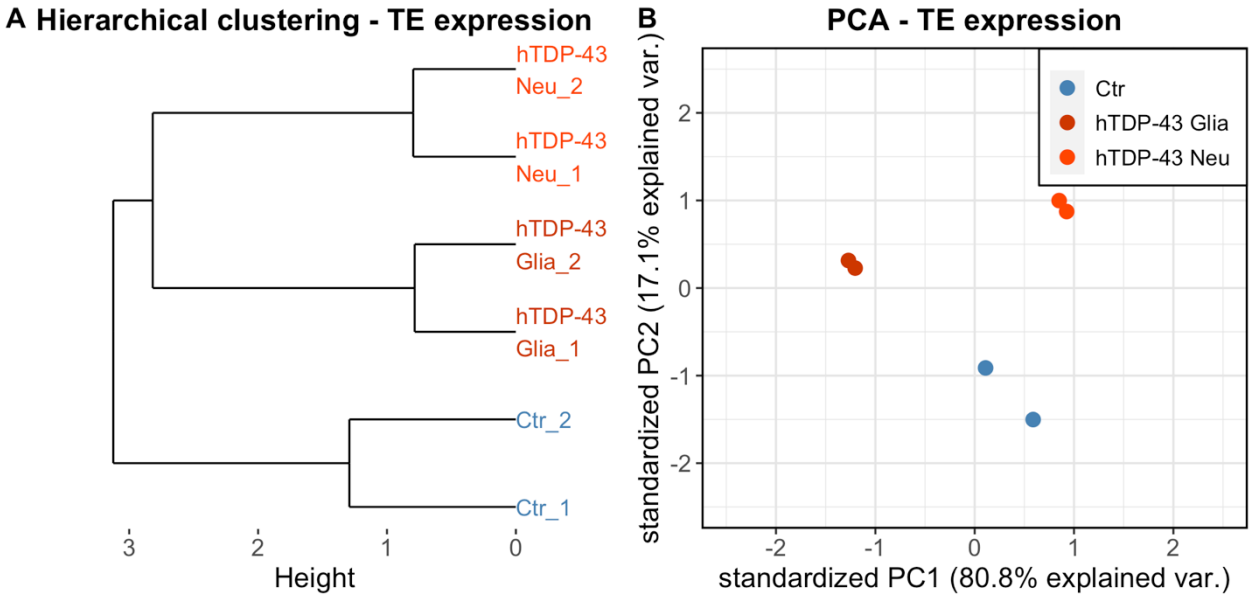


Figure-2.3: hierarchical clustering and PCA performed on TE expression values calculated by TEspeX.
(**A**) Hierarchical clustering showing the separation between Ctr and hTDP-43 expressing samples based on TEspeX-calculated TE expression. Ctr (blue) and hTDP-43 expressing samples (red and orange) are clustered in the two main dendrogram branches. Additionally, samples expressing hTDP-43 in glial (red) and neuronal cells (orange) cluster in two different sub-branches. (**B**) PCA showing the separation of Ctr, hTDP-43 glia and hTDP-43 neuron in the two-dimensional space defined by the first two principal components. The separation of the Drosophila samples expressing hTDP-43 in neuron is mainly driven by the expression of COPI_DM_I and Gypsy1-I_DM retrotransposons.

Next, to identify the TE subfamilies whose expression is differentially regulated upon the overexpression of hTDP-43, the differentially expressed (DE) TEs between control samples and hTDP-43 expressing samples in glial or neuronal cells have been identified. Globally, 35 and 37 TE subfamilies resulted differentially expressed in hTDP-43 glial and neuronal cells, respectively, compared to controls (FDR<0.05 and $\log_2$FC >1 or <-1) (**Figure-2.4A and B**). Consistent with the results displayed by Krug and colleagues, the TEspeX analysis highlighted how, in both glial and neuronal cells, the majority of the TEs resulting differentially expressed displayed an upregulation, rather than a downregulation, upon the expression of hTDP-43 (**Figure-2.4A and B**). This is particularly true for the glial cells where 28 out of 35 DE TEs (80%) resulted upregulated in the analysis made by TEspeX being consistent with the 23 upregulated TEs out of the 29 DE (79%) highlighted by Krug. Moreover, according to the TEspeX analysis, in both glial and neuronal cells, almost the totality of the upregulated TEs resulted annotated as retrotransposons (LINE and LTR), furtherly confirming the Krug data (**Figure-2.4C and D**).

Additionally, in their manuscript, Krug and colleagues highlighted how the expression of 5 TEs is particularly elevated in both cell types upon the hTDP-43 expression. TEspeX results highlighted how 4 out of the 5 TEs (HETA, HMSBEAGLE_I, GTWIN_I and BEL_I [also called 3S18]) resulted significantly upregulated also in the present analysis (**Figure-2.4C and D**) with the not validated one (MDG3_DM) showing a significant FDR (FDR=3E-03) but not a $\log_2$FC > 1 (0.47) and thus not belonging to the set of upregulated TEs. Furtherly confirming the Krug data, the 4 TEs identified by the authors of the paper as having high expression exclusively in neuronal cells upon hTDP-43 overexpression, resulted significantly upregulated also in the TEspeX-based analysis (TART-A, TAHRE, STALKER2_I and MDG1_I) (**Figure-2.4D**). Finally, the *gypsy* element (here called GYPSY_I) identified by Krug as the TEs that mostly contributes to the hTDP-43 mediated toxicity, resulted one of the most significantly upregulated TEs upon the hTDP-43 expression in glial cells in the TEspeX analysis (**Figure-2.4A**).

Overall, TEspeX successfully recapitulated the upregulation of retrotransposons upon the overexpression of hTDP-43 in both glial and neuronal cells observed by Krug and colleagues. Moreover, almost all the TEs highlighted by Krug and colleagues as remarkably transcribed upon hTDP-43 expression were identified also by TEspeX. Together, all the transcriptomic observations highlighted by Krug and colleagues in their manuscript were confirmed by TEspeX.

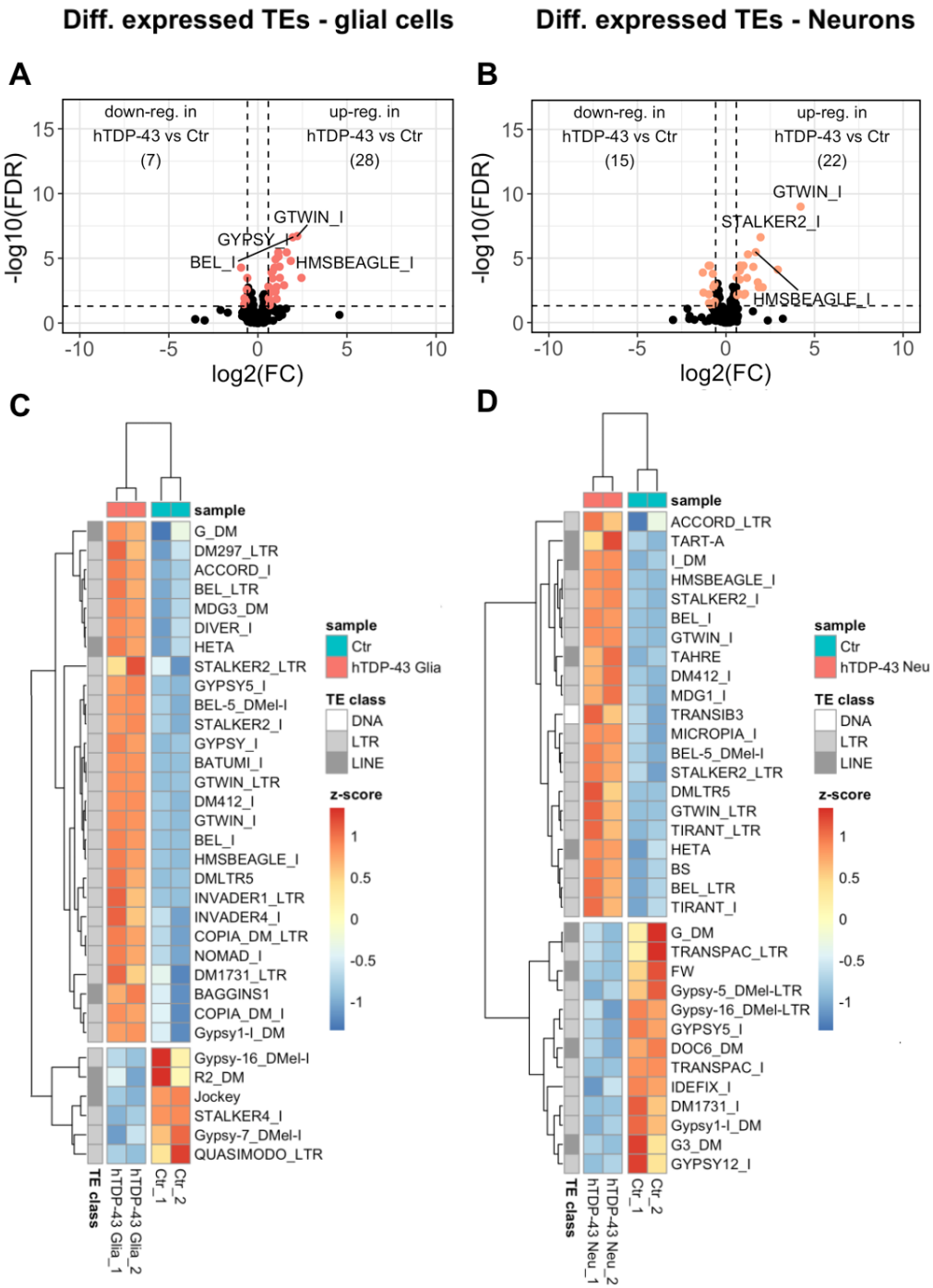


Figure-2.4: differentially expressed TEs upon expression of hTDP-43 in either glial or neuronal cells.

(**A**) Differentially expressed TEs upon expression of hTDP-43 in glial cells. Of the total 35 differentially expressed TEs, 7 resulted down and 28 resulted upregulated in hTDP-43 expressing samples compared to controls. (**B**) Differentially expressed TEs upon expression of hTDP-43 in neurons. Of the total 37 differentially expressed TEs, 15 resulted down and 22 resulted upregulated in hTDP-43 expressing samples compared to controls. (**C**) Heatmap showing the expression of the differentially expressed TEs upon expression of hTDP-43 in glial cells. (**D**) Heatmap showing the expression of the differentially expressed TEs upon expression of hTDP-43 in neuronal cells. In both (**C**) and (**D**) hierarchical clustering performed on both samples (columns) and TEs (rows) successfully groups control samples together (blue) and separately from hTDP-43 samples (red). Similarly, down and upregulated TEs are grouped in two different portions of the heatmap. Squares on the left of the heatmap indicated the TE family (DNA - white, LTR – light grey, LINE – dark grey). Expression values reported as scaled normalised TMM counts.

2.4 Conclusions

Given the repetitive nature of TEs and their impact in prompting the evolution of coding and non-coding portions of the eukaryote genomes, a high fraction of TE-derived sequences is embedded in coding and non-coding transcripts. Consequently, upon transcription of these coding and non-coding genes, the embedded TE sequences might be passively co-transcribed as part of the transcripts and counted in the TE expression quantification from RNA-seq reads (Lanciano & Cristofari, 2020). To date, only few bioinformatics tools estimating the TE expression from RNA-seq reads perform a correction on such phenomenon. Nevertheless, such few tools have been implemented to work exclusively on specific human TE subfamilies. Trying to overcome this issue, TEspeX has been developed and here described. To avoid the counting of RNA-seq reads deriving from TE sequences embedded in coding and non-coding transcripts, TEspeX discards the reads mapping with best alignment score on any annotated transcript. To test the functionality of TEspeX in applying such correction, TE expression has been measured in RNA-seq reads *in silico* generated from coding and non-coding transcripts in *Drosophila*, mouse and human. Importantly, quantification of TE expression in such RNA-seq datasets showed how TEspeX successfully quantified no TE expression suggesting its correct functionality in applying such correction. Moreover, in recent years, several studies have described how retrotransposons undergo a transcriptional activation in different neurodegenerative diseases such as ALS and Alzheimer's disease in model organisms as well as human patients (Dembny et al., 2020; Guo et al., 2018; Krug et al., 2017; W. Li et al., 2012; Prudencio et al., 2017; Sun et al., 2018; Y.-J. Zhang et al., 2019). To test the TEspeX capability in reproducing the TE transcriptional activation in this biological context, RNA-seq reads have been retrieved from a publicly available dataset comprising *Drosophila* control heads as well as heads expressing hTPD-43 in either glial or neuronal cells generated by Krug and colleagues (Krug et al., 2017). Consistent with the results highlighted by the authors of the manuscript, TEspeX successfully detected the retrotransposon upregulation upon the expression of hTDP-43 in both glial and neuronal cells. Importantly, 4 out of the 5 TEs highlighted by Krug and colleagues as particularly expressed in both cell types upon the hTDP-43 overexpression were successfully detected by TEspeX with the not validated one showing a significant FDR but not a $\log_2$FC > 1. Similarly, TEspeX succeed in detecting the significant

upregulation of 4 TEs particularly expressed in neuronal cells identified by Krug as well as the highly significant upregulation of the *gypsy* element in glial cells being, once again, consistent with the Krug and colleagues results.

Concluding, TEspeX is a new bioinformatics tools for the quantification of the TE expression in an un-biased manner towards the counting of the RNA-seq reads deriving from TE-fragments embedded in coding/non-coding transcripts. Tests performed both *in silico* and on a publicly available RNA-seq dataset confirmed the correct functioning of the tool.

Chapter 3
Exploratory analysis of transposable element expression in the *C. elegans* early embryo

3.1 Introduction

TEs are repetitive DNA sequences spread among the genomes of almost all the eukaryotes (Wicker et al., 2007). TEs can be classified in DNA transposons and retrotransposons according to their mechanism of mobilisation. DNA transposons are composed by DNA and rolling-circle (RC) elements and mobilize through a DNA intermediate, while retrotransposons are composed by Long Terminal Repeats (LTR) and non-LTR (LINE and SINE) elements that take advantage of an RNA intermediate for their mobilisation (Feschotte, 2008; Wicker et al., 2007). TEs make up a large portion of human and murine genomes (40–45%) and despite having been understudied and often considered as *junk* and selfish elements, it is currently believed that they have played and continue to play important roles in the biology and evolution of Metazoan (Chuong et al., 2017; Feschotte, 2008; Johnson & Guigó, 2014; Perrat et al., 2013; Piacentini et al., 2014). One of the first observation of the TE existence and activity in Metazoan was made in *Drosophila melanogaster* where specific outcrosses lacking the silencing of the DNA transposon P-element displayed sterility and germline abnormalities. This discovery allowed to define the molecular mechanism behind this process nowadays known as hybrid dysgenesis (Engels, 1983). More than ten years later, *Caenorhabditis elegans* (*C. elegans*) mutants, deficient for RNA interference (RNAi) pathway, were described by Mello and Fire to display an increased TE mobilization thus proposing that the RNAi system has evolved also as a defence response to protect the germline from TE activity (Tabara et al., 1999). Currently, although it might represent a driving force in genome evolution, TE activity in the gonads is widely accepted to be mostly inhibited by the PIWI/piRNA pathway (Aravin et al., 2007) with TEs resulting expressed and active during embryogenesis (reviewed in Garcia-Perez et al., 2016; Rodriguez-Terrones & Torres-Padilla, 2018) as well as in the adult central nervous system (CNS) (Baillie et al.,

2011; Coufal et al., 2009; Erwin et al., 2014, 2016; Muotri et al., 2005; Perrat et al., 2013; Richardson et al., 2014). TEs have indeed been reported to play crucial roles during embryogenesis where they modulate gene expression acting as regulatory elements, providing promoters and binding sites, regulating chromatin accessibility, and physically interacting with transcripts (Kunarso et al., 2010; Rodriguez-Terrones & Torres-Padilla, 2018). Additionally, TEs have been described to be involved in diverse biological processes during mammalian embryogenesis such as pluripotency maintenance, embryo viability and immune response priming (Garcia-Perez et al., 2016; Grow et al., 2015; Hackett et al., 2017; Percharde et al., 2018). According to these studies, the complete lack of expression as well as the uncontrolled over-expression of TEs are not compatible with the proper development of the mammalian embryos thus suggesting that, in this biological context, TEs are necessary and crucial players rather than accessory elements. Finally, TEs have also been suggested to play a dual role in the CNS of organisms such as *Drosophila*, mouse and human. On one hand, activity of retrotransposons in CNS has been linked with the generation of somatic mosaicism in human neurons (Baillie et al., 2011; Coufal et al., 2009; Erwin et al., 2014, 2016; Muotri et al., 2005; Perrat et al., 2013; Richardson et al., 2014), furtherly proposed to be correlated with the evolution of cognitive capabilities (Baillie et al., 2011; Erwin et al., 2014, 2016). On the other, alteration of their expression and activity have been associated to neurodevelopmental and neurodegenerative disorders (Guo et al., 2018; Krug et al., 2017; Sun et al., 2018; Tan et al., 2018).

C. elegans is a ~1 mm long nematode largely used as model organism. Its maintenance under laboratory conditions is simple as the transparent nematode is characterized by a short generation time (3-4 days), its food source is *Escherichia coli* and up to 1,000 worms can be cultured at the same time in a 55mm petri dish (Corsi et al., 2015). Additionally, *C. elegans* gene manipulation can be carried out in simple and very effective ways (Fire et al., 1998; Grishok & Mello, 2002). The adult is composed of about 1,000 somatic cells, 302 of which are neurons. *C. elegans* genome encodes ~20,000 protein coding and 25,000 non-coding genes with approximately 15% of genome accounting for TE sequences (Laricchia et al., 2017). Unlike in *Drosophila*, mouse and human genomes where the majority of TEs are retrotransposons, *C. elegans* genome is characterised by DNA transposons. Globally, 74% of *C. elegans* TEs are annotated as DNA transposons, 16% as

RC transposons and 10% as retrotransposons (1% SINE, 4% LINE, 5% LTR). According to literature, the Tc/Mar family (DNA TEs) is the most active, while active retrotransposition was never observed under laboratory conditions (Bessereau, 2006; Laricchia et al., 2017). *C. elegans* embryogenesis lasts for ~16 hours, and, during the early stages, five asymmetric divisions produce six founder cells: AB, MS, E, C, D, and P4. In more details a P0 zygote cell gives rise to a larger anterior cell, AB, and a smaller posterior blastomere, P1 (2-cell stage). P1 cell undergoes an asymmetric division that gives rise to EMS and P2 daughter cells, while the AB cell, through a symmetric division, gives raise to ABa and ABp (4-cell stage). Subsequent asymmetric cell divisions of EMS into MS and E, of P2 into C and P3, and symmetric divisions of ABa and ABp, which generate ABal, ABar, ABpl and ABpr, characterize the 8-cell stage. The further divisions of the 8 cells complete the generation of the founder cells whose descendants will produce specific cell types (16-cell stage) with germline cells deriving from posterior descendant cells and somatic tissues deriving from the anterior ones (Sulston et al., 1983). The *C. elegans* maternal to zygotic transition (MZT) begins in the 4-cell stage yet occurring differently from the majority of the other Metazoans as it results to be regulated post-transcriptionally and post-translationally rather than transcriptionally and with zygotic genome activation (ZGA) occurring with different timings in somatic and germline precursor cells (Robertson & Lin, 2015). For these reasons, the identification and definition of the transcriptionally dynamics characterising the *C. elegans* MZT may result hardly definable and is not object of this study.

TE expression dynamics occurring in the *C. elegans* early embryo have been investigated using the single-cell RNA sequencing (scRNA-seq) dataset generated by Tintori *et al.* in 2016 (Tintori et al., 2016). For this purpose, taking advantage of the TEspeX bioinformatics pipeline (described in Chapter 2 of this thesis), *if*, *when* and *where* each specific class of TEs is expressed during *C. elegans* development and their potential correlations with the expression of protein coding genes have been investigated.

3.2 Results and discussion

A bioinformatics pipeline to specifically measure TE expression levels

Taking advantage of the scRNA-seq dataset published by Tintori *et al* (Tintori et al., 2016), TE expression was quantified in all the sampled cells using TEspeX. The input dataset is composed of 164 samples subdivided among the 31 different cell types characterizing 5 early embryo cell stages (1-, 2-, 4-, 8- and 16-cell stages). The TEspeX pipeline is designed in order to exclude from the TE expression quantification those reads that may derive from TE fragments embedded in coding/non-coding transcripts (**Figure-3.1A and B**). TEspeX subdivides the mapped reads in *TE-specific* and *TE-non-specific* with the first group being characterised by reads confidentially transcribed by TE *per se* and the second one featured by reads likely deriving from the transcription of TE fragments embedded in coding/non-coding transcripts. On average, about 80% of reads were mapped against the whole reference transcriptome obtained from the union of coding, non-coding and TE transcripts. TE expression resulted low but detectable with a median number of *TE-specific* reads of 0.1% across all the samples. Interestingly, about 20% of the reads mapping with at least one best alignment on TEs belongs to the *TE-non-specific* read group. For these reads it is not possible to determine whether they originated from a coding/non-coding transcript or a TE and therefore, keeping them into account, might cause biased expression level calculations. TE expression levels were also quantified using SalmonTE (Jeong et al., 2018) in order to compare the TEspeX results with the ones obtained from a published and well-established tool. Globally, SalmonTE confirmed the TE expression level trends highlighted by TEspeX (**Figure-3.1C**). However, especially in the AB descendant cells of the 16-cell stage, SalmonTE indicated a generally higher TE expression levels with respect to TEspeX. To better understand the origin of such a difference, the reads defined as *TE-non-specific* by TEspeX were selected and quantified in each sample. Intriguingly, the *TE-non-specific* reads resulted more abundantly quantified in AB-descendant cells (16-cell stage), which correspond to the samples with the highest difference between SalmonTE and TEspeX (**Figure-3.1D**). These results suggest that the differences observed between the two pipelines are mainly due to the different usage of *TE-non-specific* reads with SalmonTE using, to measure TE expression levels, also reads likely deriving from TE fragments embedded in coding/non-coding transcripts. Interestingly, 16-cell stage AB cells give also rise to neurons (Altun, Z.F. &

Hall, D.H., 2011; Hobert, 2010; Tintori et al., 2016), which are known to be characterised by the expression of a high number of long non-coding RNAs (lncRNAs) which in turn are enriched for TE fragments (Chuong et al., 2017; Kapusta et al., 2013).

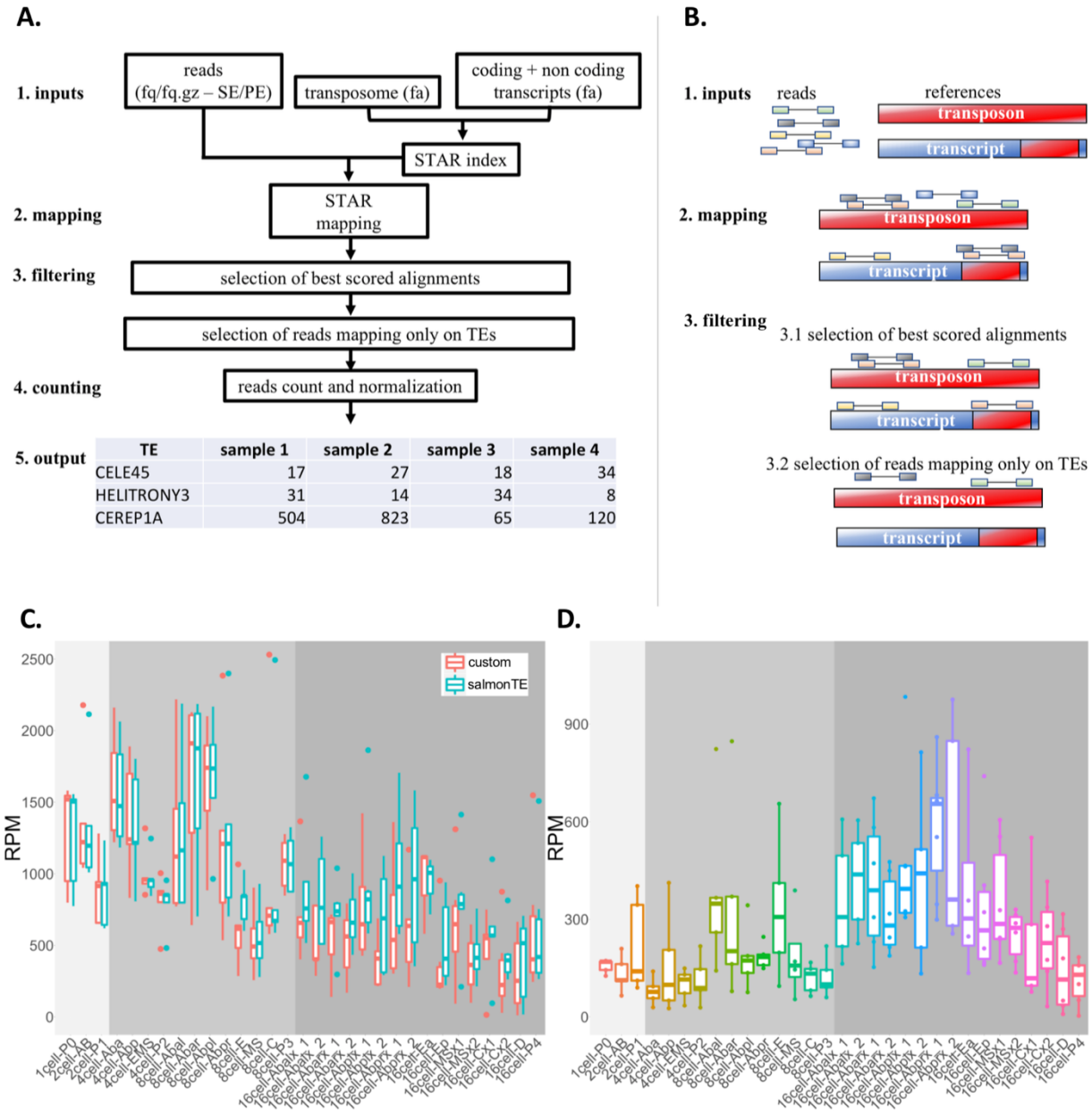


Figure-3.1: bioinformatics pipeline for the quantification of read specifically mapping on TEs.
(A) and **(B)** Workflow and schema of the TEspeX pipeline. Reads are mapped, allowing multimapping, against the reference transcriptome (composed by annotated coding and non-coding transcripts [blue] and TE consensus sequences [red]). Best scoring alignments are selected and then, to avoid selection of *TE-non-specific* reads, reads mapping with best scoring alignments both on transposome and transcriptome are discarded. **(C)** Global TE expression levels calculated for every cell type using TEspeX (custom) and SalmonTE. **(D)** Quantification of *TE-non-specific* reads used to assess whether the increased TE expression in AB descendant cells of the 16-cell stage highlighted by SalmonTE derives from *TE-non-specific* reads quantification.

TE expression changes among the stages of the *C. elegans* early embryo

Having assessed the consistency in the TEspeX TE quantification in the *C. elegans* early embryo, TE global expression profiles in each of the 31 cell types were inspected (**Figure-3.2A**). TE abundance resulted particularly high in the 1-, 2-, 4- and 8-cell stages. Intriguingly, the 1- and 2-cell stages are transcriptionally inactive (Osborne Nishimura et al., 2015) thus proposing the TE mRNAs as a component of those maternal transcripts deposited in the oocyte cytoplasm as needed by the embryo in the initial developmental stages characterised by a quiescent zygotic genome. To define whether the different cells are defined by the expression of specific TE classes a principal component analysis (PCA) was performed on the expression levels of all the *C. elegans* TEs belonging to DNA, LTR, LINE and SINE classes. From the principal component analysis it resulted that the 164 cells are subdivided in two major groups according to their TE expression with the first one composed by cells belonging to the initial developmental stages (1-, 2-, 4- and 8-cell stages), and the second one principally constituted by cells from the 16-cell stage (**Figure-3.2B**). LTR expression determines the grouping of 1-, 2-, 4-, 8-cell stages, while non-LTR retrotransposons (SINE and LINE) expression determines the separation of 16-cell stage from the other cell stages, indicating that these two groups of elements have rather opposite expression dynamics. These results support the observation that LTR and non-LTR retrotransposon expression might be differentially regulated in the *C. elegans* early embryo.

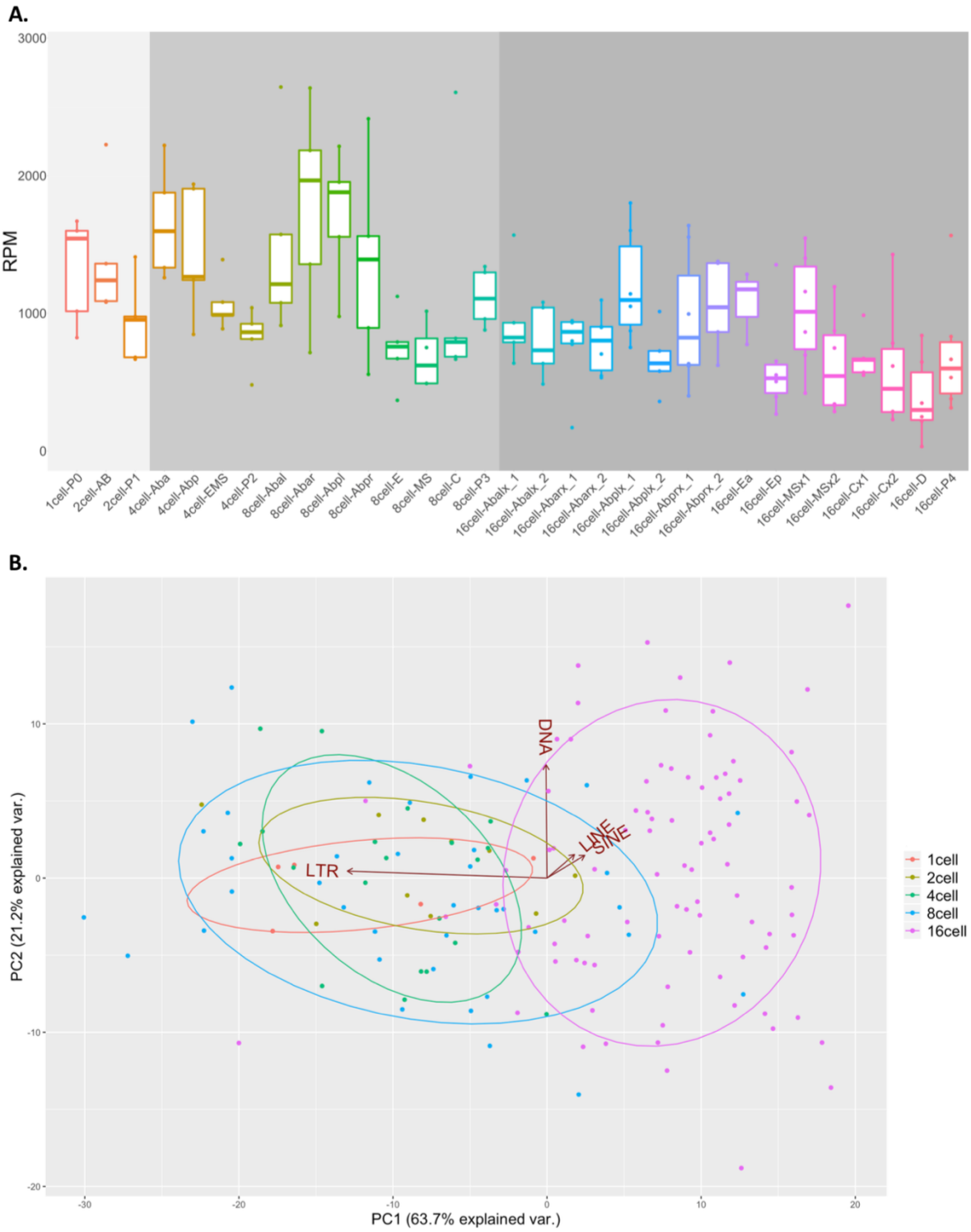


Figure-3.2: TE global expression profile among the 31 C. elegans early embryo cell types.
(A) Normalised cumulative TE expression values for all the 31 *C. elegans* early embryo cell types. Cells belonging to 1- and 2-cell stages (light grey background) are transcriptionally inactive and show, together with cells of the 4- and 8-cell stages (grey background), the highest levels of TE expression. TE expression levels decrease in the 16-cell stage (dark grey background) where embryo cell fate starts to be determined. **(B)** PCA analysis showing the distribution of all the 164 analysed cells according to their TE expression. LTR expression determines the grouping of 1-, 2-, 4-, 8-cell stages, while non-LTR retrotransposons (SINE and LINE) expression determines the separation of 16-cell stage from the other cell stages.

LTR expression is higher during stages associated to pluripotency maintenance and might activate the embryo innate immune response

To better investigate TE classes expression patterns among the different cell types and stages, TE expression levels were inspected according to the four main transposon classes (LTR, LINE, SINE and DNA transposons).

LTR retrotransposons showed the highest expression levels in the *C. elegans* embryo when compared to the other TE classes being, overall, highly abundant especially in the initial stages of *C. elegans* embryo development (1-, 2-, 4- and 8-cell stages) (**Figure-3.3A**). In particular, LTR are highly expressed in the zygote (1-cell P0) and in almost all the AB cells of the 2-, 4- and 8-cell stages. Intriguingly, LTR expression decreases strongly in the 16-cell stage. In particular, CER1 and LTRCER1 elements resulted the two most expressed LTR retrotransposons. CER1 and LTRCER1 show similar expression profiles that recapitulate the general LTR expression profile, as they both result highly expressed in 1-, 2-, 4- and 8-cell stage and lowly expressed in 16-cell stage. Although the gastrulation process in *C. elegans* begins at the 26-cell stage (Nance et al., 2005), at the 16-cell stage the fate of all the embryo cells starts to be determined (Maduro, 2010; Sulston et al., 1983) and consequently the number of pluripotent cells drops down. The deep decrease of LTR expression in correspondence of the 16-cell stage may indicate that LTR are mostly expressed in undifferentiated cells, suggesting a role for LTR in the maintenance of pluripotency in *C. elegans* as already reported in higher organisms (Göke et al., 2015; Hackett et al., 2017). In particular, in mouse and human embryonic stem cells (hESCs), different classes of TEs are specifically expressed across a transcriptional spectrum of pluripotency (Göke et al., 2015; Hackett et al., 2017) with specific ERVs re-activation during somatic cells reprogramming into induced pluripotent stem cells (iPSCs) (Friedli et al., 2014). In addition to pluripotency, it has also been shown that LTR-derived nucleic acids may play a role in the activation of innate immune pathways in mammals (Kassiotis & Stoye, 2016). *C. elegans* lacks an adaptive immune system, however an innate immune system able to respond to external insults from bacteria, fungi and viruses has been described. In detail, the *C. elegans* innate immune system is composed by anti-viral and anti-microbial pathways: the anti-viral response is activated by viral double-strand RNA (dsRNA) and is mediated by the RNAi machinery, while the anti-microbial response is composed by different pathways whose induction led to the activation of secreted

effector proteins such as C-type lectin anti-microbial peptides (AMPs) (Ermolaeva & Schumacher, 2014). Although LTR retrotransposons activity in relation to infections has never been reported in *C.* elegans, in higher organisms it has been suggested that LTR elements may have an immuno-protective role triggering the innate immune system and thus activating the embryo to respond to pathogens. For instance, the human LTR retrotransposon HERVK has been described to encode a small accessory protein, Rec, homologous to HIV Rev, which allows nuclear export of viral RNAs triggering the innate antiviral responses through the detection of cytosolic viral RNA/DNA and proteins. Additionally, Rec overexpression in hESC resulted in an increased expression of viral restriction factors thus suggesting that the HERVK element might activate the innate immune response providing an immunoprotective effect and thus defending the human embryos from different classes of viruses (Grow et al., 2015).

LINE elements are mainly expressed in AB and E lineage cells

LINE elements resulted expressed in a small number of cell types mainly belonging to the 16-cell stage thus suggesting their expression may be highly regulated to occur specifically in defined cell subpopulations (**Figure-3.3B**). In particular, LINEs resulted expressed in E and E precursor cells (4-cell stage EMS cell, 8-cell stage E cell and 16-cell stage Ea and Ep cells) and in several AB cells of the 16-cell stage. Intriguingly, from the E lineage the intestine cells are generated (McGhee, 2007; Tintori et al., 2016), while the AB lineage gives rise to neuronal and non-neuronal tissues characterized by high concentration of nervous connections such as pharynx and epidermis (Tintori et al., 2016; Altun, Z.F. & Hall, D.H., 2011; Hobert, 2010; Chisholm & Xu, 2012; Altun, Z.F. & Hall, D.H., 2009). To date LINE expression in intestine precursor cells has never been described, whereas the expression of LINE in neurons and nervous system associated tissues has already been observed for higher organism (Baillie et al., 2011; Coufal et al., 2009; Erwin et al., 2014, 2016; Muotri et al., 2005; Perrat et al., 2013; Richardson et al., 2014) and will be discussed in the next paragraph. Analyses made at the single element level highlighted no LINE element displaying an expression profile capable to recapitulate the LINE global expression pattern. The general expression profile observed is the sum of different elements showing variable and element-specific expression dynamics. LINE2A and LINE2C1 are mostly expressed in the 4-cell stage EMS cell and in 16-cell stage MS cells, LINE2B is expressed in the 8-cell stage E cell and in the 16-cell

stage AB and MSx1 cells while LINE2F, that have an expression of ~5-fold with respect to LINE 2A, 2C1 and 2B, seems to be exclusively expressed in Ea and Ep cells of the 16-cell stage. This may suggest that different LINE elements might play different roles during *C. elegans* embryogenesis.

SINE elements are mainly expressed in AB lineage cells

SINE elements are expressed at higher levels with respect to LINE, but lower than LTR and DNA transposons (**Figure-3.3C**). SINE class in the *C. elegans* reference genome is composed of 2 elements (SINE1 and CELE45), with CELE45 being the only one resulting expressed. CELE45 is highly expressed in all the 16-cell stage AB cells, suggesting its specific expression in tissues deriving from this lineage as neurons, pharynx and epidermis precursors. Intriguingly, SINE CELE45 and LINE LINE2B elements display similar expression profiles with both elements showing expression in cell precursors giving rise to tissues characterized by a high concentration of nervous connections such as neurons, pharynx and epidermis. Expression and activity of non-LTR retrotransposons in neuronal and neuronal precursor cells have already been described in several Metazoans such as *Drosophila*, mouse and human and it is thus tempting to speculate a conserved mechanism at the basis of the observation of this phenomenon (Coufal et al., 2009; Muotri et al., 2005; Perrat et al., 2013). In particular, in this context, the expression and activation of non-LTR elements in *C. elegans* nervous cells during development may be associated with neuronal cell fate specification, leading to neuronal cells diversity and possibly affecting neural plasticity and synapsis formation.

DNA transposons have a heterogeneous expression profile

DNA transposons resulted expressed at higher levels with respect to SINE and LINE but lower than LTR (**Figure-3.3D**). DNA transposons are the most abundant and the unique described active TE class in the *C. elegans* genome (Bessereau, 2006; Laricchia et al., 2017). Their global expression in the *C. elegans* early embryo results relatively constant throughout the analysed stages and cell types. At the single element level, Chapaev1, CEMUDR1, PALTA3, and PALTTTAAA3 resulted the most expressed elements with specific expression profile characterising each of the four TEs. In particular, Chapaev1 resulted constantly expressed among the early embryo cell types with its expression recapitulating the overall expression of DNA transposons. The CEMUDR1 expression

profile resulted to be similar to the one displayed by LTR elements (**Figure-3.3A**) with detectable expression measured in 1-, 2-, 4- and 8-cell stages. Finally, PALTA3 and PALTTTAAA3 elements are lowly expressed in 1-, 2- and 4-cell stages, increase their expression at 8-cell stage reaching the peak in the AB cells of the 16-cell stage. Overall, PALTA3 and PALTTTAAA3 expression profile is similar to the one showed by LINE2B and CELE45. Together, these results suggest that DNA transposons have a heterogeneous expression profile that can be divided in i) constant, ii) LTR-like and iii) non LTR-like. DNA transposons are therefore the only TE class constantly expressed in all the cell types of the *C. elegans* early embryo.

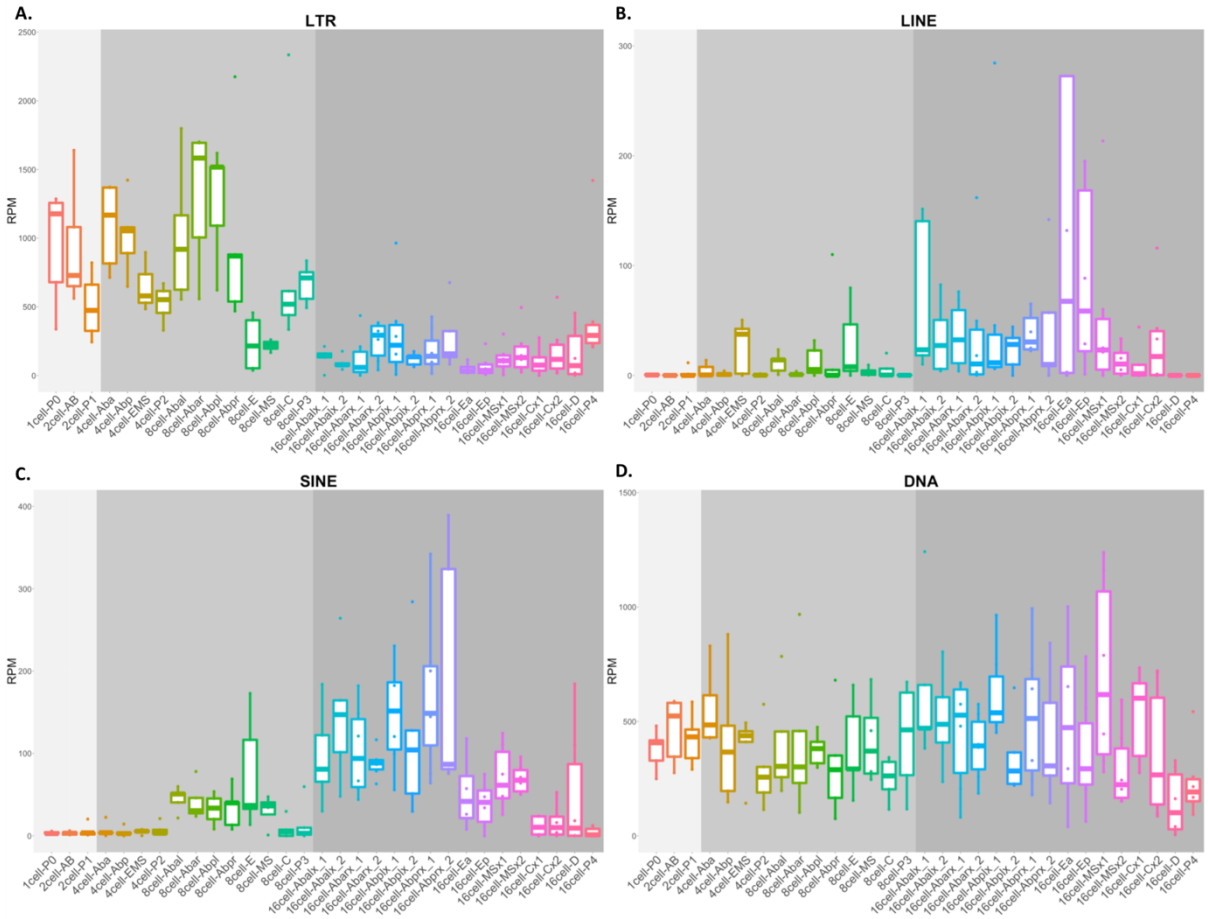


Figure-3.3: LTR, LINE, SINE and DNA transposon expression in the *C. elegans* early embryo.
(**A**) Expression of LTR. LTR elements are expressed in the initial stages (1-, 2-, 4- and 8-cell stages). (**B**) Expression of LINE. LINE elements are mostly expressed in E precursor and E descendant cells (EMS cell 4-cell stage, E cell 8-cell stage and Ea and Ep cells 16-cell stage) and in AB descendant cells at 16-cell stage. (**C**) Expression of SINE. In the *C. elegans* early embryo SINE class is characterized by the expression of a single TE (CELE45) which appears to be specifically expressed in AB descendant cells at the 16-cell stage. (**D**) Expression of DNA TE. DNA transposons show, as a whole, a constant expression profile throughout the *C. elegans* early embryo analysed cell types.

Expression of LTR elements correlates with the expression of genes associated to the innate immune response

In the latest years, several studies reported that, particularly during the embryogenesis, TEs may modulate gene expression (Garcia-Perez et al., 2016; Grow et al., 2015; Hackett et al., 2017; Rodriguez-Terrones & Torres-Padilla, 2018). Following this observation, TE expression profiles were statistically correlated with the gene ones. Although this analysis does not specifically elucidate any direct interaction between TEs and genes, it can highlight similarity in expression profiles that may suggest functional relationships. TE expression profiles were calculated as previously described by using the TEspeX pipeline. The gene expression profiles, instead, were retrieved from the work published by Tintori *et al*. (Tintori et al., 2016). To select TEs and genes with reproducible expression levels among replicates of the different cell types TEs and genes with expression values higher than 25 normalised counts in at least 3 replicates of at least 1 cell type were selected. This led to the selection of 11 TEs (**Table-3.1**) and 6,580 genes. Pearson correlation was then calculated and the gene/TE pairs showing an expression correlation with $R^2 >= 0.4$ or $<= -0.4$ and an FDR < 0.0001 were selected. This resulted in 1,300 positively and 169 negatively correlated gene/TE pairs. The 1,300 positive correlations are determined by 1,097 non redundant genes: 909 of these are correlated with 1 TE, 173 with 2 and 15 with 3 TE. The 169 negative correlations are determined by a set of 143 non-redundant genes, of which 117 are correlated with 1 TE and 26 with 2 TE. The correlation analysis highlighted that the LTR elements CER1 and LTRCER1, the SINE CELE45 and the DNA CEMUDR determine the highest number of correlations with CEMUR showing exclusively positive correlations while CELE45, CER1 and LTRCER1 both positive and negative ones (**Table-3.1**).

TE	Class	Correlations	Positive corr.	Negative corr.
CER1	LTR	363	323	40
LTRCER1	LTR	286	248	38
CEMUDR1	DNA	202	202	0
CELE45	SINE	197	106	91
PALTTTAAA3	DNA	104	104	0
PALTA3	DNA	96	96	0
CER3-1	LTR	84	84	0
LINE2F	LINE	77	77	0
CEREP1A	DNA	48	48	0
TC5	DNA	8	8	0
Chapaev-1	DNA	4	4	0

Table-3.1: number of positive and negative correlations for the 11 selected TEs.
1st column: list of the 11 TEs with RPM >25 in at least 3 replicates of at least 1 cell type. 2nd column: TE classes (DNA, LINE, SINE, LTR). 3rd column: total number of correlations between TEs and genes with RPKM >25 in at least 3 replicates of at least 1 cell type. 4th column: number of positive correlations. 5th column: number of negative correlations

To identify biological pathways associated to genes correlating with TE an enrichment analysis was performed on the genes belonging to the set of selected correlations. This analysis highlighted 66 pathways significantly enriched in the groups of genes determining the identified correlations. Of these, 36 pathways result associated to genes positively correlated with 5 TE (CEMUDR, PALTTTAAA3, LINE2F, LTRCER1 and CER1) (**Figure-3.4A**) whereas 31 pathways resulted associated to genes negatively correlated with a single TE (CER1) (**Figure-3.4B**). The enriched pathways resulting by positive correlations can be classified in 7 main groups: DNA repair, immune system, metabolism, metabolism of proteins, metabolism of RNA, signal transduction, and vesicle-mediated transport. The enriched pathways resulting by genes negatively correlated with CER1 can be classified in 7 main groups: cell cycle, DNA replication, immune system, metabolism of proteins, metabolism of RNA, signal transduction, and transport of small molecules. Here, it is important to point out that, in some cases, pathway annotations for *C. elegans* might have been inferred by homology, transferring the annotations of homologous genes from

more complex species. Results must therefore be interpreted with care and translated to the correct biological system. This is especially true for pleiotropic genes, with multiple functions, belonging to multiple pathways in complex organisms. The high number of functions for such genes is likely the result of their evolutionary recruitment into novel biological processes during the route leading to increased organismal complexity. For instance, our analysis identified a significant positive correlation between CER1 and innate immune system genes, a result in agreement with a possible involvement of CER1 in the embryonic activation of the innate immune response in *C. elegans*. On the other hand, this element also results negatively correlated with genes associated to the adaptive immune system, which is unlikely as *C. elegans* does not possess an adaptive immune response. However, these same genes are also annotated as belonging to the ubiquitination pathway, a function consistent with the biological system under analysis. Taking all this into account, it is tempting to suggest that the correlation analysis here performed may support the conclusion that genes associated to the innate immune response are significantly enriched among those whose expression correlates with LTR elements, reinforcing our previous observations.

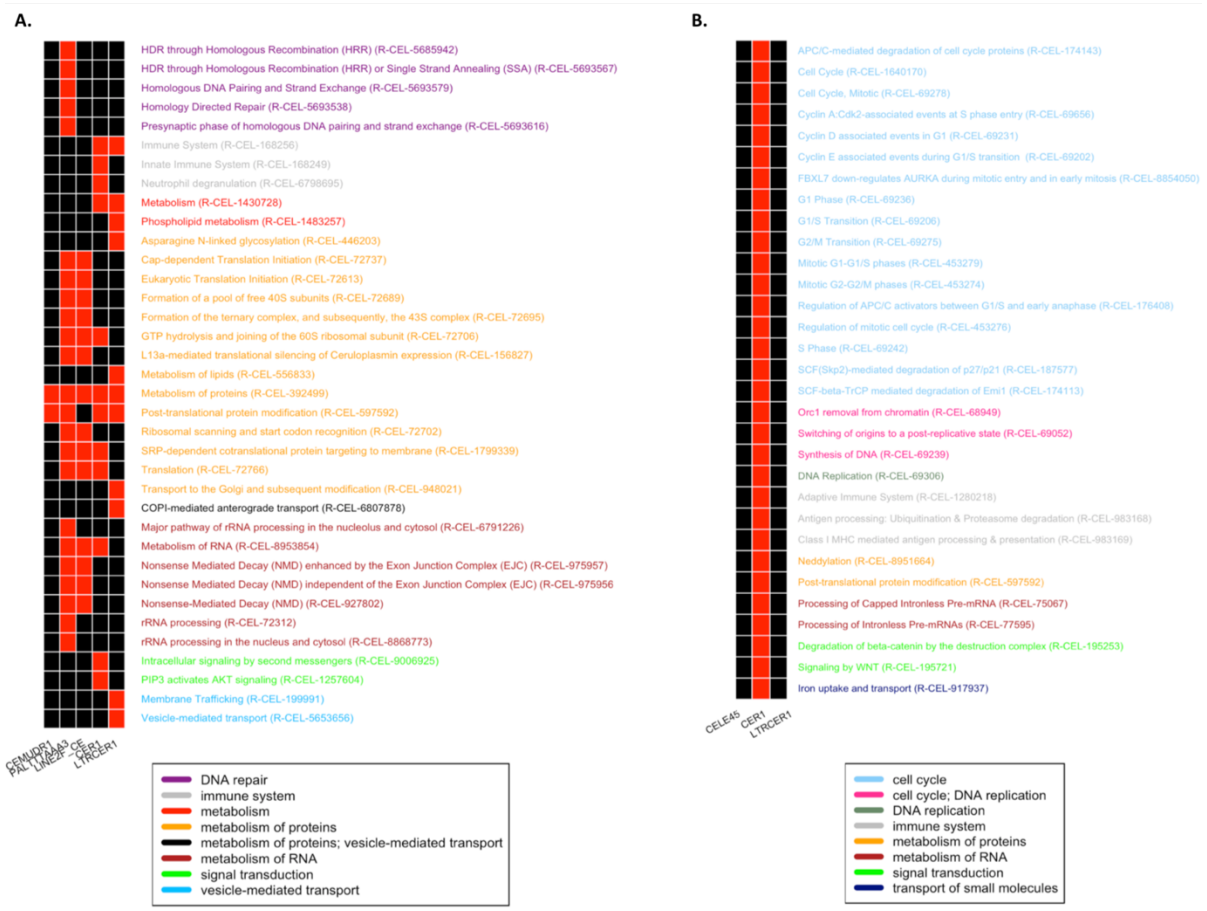


Figure-3.4: pathways enriched in genes positively and negatively correlated with TEs.
(**A**) Significantly enriched pathways associated to genes positively correlated with CEMUDR1, PALTTTAAA3, LINE2F, CER1 and LTRCER1. (**B**) Significantly enriched pathways associated to genes negatively correlated with CER1. Red color means presence, black absence.

3.3 Conclusions

Several studies have recently reported the expression of TE in mammalian embryos and the CNS suggesting their role in fundamental biological processes such as pluripotency maintenance, embryo viability and differentiation, brain functioning, evolution and diversification (Feschotte, 2008; Garcia-Perez et al., 2016; Hackett et al., 2017; Grow et al., 2015; Coufal et al., 2009; Erwin et al., 2016, 2014; Richardson et al., 2014; Baillie et al., 2011). In this study TEspeX, a recently developed bioinformatics pipeline able to quantify reads specifically mapping on TE, was used to investigate the TE expression in the *C. elegans* early embryo, from zygote to 16-cell stage. Outcomes resulting from the analyses suggest that, especially in neural tissues, a portion of reads mapping on TE cannot be distinguished by reads deriving from TE fragments embedded in annotated transcripts. These non-specific reads should therefore be discarded to avoid biases in the estimation of TE expression. In addition, the data shows that TE are expressed in the *C. elegans* embryo and that, despite their low level of expression, they present different expression profiles in different embryonic stages and cell types, suggesting a specific regulation during early development. A clear split of developmental TE expression levels in two phases is observable and, importantly, it results characterized by the expression of two different TE families: LTR and non-LTR. LTR elements resulted to be mostly expressed in the initial stages (1-, 2-, 4-, 8-cell stages). In particular, according to timing and territories of expression, LTR expression (mainly LTRCER1 and CER1 elements) in the initial developmental stages might play a role in the maintenance of pluripotency and/or in the innate immune response activation. On the other hand, non-LTR elements such as LINEs resulted mostly expressed in intestine precursor cells (E lineage) and, together with CELE45 (SINE), in 16-cell stage AB cells, the ones giving rise to neurons and tissues connected with nervous system. These results are consistent with the observations reporting the expression of non-LTR elements in nervous tissues of other organisms like fruitfly, mouse and human (Baillie et al., 2011; Coufal et al., 2009; Erwin et al., 2014, 2016; Muotri et al., 2005; Perrat et al., 2013; Richardson et al., 2014). DNA transposons are the most abundant TE fixed in the *C. elegans* genome and appear to be the only class of TEs expressed in all the cell types of the *C. elegans* early embryo.

Taken all together, these results suggest that, despite the low level of expression, TE transcription is finely regulated during the early embryo development of *C. elegans* and might be involved in specific developmental functions additionally having an immune-protective role. To further corroborate these results, a first indicative experiment would consist in silencing the most expressed LTR elements, LTRCER1 and CER1, followed by the measuring of the embryo susceptibility to viral and bacterial attacks and its capability to correctly develop and differentiate as similarly done in mouse embryo (Park et al., 2004) and in human ESCs and iPSCs (Lu et al., 2014).

3.4 Methods

Data collection and pre-processing

To quantify TE expression in the *C. elegans* early embryo, raw scRNA-seq public data was retrieved from Tintori *et al.* (Tintori et al., 2016). The dataset is representative of the 1-, 2-, 4-, 8- and 16-cell stages and is totally composed by 219 single cells accounting for 31 different cell types represented by 5-9 replicated each. Raw read files were downloaded from ENA EBI database using the PRJNA312176 accession code and 55 samples were discarded as not passing quality filters regarding whole embryo mRNA mass as state in the original manuscript.

TE expression analysis

TE expression analysis was measured using the TEspeX bioinformatics pipeline. Coding and non-coding transcript files were retrieved from Ensembl database (version 93 - WB235 genome version) (Zerbino et al., 2018) while TE consensus sequences were retrieved from RepBase database (Bao et al., 2015) discarding 16 sequences annotated as Satellite (SAT). As described in detail in Chapter2, reads are mapped by TEspeX against a reference transcriptome obtained merging the three input fasta files and assigning primary alignments score to the best scoring alignments. Reads deriving from best scoring alignments are then selected and considered as TE-specific when showing best scoring alignments exclusively against TE sequences and not against coding/non-coding transcripts. TE-specific reads have then been quantified on each TE sequence. Raw counts have finally been normalized on the total number of mapping reads and multiplied by 1,000,000 obtaining expression values indicated as reads per million mapped reads (RPM). Additionally, to test the accuracy of TEspeX in measuring TE expression in the *C. elegans* early embryo development, the same analysis was carried out taking advantage of the recently published SalmonTE pipeline (Jeong et al., 2018). SalmonTE measures TE expression levels quantifying RNA-seq reads on a set of provided TE consensus sequences using the Salmon tool (Patro et al., 2017). First, using the SalmonTE *index* mode (--te_only parameter) the index file was created providing as input the *C. elegans* TE consensus sequences file downloaded from RebBase database and used as input in the previously described TEspeX analysis. Then, taking advantage of the SalmonTE *quant* mode (--exprtype=count parameter) TE expression values were estimated. Finally, TE in

common between the two analyses were selected and the results generated by the two tools were compared.

TE/gene expression correlation and pathways analysis

To infer possible links between TE consensus sequence and *C. elegans* protein coding gene expression a correlation analysis between the normalised expression levels of the two group of features was performed. TE element expression values were quantified by TEspeX as previously described while gene expression values (RPKM) were retrieved from the Supplementary Table S2 of the paper published by Tintori *et al.* (Tintori et al., 2016). To select features showing a reproducible expression among the replicates of the same cell type, TEs and mRNAs displaying an expression value >= 25 RPM or RPKM in at least 3 replicates of at least 1 cell type were selected. Next, pairwise correlation analysis between TEs and coding genes using Pearson correlation test was performed calculating correlation coefficients using the *pearsonr* function of the *scipy* Python module (*stats* sub-module) and selecting correlations with R >= 0.4 or <= -0.4 and with a Benjamini & Hochberg FDR <= 0.0001. To identify potential pathway enrichment for genes involved in the selected correlations, a statistical over-representation test was performed using Panther tool (Thomas et al., 2003) (version: 13.1) specifying "*C. elegans*" input reference list, "Reactome pathways" as annotation dataset, performing the statistical test using Fisher's Exact with FDR multiple test correction and setting the significance threshold cut-off to 0.01 (FDR < 0.01).

Chapter 4

Genes and transposable elements transcriptionally activated upon zebrafish ZGA reside on highly transcribing genomic loci

4.1 Introduction

In all the Metazoans, the fertilised embryo cytoplasm and cytoplasmic factors entirely derive from the maternal oocyte (Schulz & Harrison, 2019; Wragg & Müller, 2016). These maternally provided factors support the development of the embryo during the initial developmental phases characterised by a transcriptionally quiescence of the zygotic genome (Tadros & Lipshitz, 2009). While delayed, the zygotic genome activation (ZGA) is nevertheless required for the early embryo to proceed developing (Jukam et al., 2017; Schulz & Harrison, 2019; Tadros & Lipshitz, 2009; Walser & Lipshitz, 2011). In zebrafish, the activation of the zygotic genome coincides with a sequence of events termed mid-blastula transition (MBT) (Wragg & Müller, 2016). At MBT, together with the cell division loss of synchrony and the elongation of the cell cycle length, a strong activation of zygotic transcription is promoted (Heyn et al., 2014; Wragg & Müller, 2016). However, some genes are transcriptionally activated earlier than MBT (starting at 1k-cell stage) with evidence of gene transcription being detected by *in vivo* imaging techniques as early as in the 64-cell stage and by transcriptomic analyses in the 512-cell stage (Giraldez et al., 2006; Hadzhiev et al., 2019; Heyn et al., 2014). These first zygotically transcribed genes, representing the ZGA minor wave, comprise micro-RNAs (miRNA) involved with the maternal transcript clearance as well as transcription factors and chromatin binding proteins playing crucial roles in the main wave of genome activation (Giraldez et al., 2006; Lee et al., 2013). Among this set of early transcribed genes, the earliest and highest expressed ones result the *miR-430* genes (Giraldez et al., 2006; Hadzhiev et al., 2019; Heyn et al., 2014). *miR-430* genes, by de-adenylating the maternal mRNAs, drive the maternal transcript clearance and their inhibition is not compatible with a proper embryonic development (Giraldez et al., 2006). Importantly, the *miR-430* genes are organised in a genomic locus characterised by a unique genomic structural organisation

(Hadzhiev et al., 2019). Indeed, the *miR-430* genes reside on a genic cluster composed by more than one hundred of almost identical *miR-430* gene copies located one immediately downstream to each other (Yavor Hadzhiev and Ferenc Müller personal communication). Additionally, not only the *miR-430* locus is characterised by a unique genomic structural organisation but also the entire long arm of the chromosome 4, on which the *miR-430* cluster reside, displays a unique and repetitive organisation (Y.H. and F.M. personal communication). The chromosome 4 long arm harbours a series of zinc-finger (*znf*) gene clusters that result transcribed during the main ZGA (White et al., 2017). Together, these observations have led to the speculation that, at ZGA, a local highly transcribing environment is generated on the long arm of the chromosome 4 thus facilitating the ZGA and probably functioning as an aggregation point for transcription factors (Hadzhiev et al., 2019).

To deeper investigate the genomic structural organisation of the genomic loci activated upon ZGA, RNA-seq raw reads have been retrieved from a publicly available dataset (generated by White et al., 2017, available at https://danio-code.zfin.org) and, upon the identification of the ZGA transcribed genes, it has been systematically addressed whether the involved genomic loci display specific structural organisations. Additionally, it has been assessed whether TEs transcription is similarly promoted upon ZGA and whether, as displayed in the mouse model, specific TE families facilitate the transcriptional activation of the nearby genes.

4.2 Results

Genes transcriptionally activated upon ZGA reside on the chromosome 4 and are enriched in genic clusters

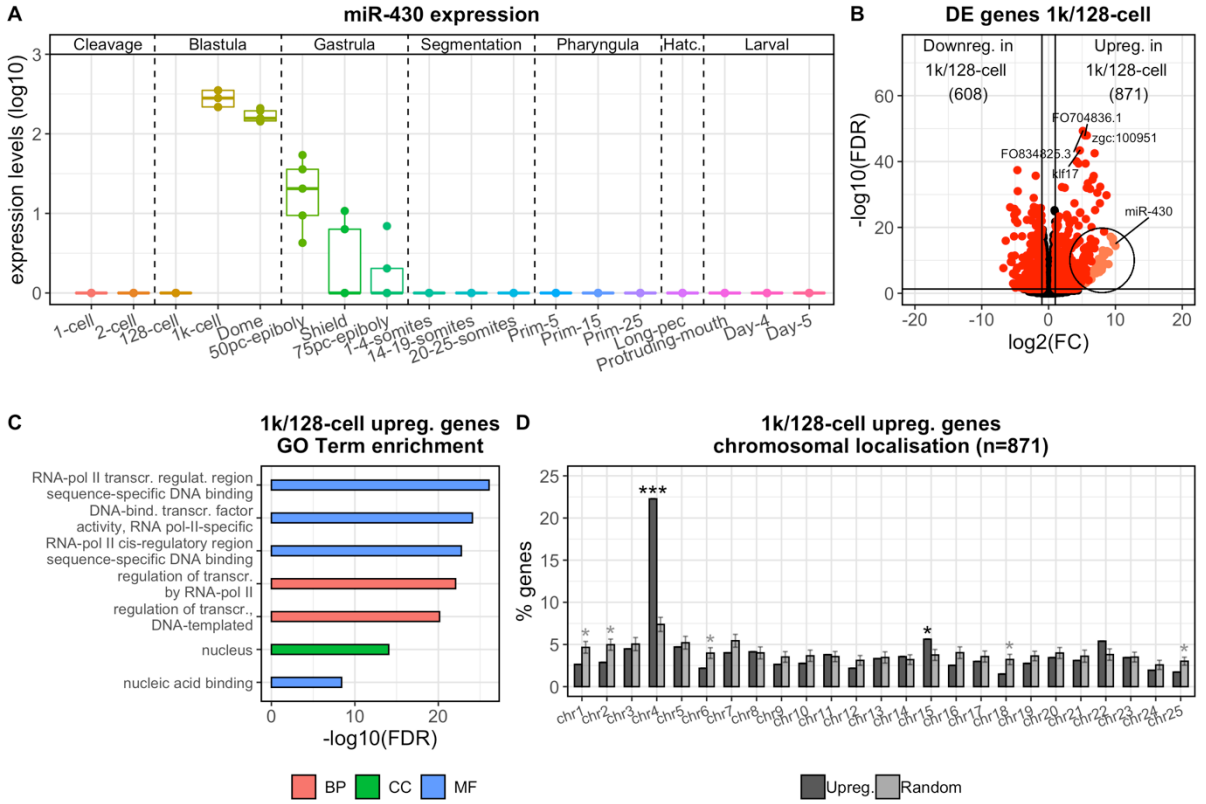
To investigate the transcriptional dynamics characterising the zebrafish development, and in particular the stages involved with the activation of the zygotic genome, raw reads have been retrieved from an RNA-seq publicly available dataset (White et al., 2017). The dataset is composed by 18 different stages, 5 replicates each for a total of 90 samples. Two developmental time points are relative to the zygote/cleavage stages, three to the blastula, gastrula, segmentation and pharyngula, one to the hatching and three to the larval stages. Having retrieved the raw RNA-seq reads, the expression levels of both coding and non-coding genes have been calculated in each of the 90 analysed samples. Since the *miR-430* has been described to be among the first genes, if not the first, to be transcribed in the zebrafish ZGA minor wave, its transcriptional profile has been used to define at which stage the ZGA is detectable in this dataset. Toward this end, the normalised reads mapping on each of the 54 *miR-430* gene copies annotated on the chromosome 4 of the zebrafish genome (version GRCz11) have been summed, for each of the 90 analysed samples. The *miR-430* transcriptional profile showed no evidence of transcription in the 1- and 2-cell stages (zygote/cleavage period) as well as in the 128-cell stage (early blastula) (**Figure-4.1A**). On the contrary, its transcription resulted strongly activated between the 128- and 1k-cell stages (mid blastula) next decreasing in the following stages (late blastula and gastrula) until resulting undetectable in the segmentation and following stages (**Figure-4.1A**). Thus, these data suggest that, at least in this dataset, the activation of the zygotic genome is detectable between the 128- and the 1k-cell stages. Considering that the zebrafish ZGA minor wave is detectable in transcriptomic data at 512-cell stage (Hadzhiev et al., 2019; Heyn et al., 2014) whereas the main wave is detectable at 1k-cell stage (Heyn et al., 2014), it is likely that between 128- and 1k-cell stages both waves are activated and therefore the resulting transcriptional signal is a mixture. From now on, I will refer to the transcriptional changes occurring between 128- and 1k-cell as 'ZGA related' without specifying whether they are minor or major waves as they both probably occur within the analysed time window.

Having defined that in this dataset the ZGA is detectable between the 128 and the 1k-cell stages, the coding and non-coding genes differentially expressed (DE) between these two time points have been identified. Almost 1,500 coding/non-coding genes resulted differentially expressed between 1k- and 128-cell stages (FDR<0.05 and log2FC >1 or <-1) with the majority of the DE genes (871 out of 1,479 genes) resulting upregulated. Assuming that the activation of the zygotic genome occurs between the 128- and 1k-cell stages, the genes resulting upregulated in the DE analysis are likely to be the genes whose transcription is activated upon zebrafish ZGA. Consistent with this observation, 52 out of the 54 *miR-430* annotated gene copies resulted among the 871 upregulated genes (**Figure-4.1B**). To define the gene ontology (GO) specifically associated to the set of the 871 upregulated genes, a GO enrichment analysis has been performed. The results highlighted 7 GO terms significantly enriched (FDR<0.05) (**Figure-4.1C**). Consistent with the analysed biological scenario, in which the most important event is the activation of the transcription of the embryo, all the enriched terms resulted associated to transcription regulation and in particular to RNA-polymerase II (RNA-pol II) transcription regulation (**Figure-4.1C**).

To investigate whether the 871 upregulated genes are homogenously distributed along the genome or reside on specific chromosomes, their chromosomal localisation has been identified and compared with that of randomly selected genes. The data showed how the upregulated genes resulted significantly enriched on chromosomes 4 and 15, with respect to the rest of the transcriptome (z-score FDR=1.6E-68 and 2.2E-02, respectively) (**Figure-4.1D**). While the fraction of the upregulated genes localised on the chromosome 15 is small (approximately 5% of the total upregulated genes), the fraction of upregulated genes residing on chromosome 4 is remarkable comprising almost 25% of such genes, including the *miR-430* cluster (**Figure-4.1D**). Consistent with previous observations (Hadzhiev et al., 2019; White et al., 2017), more than 90% of the upregulated genes on the chromosome 4 reside on its long arm furtherly reinforcing the speculation that, at ZGA, a local highly transcribing environment is generated along the long arm of the chromosome 4 probably facilitating the ZGA (Hadzhiev et al., 2019).

Next, to understand whether the upregulated genes, besides being enriched on chromosome 4, are also organised in specific genomic structural loci, the overlap of the

871 upregulated genes with genic clusters has been investigated. First, genic clusters have been defined as groups of at least 5 consecutive genes sharing at least one common functional domain. Next, the number of upregulated genes localised within such genic clusters have been calculated. The 1k- *versus* 128-cell upregulated genes resulted significantly enriched in genic clusters compared to the rest of the transcriptome with 35% of the upregulated genes resulting localised within genic clusters (z-score FDR=3.9E-69) (**Figure-4.1E**). To test the stage specificity of this feature, the same analysis has been repeated on the genes resulting upregulated in the 2 earlier (2-/1-cell and 128-/2-cell) and in the three later developmental stages (Dome/1k-cell, 50% epiboly/Dome and Shield/50% epiboly). The results displayed how before the 128-/1k-cell stages the upregulated genes are not significantly enriched in clusters either in 2- *versus* 1-cell or in 128- *versus* 2-cell (**Figure-4.1E**). On the contrary the genes resulting upregulated in the two stages following 128-/1k-cell stages (Dome/1k-cell and 50% epiboly/Dome) resulted enriched in clusters (z-score FDR=4.4E-210 and 2.9E-02, respectively) whereas in the last analysed stage (Shield/50% epiboly) they resulted significantly depleted from clusters (z-score FDR=2.9E-02) (**Figure-4.1E**). Finally, the functional domains of the specific genic clusters of the 1k- *versus* 128-cell upregulated genes have been identified. The 128-/1k-cell upregulated genes resulted significantly enriched in 13 clusters associated to 13 different functional domains (z-score FDR<0.05) with the majority of them being composed of genes with functions related with transcriptional regulation whereas the remaining sets were associated with functions related with protein degradation and immune system (**Figure-4.1F**). Importantly, 4 out of 13 genic clusters resulted composed by genes associated to zinc finger domains being thus consistent with previous observations highlighting the transcriptional activation of *znf* gene clusters during zebrafish ZGA (White et al., 2017).



F Genic clusters 1k upreg. genes are enriched in

	Pfam	Upreg. genes	Random genes	FDR	Description
1	PF00096	220	44.02	3.19e-132	Zinc finger, C2H2 type
2	PF13894	136	20.57	3.25e-129	C2H2-type zinc finger
3	PF13465	53	12.80	4.02e-29	Zinc-finger double domain
4	PF05729	67	25.19	1.22e-24	NATCH domain
5	PF00622	63	23.47	6.93e-22	SPRY domain
6	PF13516	63	23.73	3.77e-21	Leucine Rich repeat
7	PF13912	18	3.02	1.13e-19	C2H2-type zinc finger
8	PF13765	65	25.81	7.07e-19	SPRY-associated domain
9	PF07679	13	2.93	1.20e-07	Immunoglob. I-set domain
10	PF00089	8	1.63	2.68e-06	Trypsin
11	PF07686	26	11.27	1.67e-05	Immunoglob. V-set domain
12	PF00067	6	1.22	1.06e-04	Cytochrome P450
13	PF08783	7	2.94	3.19e-02	DWNN domain

Legend next page

Figure-4.1: genes transcriptionally activated upon ZGA reside on the chromosome 4 and are enriched in genic clusters.
(**A**) *miR-430* transcriptional profile. *miR-430* transcription is first detectable between the 128- and the 1k-cell stages suggesting the, in this dataset, the ZGA is detectable between these two stages. (**B**) Differentially expressed genes between 128- and 1k cell stages. Almost 1,500 genes result DE with the majority of them, including 52 out of the 54 annotated *miR-430* genes, resulting up- rather than downregulated. (**C**) GO term enrichment analysis. Results highlighted 7 GO terms enriched in the terms associated to the 1k-cell upregulated genes. The enriched GO terms are associated to biological pathways related to transcription regulation and especially to RNA-pol II-mediated transcription. (**D**) 1k-/128-cell upregulated genes chromosomal localisation. Upregulated genes are significantly enriched on chromosomes 4 and 15 while are depleted from 1, 2, 6, 18 and 25. (**E**) Fraction of upregulated genes organised in genic cluster. 1k-128-cell upregulated genes are enriched in genic clusters defined by at least 5 consecutive genes associated to the same functional domain whereas this feature is not displayed by genes resulting upregulated in earlier time points (128-/2-cell and 2-/1-cell). Genes resulting upregulated in Dome/1k-cell and in Dome/50% epiboly display the same cluster enrichment whereas the Shield/50% epiboly upregulated genes display the opposite trend being significantly depleted from genic clusters. (**F**) Genic clusters the 1k-/128-cell upregulated genes are enriched in. The majority of the clusters is composed of genes with functions related with transcription regulation whereas the remaining sets were associated with functions related with protein degradation and immune system Remarkably, 4 out of 13 genic clusters resulted composed by genes associated to zinc finger domains. In D and E mean ± standard deviation is represented by the bars. (*FDR<0.05, **FDR<0.01, ***FDR<0.001. FDR values refer to z-score derived BH FDR-corrected P-value).

Transposable element expression slightly increases upon the zygotic genome activation reaching the peak of expression during gastrulation

Having defined the coding and non-coding genes transcriptionally activated upon zebrafish ZGA, it was next investigated whether also TEs undergo a similar transcriptional activation. Given that the TE expression landscape in the zebrafish embryo is a piece of knowledge currently lacking in the literature, before addressing which TEs result differentially expressed upon zebrafish ZGA, the TE transcriptional profile in the zebrafish embryo has been defined. Toward this end, the TE expression levels have been quantified in each of the 90 analysed samples. In order to avoid the quantification of RNA-seq reads potentially deriving from the co-transcription of TE fragments as part of coding/non-coding genes, the TE expression has been calculated at the TE consensus sequence level using the TEspeX tool, capable to exclude all the RNA-seq reads mapping ambiguously on both TEs and coding/non-coding transcripts (described in the Chapter 2 of this thesis).

At first, the global TE expression profile has been assessed calculating the mean expression of all the analysed TE consensus sequences (n=2,282). The results highlighted how, before the activation of zygotic genome in 1-, 2- and 128-cell stages, a small amount of TE mRNAs is detectable (**Figure-4.2A**). These mRNAs are likely to be part of the maternal factors deposited in the oocyte cytoplasm and inherited by the zygote. Following the activation of the zygotic genome, between the 128- and the 1k-cell stages, the TE transcription is slightly increases being furtherly promoted in ZGA advanced phases between the 1k-cell and the Dome stages. After the Dome stage, during the gastrulation period, the TE expression reaches its maximum then gradually decreasing in the following stages with a moderate peak of expression at the hatching stage (**Figure-4.2A**).

The TE transcriptional dynamics have next been investigated for each of the four main TE classes (DNA, LTR, LINE and SINE). The mean TE expression of the TE consensus sequences belonging to DNA, LTR, LINE and SINE classes has thus been calculated. The expression profiles of the TE classes highlighted how SINEs are the TE class showing the highest amount of mRNAs in the transcriptionally quiescent 1- and 2-cell stages whereas in these stages a small quantity of mRNAs is detectable for DNA, LINE and LTRs (**Figure-**

4.2B). Moving from the 2- to the 128-cell stage (shift from cleavage to blastula), the SINE mRNA levels decrease, whereas the DNA, LINE and LTR ones remain constant. Intriguingly, at the ZGA onset, between the 128- and 1k-cell stages, the expression of all the TE classes slightly increases. While DNA, LINE and LTR element expression increases further between the 1k-cell and the Dome stage, the SINE expression decreases between these two stages. Moving from Dome to 50% epiboly and thus entering in the gastrulation period, the DNA, SINE and LINE element expression slightly decreases remaining constant in all the remaining stages. On the contrary, upon gastrulation, the LTR expression reaches its maximum then gradually decreasing in the following stages with a moderate peak of expression at the hatching stage (**Figure-4.2B**).

Importantly, the LTR elements resulted the class of TEs showing the highest expression levels. Therefore, to understand whether this signal is given by some specific LTR families, the expression profile of the LTR elements has been subsequently subdivided among the 6 families composing the zebrafish LTR class: Copia, ERV, LTR (general LTR classification), DIRS, Gypsy and Pao. The data showed how low levels of the mRNAs deriving from all the 6 LTR families were detectable before ZGA, in 1-, 2- and 128-cell stages (**Figure-4.2C**). Importantly, following ZGA, between 128- and 1k-cell stages, the transcription of all the families, except Copia, is slightly activated. Next, the expression of all the 6 LTR families increases between 1k- and Dome stages. Unlike for all the other LTR families where the expression increase is moderate, the ERV and general LTR families expression is remarkably enhanced between 1k- and Dome stages. Shifting from Dome to 50% epiboly and thus entering in the gastrulation period, the Copia, DIRS, Gypsy and Pao expression remains constant until the last analysed stages (larval period). On the contrary, the expression of both ERV and general LTR is furtherly increased with LTR reaching their peak of expression during gastrulation and then gradually decreasing. Interestingly, ERV shows the maximum expression levels between the gastrulation and the segmentation periods and then gradually decreases (**Figure-4.2B**).

Altogether, these data indicate that different periods of the zebrafish embryonic development are characterised by the expression of different TE classes. Before the activation of the zygotic genome the SINE elements are the one showing the highest amount of mRNAs in the zygote, probably of maternal origin. Importantly, at the ZGA

onset, the transcription of all the main TE classes (DNA, LTR, LINE and SINE) is activated, even if only modestly. Finally, in later stages, and in particular during the gastrulation period, LTR elements, and especially the general LTR and ERV families, resulted the highest expressed TE class.

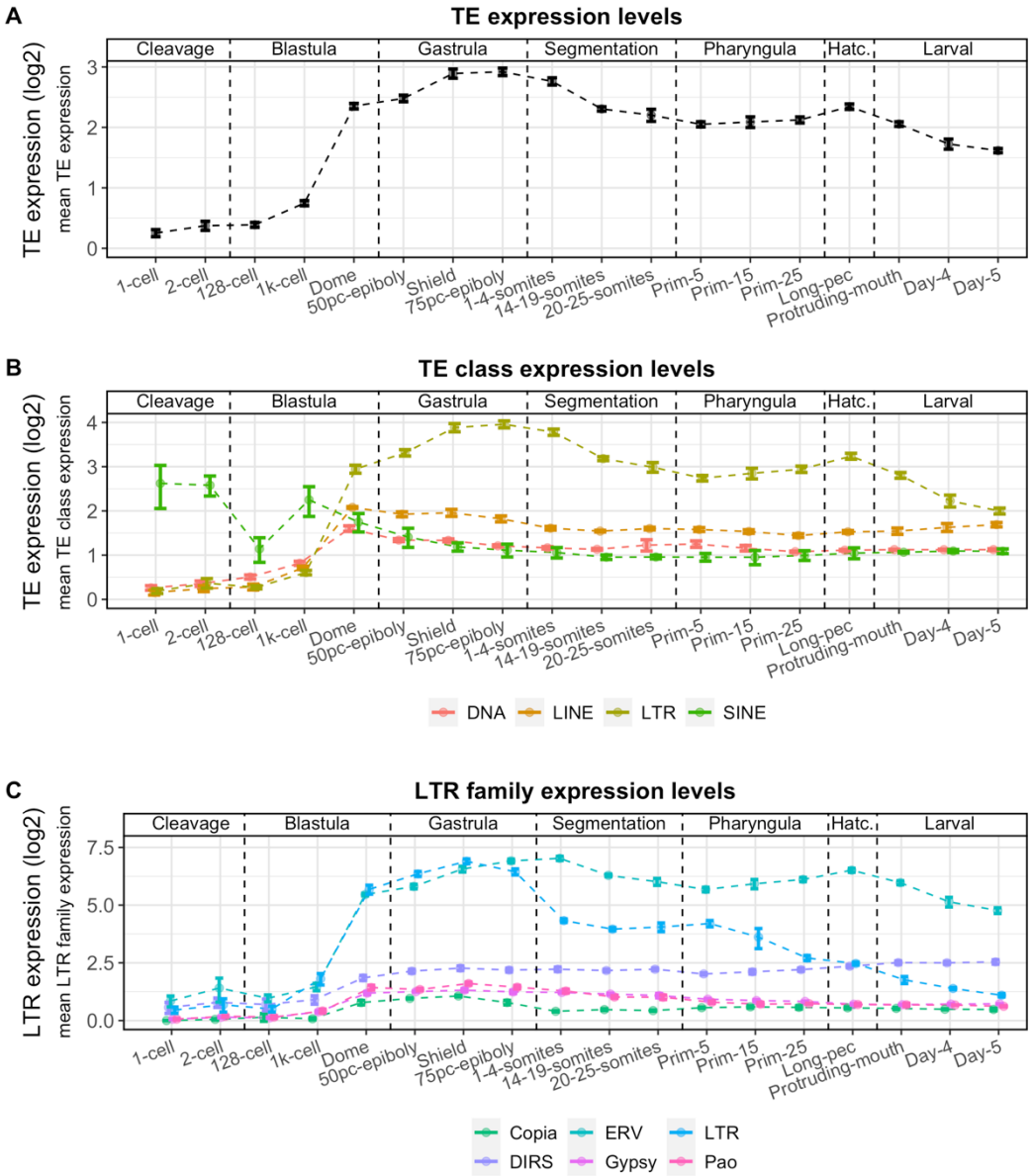


Figure-4.2: transposable element expression slightly increases upon the zygotic genome activation reaching the peak of expression during gastrulation.
(**A**) TE global expression profile. Low levels of TE-derived mRNAs are detectable before ZGA in the 1-, 2- and 128-cell stages. At ZGA (between 128- and 1k-cell stages) TE transcription is slightly enhanced being furtherly promoted between Dome and 1k-cell stage. TE expression peaks during the gastrulation period next gradually decreasing. TE expression levels are calculated as the mean of the expression of all the analysed TEs. (**B**) DNA, LINE, LTR and SINE TE classes expression profile. Except for SINEs, low levels of TE mRNAS are detectable before ZGA. At ZGA, the transcription of all the TE families is slightly enhanced with DNA, LINE and LTR expression being remarkably increased in the following stage, between 1k-cell stage and Dome. After Dome stage, entering in the gastrulation period, LTR elements reach the peak of expression whereas the expression levels of all the other TE families remain constant. For each class, the TE expression levels are calculated as the mean of the expression of all the analysed TEs belonging to that TE class. (**C**) LTR family expression profile. The expression profile of the 6 LTR families have been plotted highlighting how elements belonging to the ERV and to the general LTR families are the most expressed ones. Their transcription is remarkably increased between the 1k-cell and the Dome stages with the general LTR family expression decreasing after gastrulation whereas the ERV ones during segmentation. TE expression levels are calculated as the mean of the expression of all the analysed TEs belonging to each LTR TE family. (The expression profiles are represented as mean ± standard deviation of the TE expression calculated among the 5 biological replicates representing each of the 18 analyses samples).

Transposable elements transcriptionally activated upon ZGA reside on chromosome 4 and are enriched in genic clusters

Having defined the TE transcriptional dynamics during the zebrafish development and having highlighted how all the TE classes get transcriptionally activated, even if modestly, upon ZGA, the differentially expressed TEs between the 128- and the 1k-cell stages have been identified. At first, the DE analysis has been performed on the TE consensus sequence expression levels previously calculated with TEspeX (now on referred as 'TE consensus') and secondly it has been performed on the expression levels of each TE single locus annotated in the zebrafish genome (now on referred as 'TE loci).

At the TE consensus level, 70 TE (out of 2,282 analysed) resulted differentially expressed between the 128- and the 1k-cell stages (FDR<0.05 and log2FC >1 or <-1) (**Figure-4.3A**). Importantly, the large majority of the differentially expressed TEs resulted up- rather than downregulated upon the activation of the zygotic genome (65 out of 70). The most significantly upregulated TE resulted the DNA element hAT-N203 and, intriguingly, the second one an LTR element belonging to the Gypsy family (Gypsy117) (**Figure-4.3A**). Overall, except for SINEs, at least one TE of each of the 4 main classes (DNA, LTR, LINE and SINE) resulted differentially expressed between the analysed stages highlighting the heterogeneity of the class of TEs activated upon zebrafish ZGA and being consistent with previous observations displaying how all the TE classes get transcriptionally activated upon ZGA (**Figure-4.2B**).

To better investigate the transcriptional dynamics underlying the TE transcriptional activation upon zebrafish ZGA, and to investigate the genomic loci from where the TE transcription is activated, the expression of each TE single locus annotated in the zebrafish genome has been calculated. Intriguingly, more than 10 thousand TE loci resulted expressed in the 1k-cell stage (out of more than the 2 million analysed). However, a small fraction of such TEs resulted differentially expressed between the 128- and the 1k-cell stages with 132 loci resulting differentially expressed (**Figure-4.3B**). The majority of the differentially expressed TE loci resulted upregulated in 1k- compared to 128-cell stage (99) and, consistent with the previous TE consensus analysis, the most significantly upregulated TE locus resulted a Gypsy117 element (**Figure-4.3B**). This Gypsy elements is located on the chromosome 20 (genomic coordinates:

chr20:51,190,076-51,196,640) on the reverse strand with respect to the reference genome, and its TSS is located 99 nt upstream to the TSS of the hsp90ab1 gene, transcribed on the opposite strand of the Gypsy117 element (**Figure-4.4**). However, given the different profile of expression of the two elements and the lack of chimeric reads mapping on both the Gypsy117 element and on the hsp90ab1 gene, it is unlikely that the Gypsy element may be somehow involved with the transcription of the hsp90ab1 gene (**Figure-4.4**).

To define whether specific TE families resulted enriched in the set of upregulated TEs, compared to the rest of the transposome, the fraction of upregulated TE loci belonging to each TE family has been counted and compared with the one of the annotated TEs. The data showed how TEs belonging to the hAT (DNA) and Gypsy (LTR) families resulted significantly enriched in the set of upregulated TEs with respect to the rest of the transposome (Z test FDR=3E-02, both families) (**Figure-4.3C**). Indeed, almost 25% of the upregulated TE loci belong to the hAT family (24 TE single loci) whereas they cover approximately the 13% of the zebrafish transposome. Similarly, even if with lower fractions, almost 8% of the upregulated TE loci belongs to the Gypsy family (7 TE single loci) whereas they represent less than 3% of the zebrafish transposome (**Figure-4.3C**).

Next, starting from the previous observations highlighting that the 1k- *versus* 128-cell upregulated genes are enriched on chromosome 4, compared to the rest of the transcriptome, the chromosomal localisation of the 99 upregulated TE loci has been defined. The results displayed how the 1k-/128-cell upregulated TE loci resulted enriched on chromosomes 4 (z-score FDR=1.1E-04) as well as on chromosomes 8 and 15 (z-score FDR=9.7E-04 and FDR=1.0E-03, respectively) (**Figure-4.3D**). Importantly, also the coding/non-coding genes resulted significantly enriched on chromosome 4 and 15, thus confirming the hypothesis that, upon zebrafish ZGA, the activated loci are located on chromosome 4 and, with less statistical support, on chromosome 15.

Finally, to understand whether the upregulated TEs are organised in specific genomic structural loci, the localisation within genic clusters of the 99 1k- *versus* 128-cell upregulated TEs has been investigated. As previously shown for the upregulated genes, the 1k-/128-cell upregulated TEs resulted significantly enriched in genic clusters

compared to the rest of the transcriptome (z-score FDR=9.7E-31) (**Figure-4.3E**). Importantly, neither the 128-/2-cell nor the 2-/1-cell upregulated TEs displayed this feature. On the contrary the TEs resulting upregulated in the following three stages (Dome/1k, 50% epiboly/Dome and Shield/50% epiboly) displayed the same cluster enrichment suggesting that the enrichment in cluster may be a specific feature of the TE expressed both at the onset and after the activation of the zygotic genome (z-score FDR=0.0; 3.0E-206 and 2.9E-13, respectively) (**Figure-4.3E**).

Overall, it is intriguingly to note how, as already shown by the coding/non-coding genes, large fractions of the 1k- *versus* 128-cell upregulated TEs are enriched on chromosome 4 and in genic clusters, with respect to the rest of the transcriptome. However, it has to be taken into consideration the small absolute numbers of the loci involved (99 upregulated loci out of more than 2 million analysed) that overall suggest how not a remarkable portion of TEs is essentially transcribed at ZGA.

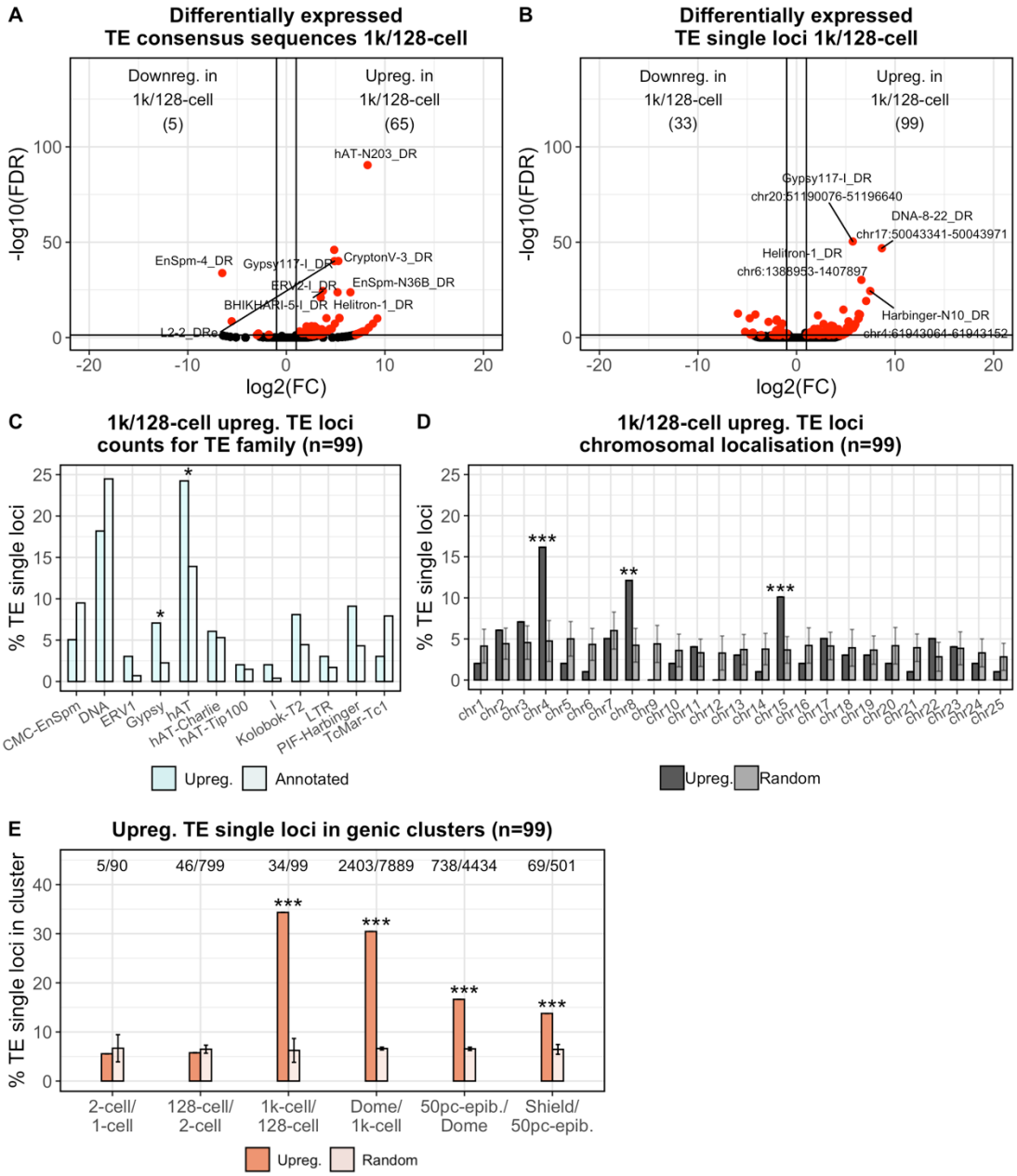


Figure-4.3: transposable elements transcriptionally activated upon ZGA reside on chromosome 4 and are enriched in genic clusters.

(**A**) Differentially expressed TE consensus sequences. Of the2,282 analysed, 70 TEs resulted DE between 128- and 1k-cell stages. Of these, 65 resulted upregulated in 1k- compared to 128-cell stage. The two TEs most significantly upregulated resulted the DNA element hAT-N203 and the LTR element Gypsy117 (**B**) Differentially expressed TE loci. Of the more than 2 million analysed, 132 TEs resulted DE with 99 resulting upregulated in 1k- compared to 128-cell stage. Consistent with (**A**), the most significantly upregulated TEs is a Gypsy117 element located on the chromosome 20. (**C**) Upregulated TE loci counted for TE family. TE belonging to the Gypsy and hAT TE families resulted significantly enriched in the set of upregulated TE loci compared to the rest of the transposome. (**D**) Chromosomal localisation of the upregulated TE loci. The 1k-128-cell stage upregulated TE loci resulted significantly enriched on chromosomes 4, 8 and 15. (**E**) Fraction of upregulated TEs overlapping genic clusters. The 1k-/128-cell upregulated TEs resulted significantly enriched in genic clusters compared to the rest of the transcriptome. Neither the 128-/2-cell nor the 2-/1-cell upregulated TEs displayed this feature. TEs resulting upregulated in the following three stages (Dome/1k, 50% epiboly/Dome and Shield/50% epiboly) displayed instead the same cluster enrichment. In C, D and E mean ± standard deviation is represented by the bars. (*FDR<0.05, **FDR<0.01, ***FDR<0.001. In C FDR values refer to Z test BH FDR-corrected P-value whereas in D and E to z-score derived BH FDR-corrected P-value).

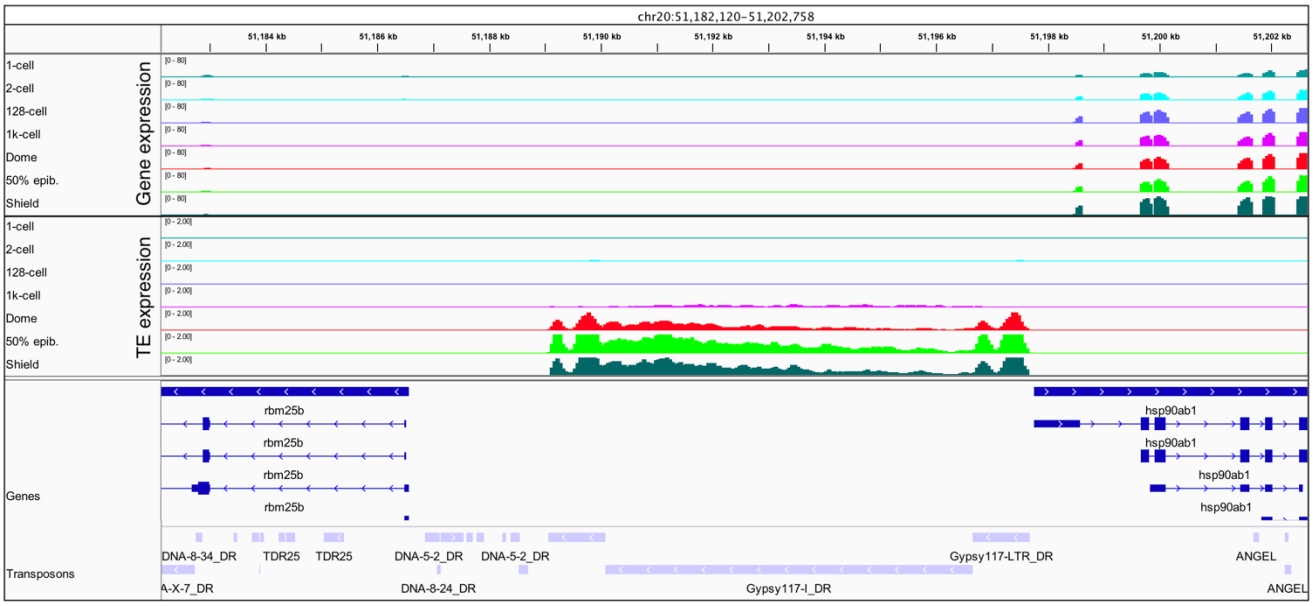


Figure-4.4: screenshot of the TE locus most significantly upregulated between 128- and 1k-cell stages.

The full-length Gypsy117 element is located between the not expressed rbm25b gene and the heat shock protein hsp90ab1. The Gypsy117 element is transcribed on the reverse strand with respect to the reference genome and its transcription is slightly activated between 128- and 1k-cell stages (blue and violet, respectively) furtherly increasing in 1k-cell (red) and Dome stages (electric green). The TSS of the Gypsy element is 99 nt upstream to putative TSS of the hsp90ab1. hsp90ab1 gene is transcribed in the opposite strand respect to the Gypsy117 element and its mRNAs are detectable since the pre-ZGA stages suggesting their maternal origin.

The ZGA upregulated genes are not enriched in TE sequences

During the earliest phases of the murine ZGA, a specific subfamily of TEs, known as MERVL, promotes the transcription of hundreds of nearby genes by providing an alternative exon/promoter and thus leading to the generation of TE-gene chimeric transcripts (Macfarlan et al., 2012; Peaston et al., 2004). Starting from this observation, it was asked whether in zebrafish a similar scenario exists. Toward this end, the 871 1k-*versus* 128-cell upregulated genes enrichment or depletion in TE sequences nearby the TSS has been investigated. First the gene TSS coordinates have been defined exploiting zebrafish early embryo CAGE-seq data and elongated by 100 nt both up- and downstream (see Methods), next the fraction of upregulated genes overlapping with the elongated TSS for each of the TE family annotated in the zebrafish genome has been calculated and compared with the one of randomly selected genes. The data highlighted how the 871 upregulated genes did not result either enriched or depleted of TE sequences of any TE family (**Figure-4.5A**). Next, instead of considering all the 871 upregulated genes together, the upregulated genes composing each of 13 genic clusters previously identified (listed in **Figure-4.1F**) have been analysed individually. The data showed how the upregulated genes of 7 out of 13 clusters displayed no TEs either enriched or depleted (**Figure-4.5B**). On the contrary, the genes organised in the remaining 6 clusters resulted enriched in a diversified set of TE families mostly belonging to the DNA TE class (TcMar, hAT-Charlie, Kolobok-T2 and CMC-EnSpm). Only the Trypsin cluster genes showed an enrichment in LTR TE families nearby the gene TSS. However, given the small number of genes composing the cluster (n=8) no general considerations can be inferred from this result (**Figure-4.5B**).

Finally, reasoning that the TE-mediated transcriptional activation of early expressed genes might be a specific phenomenon observable exclusively in the earliest phases of the ZGA, the same analysis has been performed on the genes resulting upregulated in the earlier time point. Toward this end, first the genes significantly upregulated between the 128- and the 2-cell stages have been identified and next, the fraction of upregulated genes overlapping with the elongated TSS each TE family annotated in the zebrafish genome has been calculated and compared with the one of randomly selected genes. The data showed how the 128- *versus* 2-cell upregulated genes did not result enriched in any TE sequences nearby the TSS (**Figure-4.5C**). On the contrary, the TSS of such genes resulted

significantly depleted in the DNA families DNA and CMC-EnSPM and in the SINE family 5S-Deu-L2 (**Figure-4.5C**).

Overall, these results suggest how no upregulated genes either in 1k-/128-cell or in 128-/2-cell stages are significantly enriched in TE sequences nearby the gene TSS. Nevertheless, some technical issues of this analysis are worth to be discussed here. As previously stated, the zebrafish ZGA minor wave is known to occur at different stages rather than the ones here considered (1k-/128-cell and 128-/2-cell). It is thus likely that the 128-cell stage represents a too early stage for the ZGA minor wave to be detected whereas the 1k-cell stage a too late one. Indeed, the genes upregulated between the 128- and the 2-cell stages are more likely to be genes whose expression changes due to the use of different poly-adenylation signals rather than genes activated upon the ZGA (Ulitsky et al., 2012). Instead, the genes resulting upregulated between the 1k- and the 128-cell stages are likely to be a mixture of genes activated upon both minor and major ZGA waves as the interval between the two analysed time points is quite long with the 256- and the 512-cell stages missing from this dataset (Hadzhiev et al., 2019; Wragg & Müller, 2016). Considering all this together, it is likely that such technical issues might have negatively affected the analysis outcomes.

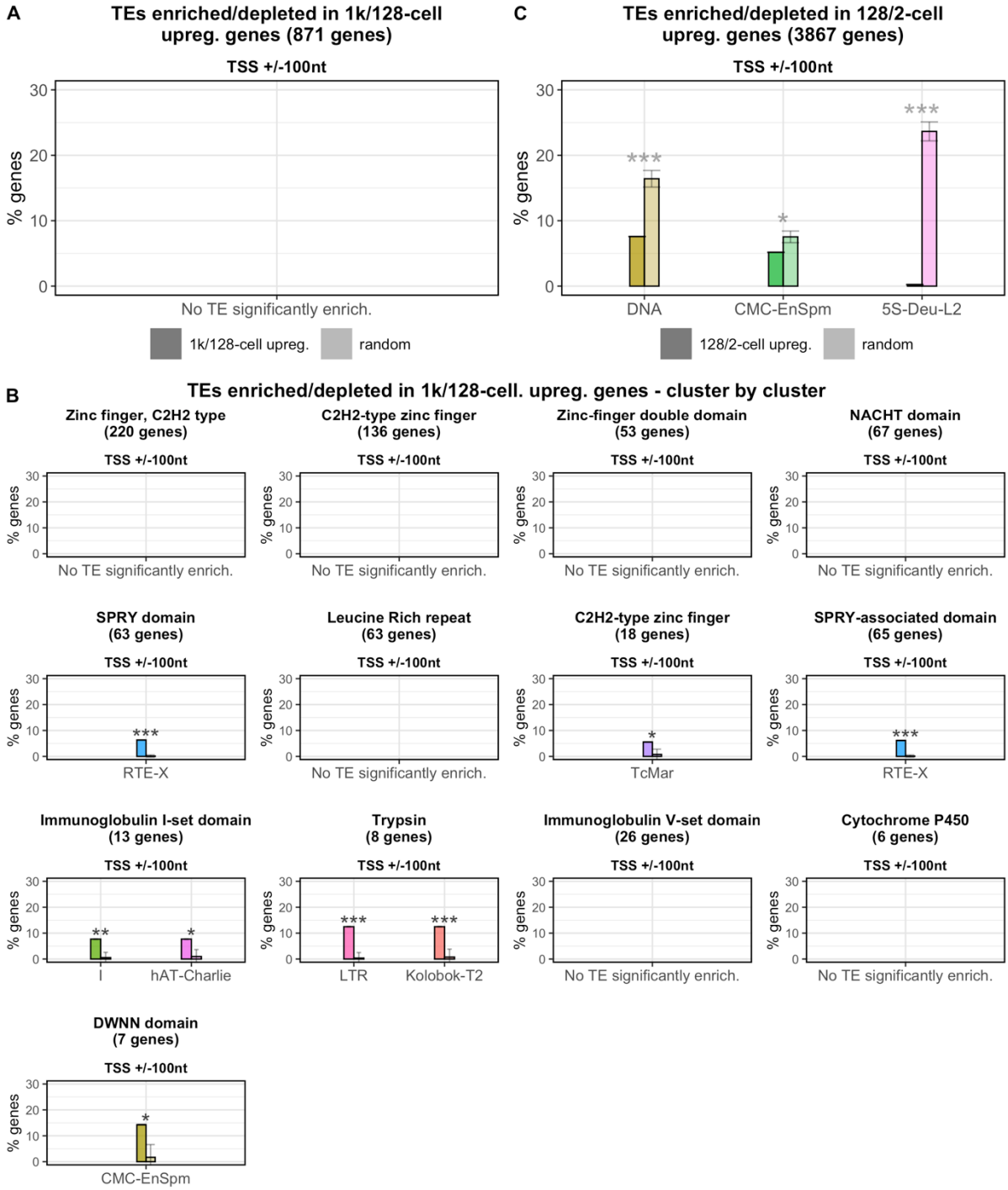


Figure-4.5: the ZGA upregulated genes are not enriched in TE sequences.
(**A**) 1k/128-cell upregulated genes TE enrichment. The fraction of 1k/128-cell upregulated genes overlapping with their TSS (+/- 100 nt) each TE family annotated in the zebrafish genome has been calculated and compared with randomly selected genes. No upregulated genes resulted enriched in any TE families (**B**) 1k/128-cell upregulated genes TE enrichment – cluster by cluster. The same analysis as in (**A**) has been performed on the genic clusters the 1k/128-cell upregulated genes are enriched in. The upregulated genes of 7 out of 13 clusters displayed no TE enrichment. The genes organised in the remaining 6 clusters resulted enriched in a diversified set of TE families mostly belonging to the DNA TE class. (**C**) 128-/2-cell upregulated genes TE enrichment. The 128-/2-cell upregulated genes resulted not enriched in any TE sequences yet resulting significantly depleted of DNA, CMC-EnSpm and 5S-Deu-L2 TE families. (*FDR<0.05, **FDR<0.01, ***FDR<0.001 - FDR values refer to z-score deriving BH FDR-corrected P-value). In all the plots mean ± standard deviation is represented by the bars. Only significant results for which at least 5% of the genes overlapped each TE family have been plotted.

4.3 Discussion

Mounting evidence has described how, during the activation of the Metazoan zygotic genome, remarkable transcriptional bursts occur (Schulz & Harrison, 2019; Tadros & Lipshitz, 2009). In the zebrafish model one of the earliest gene to be transcribed at the ZGA onset is the *miR-430* (Giraldez et al., 2006). Importantly, the *miR-430* is characterised by a unique genomic organisation being structured in a genic cluster comprising more than one hundred of gene copies (Y.H. and F.M., personal communications). However, the structural genomic organisation of other early expressed genes as well as the contribution of TEs to the activation of the zygotic genome remain to be addressed. To investigate such issues, raw reads from an RNA-seq dataset covering 18 zebrafish developmental stages, including the early embryo, have been retrieved and the transcriptional dynamics characterising coding/non-coding genes and TEs, as well as the genomic organisation of the involved loci, have been investigated.

The results highlighted how 871 coding/non-coding genes result upregulated upon ZGA, that in this dataset is detected between the 128- and the 1k-cell stages. The upregulated genes resulted involved with functions mainly related to RNA-pol II mediated transcription regulation resulting additionally enriched on chromosome 4 and in genic clusters when compared to the rest of the transcriptome. Importantly, although their activation upon ZGA resulted modest, also the 128-/1k-cell upregulated TE loci resulted additionally enriched on chromosome 4 and in genic clusters. Importantly, the localisation of a notable portion of the ZGA activated genes and TEs on the chromosome 4 is consistent with previous observations suggesting that the chromosome 4, in this specific biological context, is a transcriptional hotspot characterised by a unique organisation that creates a local highly transcribing environment, which may function as an aggregation of transcription factors (Hadzhiev et al., 2019). Additionally, the genic cluster enrichment of both 1k-/128-cell upregulated genes and TEs might be an observation furtherly supporting the hypothesis that the generation of a transcriptional dense environment might facilitate the ZGA.

Finally, it was assessed whether, besides the chromosome 4 localisation and the genic cluster organisation, also TEs may influence the transcriptional activation of the ZGA

upregulated genes by being located nearby the gene TSS as displayed by the MERVL elements during the mouse early embryogenesis. Taken together, all the data showed no evidence supporting the conservation of such a mechanism in the zebrafish embryo. Nevertheless, some technical issues might have negatively affected the analysis outcomes as the embryonic stages here analysed (1k-/128-cell and 128-/2-cell) are likely to be not completely appropriate for the detection of such an early and time-specific mechanism probably occurring during the ZGA minor wave. Indeed, the genes resulting upregulated between the 1k- and the 128-cell stages are likely to be a mixture of genes activated upon both minor and major ZGA waves whereas the 128-/2-cell upregulated genes are more likely to be genes whose expression changes due to the use of different poly-adenylation signals rather than genes activated upon the ZGA minor wave (Ulitsky et al., 2012). Considering all this together, it is likely that such technical issues might have negatively affected the analysis outcomes.

In summary, these results have extensively confirmed previous gene-specific observations suggesting how the genes activated upon zebrafish ZGA reside on transcriptionally dense genomic compartments. Additionally, these data have provided preliminary observations on how these genomic structures may be crucial for the activation of the transcription of both coding/non-coding genes and TEs. However, additional evidence has to be provided to reinforce such findings. In particular the timings by which this phenomenon occurs have to be furtherly investigated possibly producing samples from the specific ZGA minor wave stage in order to define, with a better resolution, whether TE are transcriptionally activated upon ZGA minor wave and whether they may influence the transcription of early expressed genes located nearby, as shown in the murine model.

4.4 Methods

Data collection and pre-processing

To study the transcriptional dynamics characterising the zebrafish embryonic development publicly available RNA-seq raw reads have been retrieved from the Danio Code data coordination center website (https://danio-code.zfin.org) and generated by White and colleagues (White et al., 2017). The dataset is composed by 18 different stages, 5 replicates each for a total of 90 samples (**Table-4.1**).

Sample name	Numb. of replicates	Developmental stage	Developmental period
S01	5	1-cell	Zygote
S02	5	2-cell	Cleavage
S03	5	128-cell	Blastula
S04	5	1k-cell	Blastula
S05	5	Dome	Blastula
S06	5	50% epiboly	Gastrula
S07	5	Shield	Gastrula
S08	5	75% epiboly	Gastrula
S09	5	1-4-somites	Segmentation
S10	5	14-19-somites	Segmentation
S11	5	20-25-somites	Segmentation
S12	5	Prim-5	Pharyngula
S13	5	Prim-15	Pharyngula
S14	5	Prim-25	Pharyngula
S15	5	Long-pec	Hatching
S16	5	Protruding mouth	Larval
S17	5	Day-4	Larval
S18	5	Day-5	Larval

Table-4.1: zebrafish developmental time course RNA-seq dataset.
The RNA-seq dataset from White and colleagues (White et al., 2017) has been retrieved from Danio Code website. It is composed by 90 samples totally representing 18 different developmental time points.

After having retrieved the raw RNA-seq raw reads the quality of the reads have been assessed by using FastQC (Andrews, 2010). Having detected the absence of the sequencing adapters and an overall good quality of the reads, no read trimming has been performed.

RNA-seq dataset analysis – gene expression

To quantify the gene expression values of the coding and non-coding genes annotated in the zebrafish genome, the raw RNA-seq reads have been mapped on the zebrafish genome (GRCz11 version – Ensembl 96 release) using STAR (v2.6.0c) (Dobin et al., 2013). Default parameters have been used except for the number of multimapping reads that have been set to 80 (--outFilterMultimapNmax 80). The expression of both coding and non-coding genes annotated in the genome has been then quantified using htseq-count (v0.11.2, parameters: -s reverse -m union --nonunique all) (Anders et al., 2015). Next, to identify the differentially expressed genes (DE) edgeR has been used (Robinson et al., 2010). EdgeR normalisation of raw read counts has been applied using the TMM method whereas the common, trended and tagwise dispersions have been estimated by maximizing the negative binomial likelihood (default). Next, differentially expressed genes have been identified for each pairwise comparison performing a quasi-likelihood F-tests (glmQLFit and glmQLTest). Genes have been selected as differentially expressed when showing FDR < 0.05 and log2FC < -1 or > 1 (2-fold in linear scale).

RNA-seq dataset analysis – TE expression

TE locus specific expression levels have been calculated using SQuIRE (Yang et al., 2019). First, the reference genome and the annotation datasets referring to the zebrafish danRer11 genome version have been downloaded and prepared for the subsequent analyses using the SQuIRE Fetch and Clean modules, then the trimmed reads have been mapped on the reference genome using the Map module and finally read counts have been estimated using the Count module (strandedness='1'). Elements annotated as DNA, LINE, SINE, LTR and RC have been selected and differentially expressed TE loci have been identified using edgeR as previously described. TE loci showing FDR < 0.05 and log2FC <-1 or >1 have been considered as differentially expressed.

In order to summarize the expression levels of specific TE consensus the TEspeX pipeline previously described has been used. Briefly, a reference transcriptome is built merging the RepBase TE sequences (Bao et al., 2015) and the Ensembl transcript sequences containing all the coding and non-coding annotated transcripts (Zerbino et al., 2018). Reads are then mapped on the reference transcriptome using STAR (v2.6.0c) (Dobin et al., 2013) and assigning primary alignment flag to all the alignments with the best score. All alignments flagged as primary (–F 0 × 100 parameter) are then selected using samtools (v1.3.1) (H. Li et al., 2009). To avoid counting reads mapping on TE fragments embedded in coding and/or long non-coding transcripts, reads mapping with best-scoring alignments on any Ensembl transcript are discarded using Python scripts and Picard FilterSamReads (v2.18.4) (http://broadinstitute.github.io/Picard). Selected reads mapping exclusively on TEs and in the proper orientation are finally counted in each sample. Differentially expressed TE consensus sequences have been performed using edgeR as previously described except for the library size of each sample that has been calculated providing the total number of reads mapped on the transcriptome (coding, non-coding and TE consensus sequences) instead of using the default values.

Gene ontology (GO) enrichment analysis

GO enrichment analysis has been performed by using topGO (Alexa & Rahnenfuhrer, 2019). GO enrichment analysis has been conducted on the GO terms associated to the 1k-*versus* 128-cell stage upregulated genes, using as background the GO terms associated to the whole set of coding and non-coding annotated genes. First, the statistical significance of the enrichments has been tested with the Fisher's Exact Test (algorithm='weight'). Then, GO terms associated to less than 15 significant genes have been discarded prior to FDR calculation (Benjamini & Hochberg). Significant threshold has been imposed to FDR < 0.05.

Upregulated genes chromosomal and structural organisation

To define whether the 1k- *versus* 128-cell stage upregulated genes are significantly enriched in specific chromosomes, compared to the rest of the transcriptome, the number of upregulated genes on each zebrafish chromosome (excluding scaffolds and mitochondrial chromosome) have been counted. The same analysis has been repeated selecting an equal number of randomly selected genes 100 times. Z-score has been

consequently calculated subtracting to the number of upregulated genes on each chromosome the mean number of random genes on the same chromosome and dividing by the standard deviation of the number of random genes on that chromosome (z-score describes the position of a raw score in terms of its distance from the mean, when measured in standard deviation units). P-value has next been calculated from z-score and corrected using the FDR Benjamini & Hochberg correction. FDR significant threshold has been set to 0.05.

To define whether the 1k-/128-cell stage upregulated genes are significantly enriched in genic clusters compared to the rest of the transcriptome, first genic clusters composed by at least 5 genes have been identified using ClusterScan (parameters: -n 5 -d 410000 –singletons) (Volpe et al., 2018). Next, the number of upregulated genes belonging to one of the previously defined clusters has been calculated. The genes have been considered as part of specific gene clusters when overlapping with at least the 50% of their length the cluster. The same analysis has been performed on an equal number of random genes as previously described therefore obtaining z-score, P-values and FDR-corrected P-values. The same workflow has been used to calculate the enrichment of upregulated TEs in genic cluster. The upregulated gene cluster enrichment analysis has been additionally performed cluster by cluster. Namely, the number of upregulated genes in each different cluster has been calculated and compared with the one of an equal number of randomly selected genes next calculating z-score, P-values and FDR-corrected P-values as previously described.

CAGE-seq data collection

To map at the nucleotide resolution the TSS of each gene expressed upon the zebrafish ZGA CAGE-seq raw reads have been retrieved from the Danio Code data coordination center website (https://danio-code.zfin.org). Overall, the complete dataset is composed by 12 samples comprising: unfertilised egg, fertilised egg, 64-cell, 512-cell, High, Oblong, Sphere, 30% epiboly, Shield, 14-19-somites, Prim-5 and Prim-25 stages (**Table-4.2**).

CAGE-seq	RNA-seq	Developmental period
unfertilised egg	-	Egg
fertilised egg	-	Egg
-	1-cell	Zygote
-	2-cell	Cleavage
64-cell	-	Cleavage
-	128-cell	Blastula
512-cell	-	Blastula
-	1k-cell	Blastula
High	-	Blastula
Oblong	-	Blastula
-	Dome	Blastula
30% epiboly	-	Blastula
-	50% epiboly	Gastrula
Shield	Shield	Gastrula
-	75% epiboly	Gastrula
-	1-4-somites	Segmentation
14-19-somites	14-19-somites	Segmentation
-	20-25-somites	Segmentation
Prim-5	Prim-5	Pharyngula
-	Prim-15	Pharyngula
-	Prim-25	Pharyngula
-	Long-pec	Hatching
-	Protruding mouth	Larval
-	Day-4	Larval
-	Day-5	Larval

Table-4.2: zebrafish embryo CAGE-seq dataset.
The CAGE-seq dataset, retrieved from the Danio Code website, is reported in the 1st column of the table. It is composed by 12 samples comprising the time points from unfertilised eggs to Prim-5 stage. For convenience the RNA-seq dataset previously described in Table-4.1 is reported beside in the 2nd column of the table.

Considering the discrepancies between the RNA-seq and CAGE-seq datasets, and considering that the genes whose TSS has to be identified are expressed in the early blastula stages (128- and 1k-cell stages) the TSS for both the 1k-/128-cell and the 128-/2-cell stages have been identified using the earliest blastula stage of the CAGE-seq data, the 512-cell stage.

Gene TSS identification from CAGE-seq data

To identify the gene TSS starting from raw CAGE-seq data, the sequencing reads of the 512-cell stage have been mapped on the zebrafish reference genome (GRCz11 version – Ensembl 96 release) using bowtie1 (v. 1.2.3) (Langmead et al., 2009), allowing up to 80 multi-mapping reads and selecting all the alignments belonging to the 'best' strata (parameters: -a -m 80 --best --strata). Next, the gene TSS have been identified using the CAGEr Bioconductor package as follow (Haberle et al., 2015). First, starting from the alignment bam files generated by bowtie1, the CAGE transcriptional start sites (CTSS) have been identified using the CAGEr getCTSS function. Next the co-stranded CTSS closer than 20 nt and supported by at least 0.5 normalized read counts (expressed in tpm – tag per million) have been clustered together in the so-called tag cluster (clusterCTSS function). Then the CTSS not clustering in any tag cluster (the so-called singletons) have been removed from the analysis. Finally, the CTSS with the highest expression value (the so-called dominant TSS) has been considered as the tag cluster representative CTSS in the downstream analyses. To annotate the dominant TSS identified by CAGEr to the nearest gene, to each gene transcript isoform it has been associated the dominant TSS closest to its Ensembl-annotated TSS, but not further than 1kb. Then, for each gene the dominant TSS showing the highest expression among the different isoforms has been selected. In case of annotated genes with no dominant TSS associated (e.g. genes not expressed in this developmental stage) the Ensembl-annotated TSS has been considered.

Upregulated genes/annotated TEs overlap analysis

To define the TE occupancy nearby the TSS of the 1k-/128-cell and 128-/2-cell upregulated genes, the number of upregulated gene TSS overlapping annotated TEs has been calculated. As previously described, only TEs belonging to DNA, RC, LTR, LINE and SINE classes have been considered in the analysis. The upregulated genes/annotated TEs overlap analysis has been performed identifying the gene TSS as previously described

and elongating it by +/- 100 nt. The same analysis has been repeated on randomly selected genes following the previously described methods to define a statistically significant enrichment or depletion. This analysis has been performed on the total number of 1k-/128-cell and 128-/2-cell upregulated genes as well as on the 1k-/128-cell upregulated genes subdivided in the different genic clusters they result enriched in. In this second analysis, the number of randomisations has been increased to 10,000 given the high variability deriving from the low number of genes composing each of the analysed clusters.

Statistical analysis

All the statistical analyses performed externally to previously reported software (edgeR, topGO) have been conducted either in R (v3.6.2) (R Core Team, 2018) or in python (v3.7.6) (Rossum & Drake, 2001) taking advantage of the numpy (Harris et al., 2020) and scipy (SciPy 1.0 Contributors et al., 2020) libraries. All the plots have been generated in R, using the ggplot2 library (Wickham, 2016).

Chapter 5

Transposable elements and genic clusters influence the dynamics underlying the zygotic genome activation

5.1 Introduction

During the initial phases of the Metazoan embryonic development, the zygotic genome remains transcriptional silent (Eckersley-Maslin et al., 2018; Schulz & Harrison, 2019). In absence of transcription, the development is driven by maternally provided transcripts and proteins. Nevertheless, for the embryo to continue developing, the transcriptional control must be shifted from the maternal transcripts to the zygotic genome (Schulz & Harrison, 2019). Toward this end, maternal products are degraded and the zygotic genome activated (Tadros & Lipshitz, 2009). Zygotic genome activation (ZGA) is characterised by two remarkable transcriptional waves, a minor and a major one, by which the embryo is gradually taken from a quiescent transcriptional state to a state where thousands of genes are actively expressed (Jukam et al., 2017; Lee et al., 2014). Additionally, besides the transcriptional activation, Metazoans ZGA is also characterised by multiple levels of chromatin reorganisation such as DNA methylation, histone modifications and changes in chromatin accessibility and nucleosome positioning (Eckersley-Maslin et al., 2018; Schulz & Harrison, 2019). Nevertheless, it is unclear whether the changes in chromatin are required for the ZGA or whether the ZGA is instructive to the chromatin changes (Eckersley-Maslin et al., 2018).

Although being remarkably conserved among Metazoans, ZGA occurs with timings and is coordinated by activators that appear to be species-specific. In *Mus musculus* (mouse) the ZGA minor wave occurs as early as in the early 2-cell stage whereas the major one arises in the 2-cell stage (Jukam et al., 2017; Tadros & Lipshitz, 2009). The transcriptional program activated upon murine ZGA appears to be driven by the zygotically transcribed transcription factor *Dux* (encoded by the *Duxf3* gene) (De Iaco et al., 2017; Hendrickson et al., 2017). *Dux*, once activated, coordinates a well-established transcriptional

programme that leads to the transcription of the *Zscan4*, *Prame*, and *Eif1a*-like gene families (De Iaco et al., 2017; Hendrickson et al., 2017). Importantly, the transcriptional programme activated by *Dux* is not exclusively restricted to protein-coding genes as the transcription of the retrotranspositionally inactive endogenous retrovirus (ERV) MERVL is activated in this stage (Hendrickson et al., 2017). Whether MERVL transcription is directly activated by *Dux* (Hendrickson et al., 2017) or it is activated by the *Dux* target Zscan4c (W. Zhang et al., 2019) is still uncertain with observations supporting both hypotheses. Nevertheless, given the importance of the transcriptional activation of the MERVL retrotransposons in these stages, it may be reasonable to postulate that both transcription factors, independently from each other, might contribute to the transcriptional activation of the MERVL elements. For sure, MERVL-derived sequences have been described to provide an alternative 5' exon to many genes expressed at the ZGA onset (Peaston et al., 2004). As consequence of this phenomenon, MERVL-gene chimeric transcripts are formed with MERVL promoter regulating the transcription of the chimera (Macfarlan et al., 2012; Peaston et al., 2004). From an evolutionary perspective, this suggests that MERVL amplification within the host genome may have evolved to facilitate ZGA providing one single element regulating many genes in order to coordinate their quick and stage-specific transcriptional activation (Torres-Padilla, 2020).

Nevertheless, MERVL elements are not the only transposable elements (TE) expressed in the mouse early embryo. Recent evidence has indeed shown how the LINE L1 transcription plays a key dual role in this context (Jachowicz et al., 2017; Percharde et al., 2018). On one hand, LINE L1 derived transcripts have been described to act as chromatin remodellers modulating the chromatin accessibility in the mouse 2-cell stage embryos (Jachowicz et al., 2017). On the other, LINE L1 mRNA has been proposed to be required, together with Nucleolin and Trim28 proteins, for the repression of the *Dux* activated 2-cell-specific transcriptional programme thus promoting the progression of the embryonic development beyond this stage (Percharde et al., 2018).

Finally, it is important to notice how the transcriptional activation characterising the murine ZGA is not the only molecular change occurring in this biological context. Indeed, the murine ZGA is also characterised by multiple levels of chromatin reorganisation (Eckersley-Maslin et al., 2018; Schulz & Harrison, 2019). In this context, at the ZGA minor

wave onset (early 2-cell), the embryo shows a unique chromatin landscape that features weak and noisy ATAC-seq profile, with large domains of accessible chromatin covering the expressed genes and especially the MERVL elements (J. Wu et al., 2016). On the contrary, following the ZGA minor wave, sharp peaks of accessible chromatin are observable at the gene transcription start site (TSS) (J. Wu et al., 2016). Intriguingly, although its biological relevance is still unclear, in this stage peaks of accessible chromatin are also observable at the transcription termination sites (TTS) (J. Wu et al., 2016).

Although most of the players involved in the murine ZGA are known, it is still unclear whether the genomic loci of such activated genes share common structural genomic organisation. Additionally, although the transcription of TEs, and in particular of MERVL, has been observed in this context it is still unclear whether such elements actively contribute to the activation of the zygotic genome or whether their sequences have been passively co-opted by the zygotic genome as still uncertain is the role LINE L1 elements play in this context. Toward this end, taking advantage of mouse early embryo RNA-seq and ATAC-seq publicly available dataset (J. Wu et al., 2016), the transcriptional and epigenetic dynamics characterising the murine ZGA minor wave and the genomic structural organisation of the activated loci have been characterised. These results, highlighted how, both from a transcriptional and epigenetic point of view, upon ZGA minor wave the zygotic genome is characterised by stage-specific features not reproducible in earlier or later stages. By investigating the structural organisation of the involved genomic loci, I was able to highlight that many of these genomic features are organised in genic clusters and that TE sequences might have a role in both the transcriptional activation and the chromatin opening of such loci.

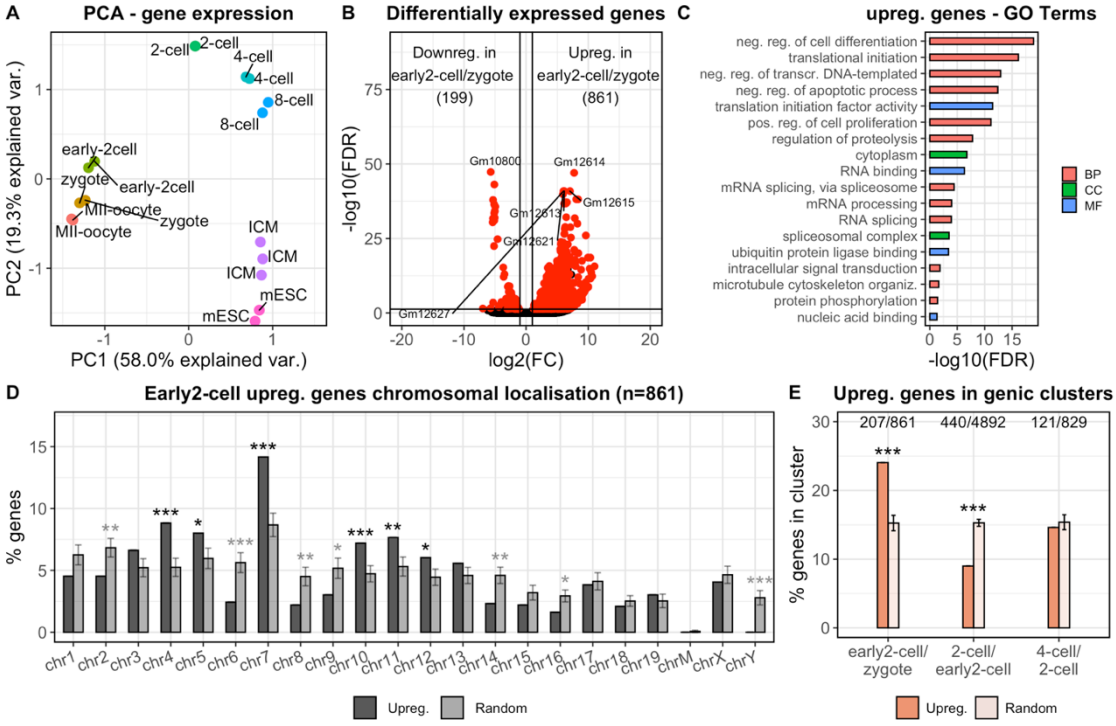
5.2 Results

Coding and non-coding genes get transcriptionally activated upon ZGA minor wave

To investigate the transcriptional dynamics characterising the mouse early embryo, the expression levels of all the annotated coding and non-coding genes have been quantified from 8 murine embryonic stages (from MII-oocyte to blastocyst cells). To identify the stages characterised by similar expression profiles a principal component analysis (PCA) has been performed on the calculated expression levels. According to the PCA first two principal components, the 8 analysed samples resulted clustered in 3 main groups (**Figure-5.1A**). The first one is composed by pre-ZGA (MII-oocyte and zygote) and ZGA minor wave samples (early 2-cell). The second one by ZGA major wave samples (2-, 4- and 8-cell stage) whereas the third one is composed by post-ZGA blastocyst cells (inner cell mass [ICM] and murine embryonic stem cells [mESC]). Although having been clustered together in the first group, a small grade of separation is appreciable between the MII-oocyte/zygote and the early 2-cell samples supporting the evidence of the ZGA minor wave occurring between the zygote and the early 2-cell stages (**Figure-5.1A**). To identify the genes whose expression is altered upon ZGA minor wave onset, the differentially expressed (DE) genes between early 2-cell and the zygote stages have been identified. Almost one thousand genes resulted differentially expressed (FDR<0.05 and log2FC >1 or <-1) with approximately 40% of them being represented by non-coding genes (**Figure-5.1B**). Notably, more than 80% of the DE genes (861 genes) resulted upregulated rather than downregulated in early 2-cell compared to the zygote stage (**Figure-5.1B**). Given that the murine ZGA minor wave begins at the early 2-cell stage, such 861 upregulated genes are likely to belong to the set of early genes transcribed upon the activation of the zygotic genome. Consistent with this observation, several of the well-known *Dux* target genes were observed among the upregulated genes defined in this analysis (e.g. *Zscan4*, *Prame*, and *Eif1a*-like gene families) (De Iaco et al., 2017; Hendrickson et al., 2017). Gene ontology (GO) enrichment analysis showed how, applying the parameters and significance thresholds described in the Method section, 15 GO terms resulted significantly enriched in the set of terms associated to the upregulated genes with respect to the rest of the transcriptome (**Figure-5.1C**). The enriched terms resulted related to biological pathways such as transcription, translation, RNA and DNA binding, proteolysis and cell proliferation (**Figure-5.1C**). All of these processes are in line with the

analysed biological context characterised by an increased transcriptional and translational activity (*i.e.*, RNA binding, mRNA splicing, nucleic acid binding, translation initiation), by the degradation of the maternally deposited transcripts and proteins (*i.e.*, regulation of proteolysis, ubiquitin protein ligase binding) and by an increased cell proliferation (*i.e.*, positive regulation of cell proliferation) (**Figure-5.1C**). To understand whether the 861 early transcribed genes are homogenously distributed within the genome or belong to specific genomic loci, the chromosomal localisation of the early 2-cell upregulated genes has been assessed. Intriguingly, upregulated genes resulted enriched in specific chromosomes (4, 5, 7, 10, 11 and 12) and depleted from others (2, 6, 8, 9, 14) compared to the rest of the transcriptome (z-score FDR<0.05) (**Figure-5.1D**). This result suggests that the zygotic genome is not homogenously activated but rather the transcription starts from specific hotspots. To define whether this is a consequence of specific genomic structural organisations, the reciprocal distribution of the 861 upregulated genes has been investigated with regards to their overlap with genic clusters. First, genic clusters have been defined as groups of at least 5 consecutive genes associated to the same functional domains. Next, the number of upregulated genes located within such genic clusters have been calculated. Early 2-cell *versus* zygote upregulated genes resulted significantly enriched in genic clusters with respect to the rest of the transcriptome (z-score FDR=5.8E-15) with more than 20% of the upregulated genes (~200 genes) resulting located within a cluster (**Figure-5.1E**). Intriguingly, neither the 2- *versus* early 2-cell nor the 4- *versus* 2-cell upregulated genes displayed such pattern thus suggesting the stage specificity of this result (**Figure-5.1E**). Finally, the specific clusters the early 2-cell/zygote upregulated genes are enriched in have been identified. The upregulated genes resulted significantly enriched in 13 different clusters with the majority of them being composed of genes with functions related with transcription regulation whereas the remaining sets were associated with functions like signal transduction, proteolysis, translation, immune and RNA binding (**Figure-5.1F**).

Together, these data displayed how the ZGA transcriptionally activated genes appear to be not randomly distributed along the genome being instead enriched in genic clusters.



F

Genic clusters early2-cell upreg. genes are enriched in

	Pfam	Upreg.	Random	FDR	Description
1	PF00651	20	0.48	3.04e-211	BTP-POZ domain
2	PF02847	7	0.11	1.46e-105	MA3 domain
3	PF00076	20	0.82	7.44e-85	RNA recognition motif.
4	PF01176	21	1.11	2.20e-80	Translation initiation factor 1A/IF-1
5	PF07270	10	0.42	6.67e-53	Protein of unknown function
6	PF00143	8	0.92	2.81e-14	Interferon alpha/beta domain
7	PF05296	15	2.60	5.23e-14	Taste receptor protein (TAS2R)
8	PF02023	10	1.52	7.23e-13	SCAN domain
9	PF00046	11	2.80	1.48e-07	Homeobox domain
10	PF00069	12	3.19	8.52e-07	Protein kinase domain
11	PF03402	17	7.78	9.36e-04	Vomeronasal organ pheromone recept. fam.
12	PF00096	14	5.76	2.16e-03	Zinc finger, C2H2 type
13	PF13465	28	19.36	4.80e-02	Zinc-finger double domain

Figure-5.1: coding and non-coding genes get transcriptionally activated upon ZGA minor wave. (**A**) PCA performed on the coding/non-coding gene expression values. PCA separates the analysed samples in 3 major groups. The pre-ZGA/ZGA minor wave one (MII-oocyte, zygote and early 2-cell), the ZGA major wave/post-ZGA one (2-, 4-, 8-cel; stages) and the post-ZGA one (ICM and mESC). Although having been clustered together, a small grade of separation is appreciable between the pre-ZGA (MII-oocyte and zygote) and ZGA minor wave onset samples (early 2-cell) (**B**) Differentially expressed genes between early 2-cell and zygote cell stages. More than 1,000 genes result DE with more than 80% of them resulting upregulated rather than downregulated. (**C**) GO term enrichment analysis. Results highlighted 15 GO terms enriched in the terms associated to the early 2-cell upregulated genes. The enriched GO terms are associated to biological pathways such as transcription, translation, RNA and DNA binding and proteolysis. (**D**) Early 2-cell/zygote upregulated genes chromosomal localisation. Upregulated genes are significantly enriched in specific chromosomes while being depleted from others with respect to the rest of the transcriptome. (**E**) Number of upregulated genes organised in genic cluster. Early 2-cell/zygote upregulated genes are enriched in genic clusters defined by at least 5 consecutive genes associated to the same functional domain whereas this feature is not displayed neither by the 2-/early 2-cell nor by the 4-/2-cell upregulated genes. (**F**) Genic clusters the early 2-cell upregulated genes are enriched in. The majority of such clusters is composed of genes with transcriptional regulatory functions whereas the remaining sets are associated with signal transduction, proteolysis, translation, immune and RNA binding functions. (D and E: mean ± standard deviation is represented by the bars. *FDR<0.05, **FDR<0.01, ***FDR<0.001. In D, E and F FDR values refer to z-score derived BH FDR-corrected P-value. 100 randomisations).

Transposable elements are transcriptionally activated upon ZGA minor wave

To deeply characterise the transposable element (TE) expression in the mouse early embryo and especially upon ZGA minor wave onset, the expression of every TE fragment (from now-on called TE locus) annotated in the murine genome has been quantified in each of the analysed samples. To address whether the 8 analysed samples are characterised by different TE transcriptional profiles, a PCA has been performed on the TE locus expression levels. As already observed for the coding/non-coding genes (**Figure-5.1A**), the 8 samples resulted characterised in three major groups (**Figure-5.2A**). As before, the first group is composed by pre-ZGA/ZGA minor wave samples with the early 2-cell samples forming a subgroup slightly separated from the pre-ZGA samples and thus supporting the evidence of ZGA minor wave occurring between these stages. The second group is instead composed by ZGA major wave samples (2-, 4- and 8-cell) and the third one by post-ZGA samples (ICM and mESC) (**Figure-5.2A**). Having confirmed the different TE transcriptional profile characterising the pre-ZGA and ZGA minor wave samples, a TE locus differentially expressed analysis between early 2-cell and zygote stages has been performed. On total, more than 2,600 TE loci resulted differentially expressed between the early 2-cell and the zygote samples (FDR<0.05 and log2FC >1 or <-1) (**Figure-5.2B**). Intriguingly, more than 96% of such DE loci (2,589) resulted upregulated in the early 2-cell sample thus suggesting a strong activation of the TE transcription upon murine ZGA minor wave (**Figure-5.2B**). Importantly, more than 50% of such upregulated TE loci (1,386 out of 2,589 loci) resulted annotated as ERVL whereas they represent less than 5% of the total annotated TEs in the murine genome being in line with previous observations (**Figure-5.2C**) (De Iaco et al., 2017; Hendrickson et al., 2017; Peaston et al., 2004). To confirm the stage-specificity of the observed ERVL transcriptional activation in the early 2-cell stage, the upregulated TE loci between 2- and early 2-cell as well as between 4- and 2-cell stages have been identified and classified according to their TE family. The fraction of ERVL loci resulting upregulated at these stages decreases with the proceeding of the development. While more than 50% of the total TE upregulated loci are annotated as ERVL in early 2-cell *versus* zygote stages, less than 20% and less than 1% of the upregulated TE resulted annotated as ERVL elements in 2- *versus* early 2-cell and in 4- *versus* 2-cell stages, respectively (**Figure-5.2C**). Proceeding with the development, more than four thousand TE loci, heterogeneously distributed among the different TE families, resulted upregulated upon the ZGA major

wave (between the early 2- and the 2-cell stages), whereas a small portion of TEs resulted upregulated between the 4- and the 2-cell stages (n=411) with a prevalence upregulation of ERVK elements (**Figure-5.2C**).

To assess whether the early 2-cell upregulated genes are homogenously distributed along the genome or they reside on loci characterised by specific genomic structural organisation, the number of the early 2-cell/zygote upregulated TE loci located within genic clusters has been calculated. Intriguingly, as already displayed by the coding/non-coding genes, early 2-cell/zygote upregulated TE loci resulted significantly enriched in genic clusters with respect to the rest of the transcriptome (z-score FDR=4.1E-34) with more than 300 TE loci (12% of the total upregulated loci) resulting located within a cluster (**Figure-5.2D**). However, this pattern does not appear to be stage-specific as 2-/early 2-cell and 4-/2-cell upregulated TE loci resulted enriched in clusters as well (z-score FDR=1.9E-55 and FDR=2.8E-42, respectively) (**Figure-5.2D**).

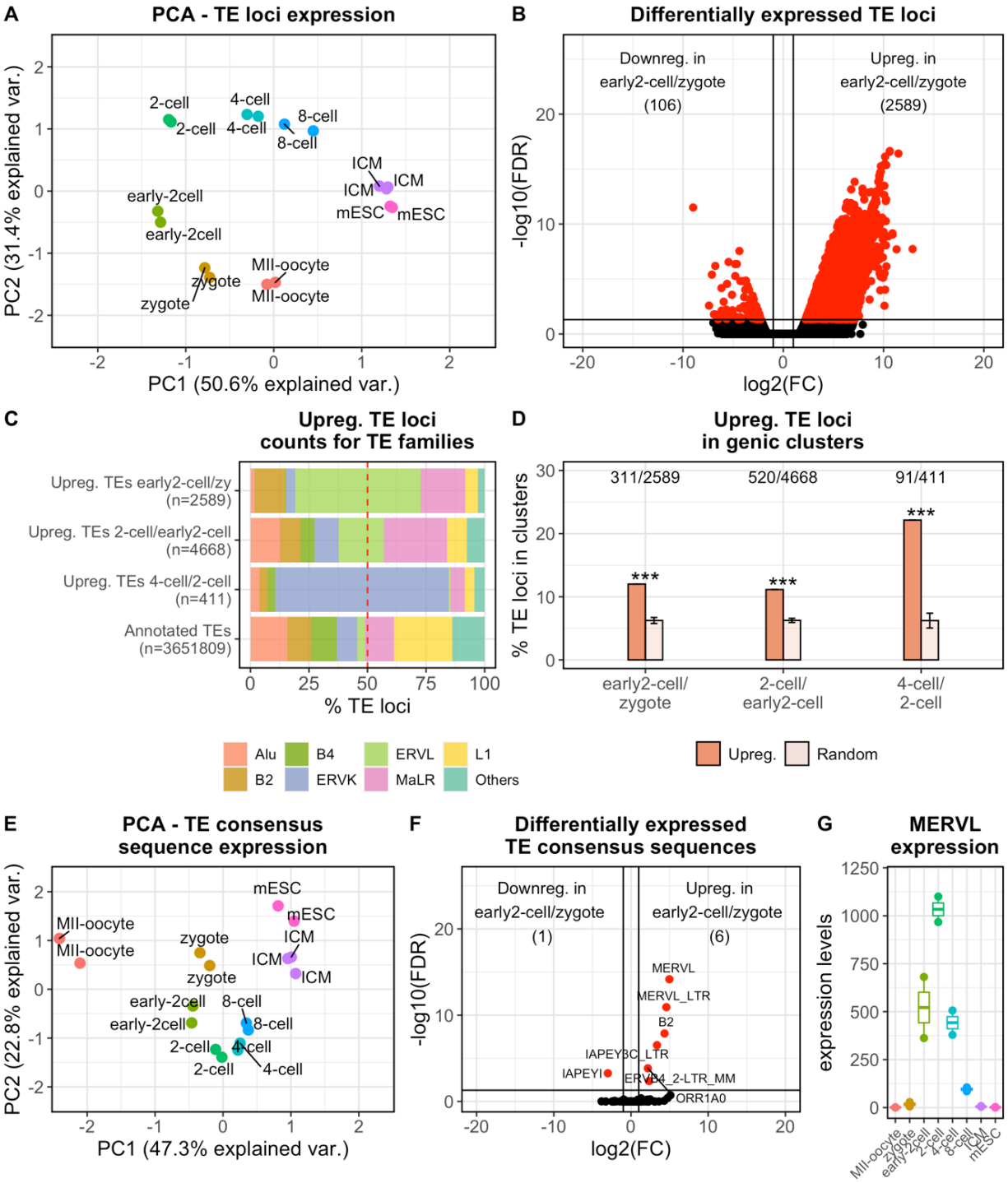
Together, these results displayed how during ZGA minor wave the transcription of thousands TE loci, and in particular of ERVL, is activated. Importantly, these TE loci resulted enriched in genic clusters compared to the rest of the transcriptome, even though this feature does not appear to be stage specific.

MERVL are specifically transcribed upon the murine ZGA minor wave

The ERVL transcriptional activation upon murine ZGA minor wave previously described is consistent with earlier observations (De Iaco et al., 2017; Hendrickson et al., 2017; Peaston et al., 2004). However, given the tendency of ERVL in generating chimeric transcripts with early expressed genes by providing alternative promoter/exon (Macfarlan et al., 2012; Peaston et al., 2004), it is not clear whether ERVL are transcribed as exonised TE fragments part of coding/non-coding transcripts or as independent transcripts. Additionally, considering that the tool previously used to quantify the TE locus expression does not apply any correction on the potential transcription of TE fragments as part of coding/non-coding transcripts, the observed results may be a reflection of the transcription of TE fragments embedded in coding/non-coding genes rather than representing specific TE expression. Therefore, the TE expression levels have been additionally calculated at the TE consensus sequence level (now-on TE consensus) using the TEspeX tool, capable to exclude all the RNA-seq reads mapping ambiguously on both TEs and coding/non-coding transcripts (described in the Chapter 2 of this thesis). Overall, the PCA analysis recapitulated the one observed for coding/non-coding genes and TE loci even though some differences are observable with MII-oocyte sample clustering separately from all the other samples and zygote and early 2-cell samples resulting grouped nearby the ZGA major wave samples (2-, 4-, 8-cell) (**Figure-5.2E**). Nevertheless, such differences are likely to be due to the lower variance explained by the first two components in this analysis (~70% variance explained) with respect to the previous ones (~77% for coding/non-coding genes and 82% for TE loci) (**Figure-5.2E**). Next, the differentially expressed TE consensus between early 2-cell and zygote stages has been identified. Out of the 301 analysed, 7 TE consensus resulted differentially expressed between the analysed stages (**Figure-5.2F**). In line with the TE loci results, the majority (6 out of 7) of the DE TE consensus resulted upregulated rather than downregulated upon murine ZGA minor wave (early 2-cell *versus* zygote) (**Figure-5.2F**). Importantly, the 2 most significantly upregulated TE consensus sequences resulted two ERVL subfamilies, MERVL and its corresponding solitary LTR portion MERVL_LTR thus suggesting MERVL are transcribed in a specific manner rather than as exonised fragments part of coding/non-coding transcripts (**Figure-5.2F**). Finally, the MERVL expression profile excluding from the quantification all the RNA-seq reads possibly deriving from the transcription of coding/non-coding genes has been investigated

(**Figure-5.2G**). The results showed how the MERVL transcription appears to be specifically confined in the early 2- and 2-cell stages. MERVL-derived mRNA is almost undetectable in MII-oocyte and zygote stages, then, upon ZGA minor wave (early 2-cell) MERVL expression increases reaching the peak in the 2-cell stage (ZGA major wave). MERVL-derived mRNA levels then gradually decrease in the following stages until the undetectability in blastocyst cells (ICM and mESC) (**Figure-5.2G**).

Altogether, these data showed that the MERVL transcription is specifically activated upon murine ZGA minor wave. Thus, in this context, MERVL are transcribed in a specific manner rather than as exonised fragments part of coding/non-coding transcripts.



Legend next page

Figure-5.2: transposable elements are transcriptionally activated upon murine ZGA minor wave. (**A**) PCA performed on the TE locus expression. PCA shows the separation between the samples mainly according to their pre-ZGA/ZGA minor wave (MII-oocyte, zygote and early 2-cell), ZGA major wave (2-, 4- and 8-cell stage) and post-ZGA samples (ICM and mESC). Nevertheless, supporting the evidence of the ZGA minor wave occurring between early 2-cell and zygote stages, within the pre-ZGA/ZGA minor wave group, the early 2-cell samples form a specific subgroup clustering differently from the pre-ZGA samples (MII-oocyte and zygote). (**B**) Differentially expressed TE loci between early 2-cell and zygote stages. More than 2,600 TE loci result differentially expressed between early 2-cell and zygote cell. The large majority (>96%) of the TE loci show an upregulation, rather than a downregulation, upon murine ZGA minor wave (early 2-cell). (**C**) Number of the upregulated TE loci counted for TE families. More than half (1,386) of the upregulated TE loci in early 2-cell/zygote stages are annotated as ERVL. The fraction of upregulated ERVL loci decreases in 2-/early 2-cell and in 4-/2-cell stages with <20% and <1% of the total upregulated TE loci being annotated as ERVL, respectively. (**D**) Number of upregulated TE loci organised in genic cluster. Early 2-cell/zygote upregulated TEs are enriched in genic clusters defined by at least 5 consecutive genes associated to the same functional domain. The same feature is displayed by both the 2-/early 2-cell and the 4-/2-cell upregulated TEs. (**E**) PCA performed on the TE consensus sequence expression. Data confirm the separation previously observed in (A) even though some differences are observable with MII-oocyte sample clustering separately from all the other samples and zygote and early 2-cell samples resulting grouped nearby the ZGA major wave samples (2-, 4-, 8-cell). These differences are likely a consequence of the smaller variance explained by this PCA compared to the previous ones (**F**) Differentially expressed TE consensus sequences. Of the 301 analysed, 7 TEs resulted differentially expressed with 6 out of 7 resulting upregulated in early 2-cell compared to zygote stages. The two most significantly upregulated TEs resulted the MERVL and its corresponding solitary LTR portion MERVL_LTR. (**G**) MERVL consensus sequence expression profile calculated by TEspeX and thus excluding RNA-seq reads mapping ambiguously on both TE sequences and coding/non-coding transcripts. MERVL expression is activated upon murine ZGA minor wave (early 2-cell) and it is specifically confined in early 2-cell and 2-cell stages. (D: mean ± standard deviation is represented by the bars. ***FDR<0.001. FDR values refer to z-score derived BH FDR-corrected P-value. 100 randomisations).

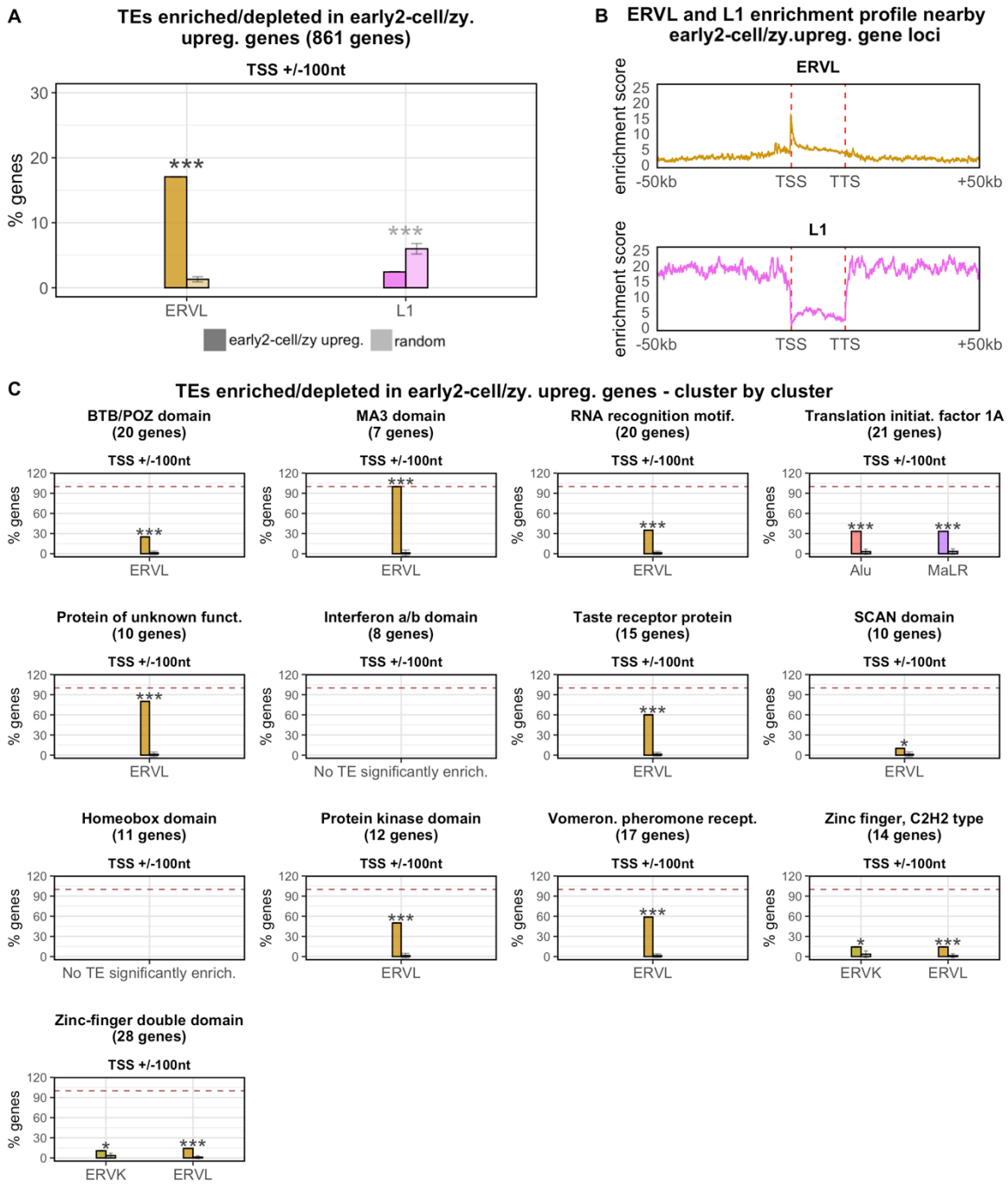
Early 2-cell/zygote upregulated genes are enriched in ERVL and depleted of LINE L1 sequences

Having defined the transcriptional dynamics characterising the murine ZGA minor wave, it was next investigated whether specific TE families, independently from the TE expression, are located in the proximity of the TSS of the early 2-cell/zygote upregulated genes. Toward this end, the fraction of upregulated genes overlapping with the TSS for each TE family annotated in the murine genome, has been calculated and compared with the one of randomly selected ones. Instead of considering the gene TSS as a single-nucleotide point, the genomic coordinates of the TSS have been elongated both upstream and downstream by 100 nt. The results showed how almost 20% of the early 2-cell upregulated gene TSS (147 genes) is significantly enriched in ERVL sequences with respect to the rest of the transcriptome (z-score FDR<1.8E-308) (**Figure-5.3A**). These data are overall in line with previous observations describing the role of the MERVL elements, a specific subfamily of ERVL, in regulating the transcription of nearby genes through the generation of MERVL-gene chimeric transcripts (Macfarlan et al., 2012; Peaston et al., 2004). Next, the result was furtherly confirmed by the visualisation of the ERVL enrichment profile nearby the genomic loci occupied by the early 2-cell/zygote upregulated genes that showed how the upregulated genes, besides being enriched in ERVL sequences at the TSS level, are also covered by ERVL sequences all along the gene body (**Figure-5.3B** top panel). The upregulated gene resulted enriched in no other TE families nearby the TSS. However, they resulted significantly depleted of LINE L1 sequences in proximity of their TSS (z-score FDR=6.0E-05) (**Figure-5.3A**). The visualisation of the LINE L1 enrichment profiles nearby the loci occupied by the upregulated genes furtherly confirmed this result showing how the upregulated genes resulted depleted of LINE L1 sequences especially at the TSS level but also all along the entire gene body (**Figure-5.3B** bottom panel). This result might be of particular interest especially when considering previous observations describing how LINE L1 transcripts act as chromatin remodellers in the mouse early embryo context (Jachowicz et al., 2017). Therefore, it may be possible that, in order to be successfully transcribed and to avoid a hypothetical LINE L1 mediated transcriptional repression, the early 2-cell/zygote upregulated genes must be depleted of LINE L1 sequences.

Next, to define whether the approximately 200 early 2-cell/zygote upregulated genes organised in genic clusters are similarly enriched and depleted of TE sequences, the same analysis has been repeated on the upregulated genes composing each of the 13 genic clusters previously identified (**Figure-5.1F**). The data showed that the TSS of the genes of 10 out of 13 clusters resulted significantly enriched in ERVL sequences when compared to the rest of the transcriptome thus confirming the results previously showed by the 861 upregulated genes (**Figure-5.3C**). Interestingly, 3 specific genic clusters displayed no ERVL enrichment at the TSS level: Translation initiation factor 1A, Interferon a/b domain and Homeobox domain. While the Interferon and Homeobox domain genes showed no TE significantly enriched nearby their TSS, approximately 30% of the Translation initiation factor 1A cluster genes displayed a significant enrichment nearby the TSS of Alu (SINE) and MaLR (LTR) sequences. Furtherly investigating this result, it become apparent that the Alu sequences were embedded in the first exon of the Translation initiation factor 1A cluster genes whereas the MaLR were located upstream to the gene TSS with no contribution to the gene transcription (**Figure-5.4**). It is thus possible to hypothesise that Alu sequences may have contributed to the evolution of the Translation initiation factor 1A cluster genes by providing splicing sites and/or alternative transcriptional start sites, as SINEs have the tendency to do (Lev-Maor et al., 2008; Schmitz & Brosius, 2011), whereas the role of MaLR in this context is still unclear. Considering all these data together, it might be proposed that ERVL-positive and ERVL-negative genes are characterised by different transcriptional timings with the ERVL acting as "transcriptional timer" activating the nearby genes in the same moment. However, whether the ERVL-negative genes are transcribed earlier or later than ERVL-positive remains to be defined.

Importantly, the analysis of the approximately 200 genes enriched in the 13 genic clusters showed that none of them is significantly depleted of LINE L1 sequences (**Figure-5.3C**). However, the lack of the confirmation of the previously observed depletion of LINE L1 sequences from the TSS of the early 2-cell/zygote upregulated genes might be the consequence of the small number of genes composing each cluster and thus considered in these cluster-by-cluster analyses.

Taken together, these results showed how the early 2-cell *versus* zygote upregulated gene genomic loci are significantly enriched in ERVL sequences and depleted of LINE L1 when compared to the rest of the transcriptome.



Legend next page

Figure-5.3: early 2-cell/zygote upregulated genes are enriched in ERVL and depleted of LINE L1 sequences.
(**A**) TE enrichments nearby the early 2-cell/zygote upregulated gene TSS. Upregulated gene TSS resulted significantly enriched in ERVL and depleted of LINE L1 with respect to the rest of the transcriptome. (**B**) Minigene showing the enrichment score of ERVL, and LINE L1 nearby the genomic loci occupied by the early 2-cell upregulated genes. As already shown in (A) the upregulated genes resulted covered by ERVL sequences especially nearby the TSS but also all along the gene body. On the contrary, upregulated genes resulted depleted of LINE L1 sequences nearby the TSS and all along the gene body. On y-axis the enrichment score is expressed as the % of upregulated genes overlapping annotated ERVL/LINE L1 sequences (**C**) TE enrichment nearby the early 2-cell/zygote upregulated gene TSS – cluster by cluster. For each of the 13 clusters the upregulated genes are enriched in the TE enrichment has been calculated. Ten out of 13 clusters showed an enrichment of ERVL nearby the TSS. On the contrary, Interferon a/b and homeobox domain cluster genes showed no TE significantly enriched nearby the TSS. Translation initiation factor 1A gene displayed enrichment nearby the gene TSS of the MaLR (LTR) and of the Alu (SINE) non-LTR. (A and C: mean ± standard deviation is represented by the bars. *FDR<0.05, **FDR<0.01 and ***FDR<0.001. FDR values refer to z-score derived BH FDR-corrected P-value. Only significant results for which at least 5% of the genes overlapped each TE family have been plotted. 100 randomisations (A), 10,000 randomisations (C)).

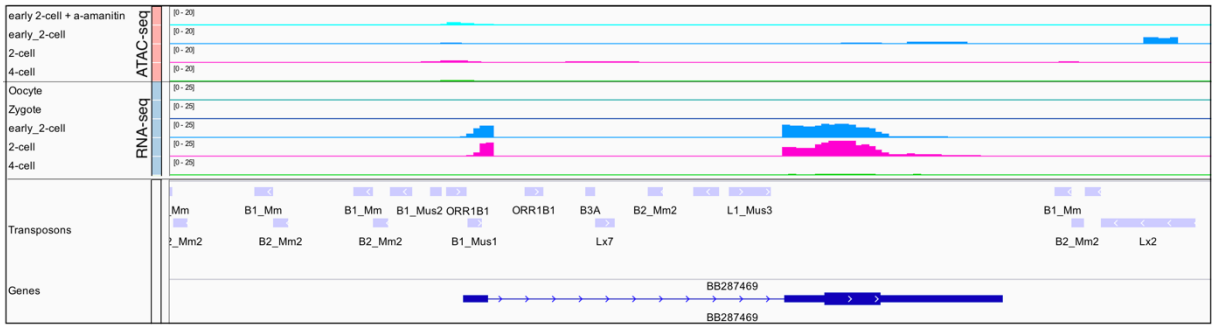


Figure-5.4: representative screenshot of the Translation initiation factor 1A cluster genes.
In the screenshot ATAC-seq tracks (red) and RNA-seq tracks (blue) are shown together with the TE UCSC and coding/non-coding gene tracks. The first exon of the BB287496 gene overlap a SINE B1 (Alu) element whereas the ORR1B1 (MaLR) element is located upstream to the B1. While it is likely that the SINE B1, by providing splicing sites, has contributed to the gene evolution the role of the MaLR ORR1B1 element remains unclear with the element not participating to the transcription of the gene. The ATAC-seq tracks show no signal in any stage whereas the RNA-seq tracks support the specific expression of the gene in early 2- and 2-cell stages.

A unique chromatin landscape characterises the mouse early embryo upon ZGA minor wave

To investigate how the accessible chromatin landscape changes during the mouse early embryo development and to define the genomic loci involved with this process, ATAC-seq raw reads have been retrieved from a publicly available dataset (J. Wu et al., 2016). The dataset is composed by one pre-ZGA sample represented by an early 2-cell embryo in which the transcription of RNA-pol II has been blocked starting from the PN3 zygote stage (early 2-cell + alpha-amanitin), one ZGA minor wave sample (early 2-cell), three ZGA major wave samples (2-, 4- and 8-cell) and two post ZGA samples (ICM, mESC).

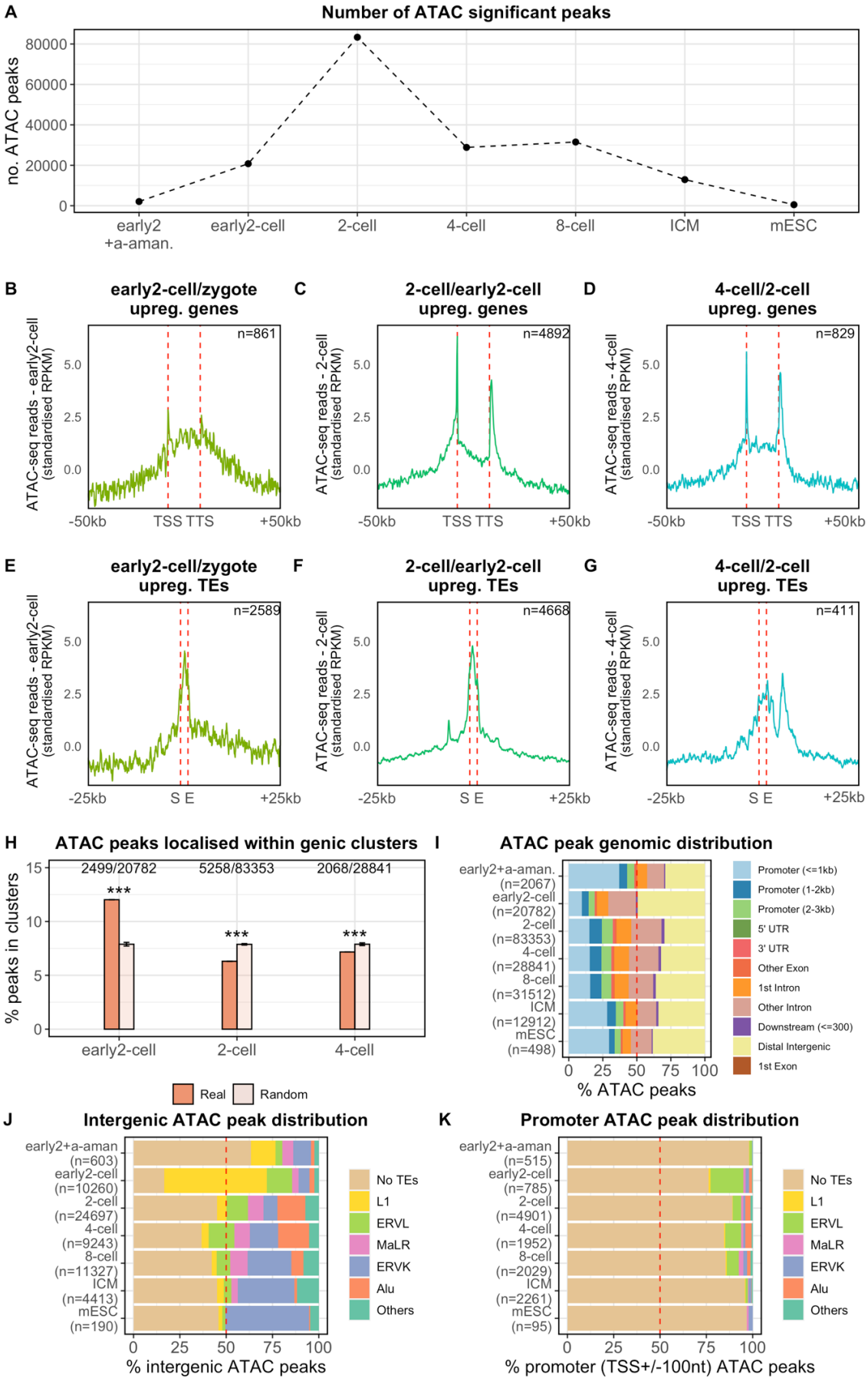
The ATAC peaks have been identified in each of the analysed samples using Genrich as described in the Methods section. Interestingly, the results showed how the number of significantly identified peaks increases by 10-fold upon ZGA minor wave (between the early 2-cell + alpha-amanitin and the early 2-cell samples) reaching the peak in correspondence of the onset of the ZGA major wave (2-cell stage) (**Figure-5.5A**). These results highlight that upon murine ZGA minor and major waves the embryonic genome undergoes a broad chromatin opening whereas proceeding with the embryo development the number of open chromatin loci gradually decreases supporting the evidence of a less permissive chromatin characterising the later stages of the development (Aoki et al., 1997; J. Wu et al., 2016) (**Figure-5.5A**). To test whether the transcription of the previously identified genes and TEs upregulated in early 2-/zygote, 2-/early 2- and 4-/2-cell stages is supported by open chromatin domains, the profile of the ATAC peaks nearby the genomic loci occupied by such genes and TEs has been visualised. The data showed the presence of sharply defined and highly supportive ATAC peaks nearby the upregulated gene TSS, in all the three analysed stages suggesting how both the ATAC- and RNA-seq datasets support the transcription of such genes (**Figure-5.5B, C and D**). As previously described by Wu and colleagues (J. Wu et al., 2016) the early 2-cell ATAC signal resulted particularly noisy and, most importantly, in all the three stages, open chromatin domains were observable also nearby the gene TTS (**Figure-5.5B, C and D**). The visualisation of the ATAC signal nearby the genomic loci occupied by the TEs resulting upregulated in the three analysed stages highlighted how, for all the stages, ATAC peaks cover the entire TE body (**Figure-5.5E, F and G**). This is likely to be the consequence of the short nature of the majority of the TEs populating the murine

genome with the TE body often located inside the ATAC peaks. Moreover, in all the three stages, ATAC peaks were observable also outside the TE body (**Figure-5.5E, F and G**). Although the interpretation of this result is still uncertain, this profile could be the consequence of the enrichment in clusters of the upregulated TEs (previously described in **Figure-5.2D**) with the signal outside the TE body arising from nearby TEs located within the same cluster. Next, starting from the observations of the genic cluster enrichment of the early 2-cell/zygote upregulated genes and TEs, the early 2-cell ATAC peaks enrichment in clusters has been investigated. The results showed that the early 2-cell stage ATAC peaks resulted significantly enriched in genic clusters with respect to the rest of the genome (z-score FDR=1.3E-132) (**Figure-5.5H**). Importantly, this feature appeared to be stage-specific with neither 2- nor 4-cell stage ATAC peaks showing such enrichment and displaying instead a significant depletion (z-score FDR=1.8E-93 and FDR=5.2E-07, respectively) (**Figure-5.5H**).

To annotate all the ATAC peaks, independently from their proximity or not with upregulated genes and TEs, the overlap between the identified peaks and the annotated genomic features (*i.e.*, genic promoter, UTR, exon, intron, immediately downstream or intergenic) has been assessed for each of the analysed stages. As expected, in all the samples, the majority of the ATAC peaks overlapped genic features and in particular the gene TSS (promoter <= 1kb) (**Figure-5.5I**). However, while for all the other samples 60-70% of the peaks resulted annotated in genic regions, in the early 2-cell stage only the 51% of the peaks overlapped genic features (**Figure-5.5I**). The remaining portion of the peaks (10,260 peaks) overlapped intergenic regions resulting in a significant enrichment of the early 2-cell ATAC peaks in intergenic regions with respect to the rest of the genome (Z-test FDR<2.2E-308). A further investigation of the localisation of these 10,260 early 2-cell intergenic peaks, showed that more than 80% of them overlapped annotated TEs (**Figure-5.5J**). Surprisingly, LINE L1 elements, and not ERVL, resulted the TEs mostly in overlap with the early 2-cell intergenic peaks with more than 50% of the intergenic peaks overlapping LINE L1 elements (5,683 peaks) resulting in a significant enrichment of the intergenic early 2-cell ATAC peaks in LINE L1 sequences (Z-test FDR<2.2E-308) (**Figure-5.5J**). Additionally, such enrichment resulted stage-specific with no other analysed embryonic stages displaying a similar pattern. When the TE composition of the early 2-cell ATAC peaks nearby the gene TSS (+/- 100 nt) was inspected, peaks were enriched in

ERVL elements (Z-test FDR=1.0E-16) but not in LINE L1 elements (Z-test FDR=2.3E-01) (**Figure-5.5K**) being consistent with previous observations showing how the early 2-cell/zygote upregulated genes are enriched nearby the TSS of ERVL sequences and depleted of LINE L1 (**Figure-5.3A**).

Altogether, these results display that, specifically in the early 2-cell stage, more than one fourth of the open chromatin domains are located in correspondence of intergenic LINE L1 elements suggesting LINE L1 implication in chromatin opening at this developmental stage.



Legend next page

Figure-5.5: a unique chromatin landscape characterises the mouse early embryo upon ZGA minor wave.
(**A**) Number of ATAC peaks in each analysed sample. The number of ATAC peaks remarkably increases (10-fold) upon murine ZGA minor wave (between early 2-cell + alpha-amanitin and early 2-cell samples) reaching the peak in the 2-cell stage and then gradually decreasing in later stages. (**B**) Minigene plot representing the early 2-cell ATAC-seq reads distribution nearby the early 2-cell/zygote upregulated genes. The ATAC signal is noisy and shows two peaks nearby the upregulated gene TSS and TTS. (**C**) Minigene plot representing the 2-cell ATAC-seq reads distribution nearby the 2-cell/early 2-cell upregulated genes. The signal displays a sharper profile than the one observed in early 2-cell. Two highly supported and sharp peaks are observable in correspondence of the upregulated gene TSS and TTS. (**D**) Minigene plot representing the 4-cell ATAC-seq reads distribution nearby the 4-cell/2-cell upregulated genes. Similarly to the 2-cell stage, peaks at both upregulated gene TSS and TTS are observable. (**E**) Minigene plot representing the early 2-cell ATAC-seq reads distribution nearby the early 2-cell/zygote upregulated TEs. The ATAC signal covers the entire TE body, with the peak probably counting the TEs (**F**) Minigene plot representing the 2-cell ATAC-seq reads distribution nearby the 2-cell/early 2-cell upregulated TEs. As in (E) the ATAC peaks entirely covers the TE body. (**G**) Minigene plot representing the 4-cell ATAC-seq reads distribution nearby the 4-cell/2-cell upregulated TEs. As in (E) and (F) the ATAC peaks entirely covers the TE body. An additional ATAC peaks is observable 5/6 kb downstream to the TE TTS. (**H**) ATAC peaks localisation within genic clusters. Early 2-cell ATAC peaks, but not the 2- and 4-cell stage peaks, are enriched within genic cluster compared to the rest of the genome. (**I**) ATAC peak genomic distribution. For each analysed stage, the ATAC peaks have been annotated as overlapping promoter, UTRs, exon, intron, immediately downstream or intergenic regions. In case of overlap, priorities have been assessed following the order reported in the figure legend. (**J**) Intergenic ATAC peak genomic distribution. The genomic coordinate of the intergenic ATAC peaks defined in (I) have been overlapped with the ones of the annotated TEs. (**K**) Promoter (+/-100nt) ATAC peak genomic distribution. The genomic coordinate of the promoter (+/-100nt) ATAC peaks have been overlapped with the ones of the annotated TEs. (E, F and G, S (start) and E (end) refer to the start and end of the TE coordinates. In H mean ± standard deviation is represented by the bars. ***FDR<0.001. FDR values refer to z-score derived BH FDR-corrected P-values. 100 randomisations).

5.3 Discussion

Mounting evidence had described how, upon ZGA minor and major waves, the transcriptional and epigenetic landscapes of the murine zygotic genome drastically change. Although most of the players involved in the murine ZGA are known, it is still unclear whether the genomic loci of such activated genes share common structural genomic organisation. Toward this end, taking advantage of mouse early embryo RNA-seq and ATAC-seq publicly available datasets (J. Wu et al., 2016), the transcriptional and epigenetic dynamics characterising the murine ZGA minor wave and the genomic structural organisation of the activated loci have been investigated.

The results showed how upon ZGA minor wave (between early 2-cell and zygote stages) a remarkable number of coding/non-coding genes (861) and TE loci (2,589) is transcriptionally activated in the mouse early embryo. These data were furtherly supported by the 10-fold increase in the number of open chromatin domains observed upon ZGA minor wave and by the enrichment of such domains in correspondence of the upregulated gene TSS and of the TE genomic loci, as displayed by the visualisation of the ATAC-seq peaks. Intriguingly, by the investigation of the structural organisation of the genomic loci of the early 2-cell/zygote upregulated genes and TEs as well as of the early 2-cell ATAC peaks, genes, TEs and peaks are not homogenously distributed along the genome resulting instead enriched in genic clusters.

Consistent with previous observations (De Iaco et al., 2017; Hendrickson et al., 2017), the TE transcriptional activation observed in the early 2-cell does not involve homogenously all the TE families with more than 50% of the upregulated TEs resulting annotated as ERVL. However, given the tendency of ERVL in generating chimeric transcripts with early expressed genes at these stages (Macfarlan et al., 2012; Peaston et al., 2004), it may be possible that the observed ERVL expression is a reflection of the transcription of ERVL as exonised fragments embedded in coding/non-coding early expressed genes rather than representing a specific ERVL transcription. Thus, the TE expression has been additionally calculated using TEspeX (described in the Chapter II of this thesis) that discards all the RNA-seq reads deriving from putative co-transcriptional events. The data remarkably showed how the transcription of the MERVL elements, a specific subfamily of ERVL, result

specifically activated upon ZGA minor wave supporting the transcription of such TEs as independent units. Moreover, consistent with previous observations (Macfarlan et al., 2012; Peaston et al., 2004), annotated ERVL were also found to be located nearby the early 2-cell upregulated gene TSS suggesting that, ERVL elements modulate the transcription of such early expressed genes. Intriguingly, while being enriched in ERVL, the early expressed genes resulted depleted of LINE L1 sequences nearby their TSS. This observation results particularly interesting especially when considering previous data describing how LINE L1 transcripts act as chromatin remodellers in the mouse early embryo (Jachowicz et al., 2017). Therefore, it may be reasonable to hypothesise that, in order to be successfully transcribed and to avoid a hypothetical LINE L1 mediated transcriptional repression, the early 2-cell/zygote upregulated genes must be depleted of LINE L1 sequences. Furtherly investigation of the chromatin landscape of the mouse embryo, showed that, specifically in the early 2-cell stage, the identified ATAC peaks are enriched in intergenic regions (49% of the total peaks) that are, in turn, enriched in LINE L1 sequences. Indeed, more than 50% of the total early 2-cell intergenic peaks (5,683 peaks), corresponding to more than one fourth of the total early 2-cell peaks, overlapped LINE L1 elements. On the contrary, the ATAC peaks located nearby the gene TSS (+/- 100nt) resulted enriched in ERVL but not in LINE L1 elements suggesting that LINE L1 might mediated the chromatin opening in the intergenic regions whereas ERVL nearby the gene TSS.

According to these observations, a possible model would define LINE L1, ERVL and genic clusters as key players in the activation the zygotic genome. In this scenario, LINE L1 may have a role in the broad chromatin opening of the early embryo with more than one fourth of the early 2-cell ATAC peaks overlapping LINE L1 elements. This observation would be consistent with previous results displaying how LINE L1 transcripts act as chromatin remodellers in the mouse early embryo (Jachowicz et al., 2017). ERVL, on the contrary, would have a more gene-specific role with almost 20% of the early 2-cell/zygote upregulated genes overlapping ERVL elements nearby the TSS. At the TSS region, it may be speculated that ERVL may drive the chromatin opening then driving, as previously described (Macfarlan et al., 2012; Peaston et al., 2004), the transcription of such genes. Importantly, the early 2-cell/zygote upregulated genes are enriched in clusters and, in turn, such cluster genes are enriched in ERVL. Thus, it may be reasonable

to propose that, upon ZGA minor wave, the transcription begins from genic clusters and is promoted by ERVL elements. Given the intrinsic genomic structure of the genic clusters and the repetitive nature of TEs, it may be kinetically convenient for the embryo, especially in these transcriptional immature phases, to initiate the transcription from such loci. Indeed, considering that the genic clusters are composed by several gene copies sharing a high sequence similarity, just one regulator/activator (possibly an ERVL) can be capable to activate many genes, at the same time.

From an evolutionary perspective, this entire scenario may be the consequence of a long-lasting evolutionary arm race between TE and host-genome. On one hand TEs, ERVL in this case, evolved *cis*-regulatory sequence in order to be transcribed, and possibly transposed, in the early stages of the embryonic development to insert new copies in the genome before the germline specification increasing their likelihood to be vertically transmitted. On the other, given the detrimental effects that TE transposition may have on the embryo physiology, the embryo has evolved mechanisms to exploit to his advantage the LINE L1 and ERVL sequences co-opting part of their sequences to regulate the transcriptional activation of the zygotic genome.

In summary, these results propose how transposable elements and genic clusters may influence the zygotic genome activation. While confirming the role of ERVL in promoting the transcription of early expressed genes, these data have extensively showed how the genomic loci from which the transcription is first initiated in the mouse early embryo reside on genic clusters. Additionally, these data have provided preliminary observations on how a consistent fraction of the early 2-cell open chromatin domains overlap LINE L1 elements. However, additional evidence has to be provided to better define the biological meanings underlying such observation. In particular, the expression profile and the genomic features of these LINE L1 have to be defined in order to define *if*, *when* and *how* these elements are expressed and influence specific transcriptional dynamics or biological pathways of the mouse embryo.

5.4 Methods

Data collection and pre-processing

To study the transcriptional and epigenetic dynamics characterising the zygotic genome during the murine MZT, RNA-seq and ATAC-seq datasets have been retrieved from Wu *et. al* (J. Wu et al., 2016). The RNA-seq dataset is composed by 8 embryonic stages dataset (MII-oocyte, zygote, early 2-cell, 2-cell, 4-cell, 8-cell, ICM, mESC), each represented by 2 or 3 biological replicates (**Table-5.1**).

Study accession	Experiment accession	Run accession	Sample title
PRJNA277361	SRX1424863	SRR2927026	RNA-Seq MII oocyte rep1
PRJNA277361	SRX1424864	SRR2927027	RNA-Seq MII oocyte rep2
PRJNA277361	SRX1424865	SRR2927028	RNA-Seq early 2-cell rep1
PRJNA277361	SRX1424866	SRR2927029	RNA-Seq early 2-cell rep2
PRJNA277361	SRX902549	SRR1840514	RNA-Seq zygote rep1
PRJNA277361	SRX902550	SRR1840515	RNA-Seq zygote rep2
PRJNA277361	SRX902551	SRR1840516	RNA-Seq 2-cell rep1
PRJNA277361	SRX902552	SRR1840517	RNA-Seq 2-cell rep2
PRJNA277361	SRX902553	SRR1840518	RNA-Seq 4-cell rep1
PRJNA277361	SRX902554	SRR1840519	RNA-Seq 4-cell rep2
PRJNA277361	SRX902555	SRR1840520	RNA-Seq 8-cell rep1
PRJNA277361	SRX902556	SRR1840521	RNA-Seq 8-cell rep2
PRJNA277361	SRX902557	SRR1840522	RNA-Seq ICM rep1
PRJNA277361	SRX902558	SRR1840523	RNA-Seq ICM rep2
PRJNA277361	SRX902561	SRR1840526	RNA-Seq ICM rep5
PRJNA277361	SRX902562	SRR1840527	RNA-Seq mESC rep1
PRJNA277361	SRX902562	SRR1840528	RNA-Seq mESC rep1
PRJNA277361	SRX902563	SRR1840529	RNA-Seq mESC rep2
PRJNA277361	SRX902563	SRR1840530	RNA-Seq mESC rep2

Table-5.1: mouse early embryo RNA-seq dataset.
The RNA-seq dataset from Wu *et al.* (J. Wu et al., 2016) is composed by 19 samples representing 8 embryonic stages, each represented by 2 or biological replicates.

The ATAC-seq dataset is composed by 7 embryonic stages (early 2-cell + alpha-amanitin, early 2-cell, 2-cell, 4-cell, 8-cell, ICM, mESC), each represented by 2 biological replicates, except for the mESC represented by a single replicate (**Table-5.2**).

Study accession	Experiment accession	Run accession	Sample title
PRJNA277362	SRX1424704	SRR2927014	ATAC-Seq early 2-cell alpha amanitin treatment rep1
PRJNA277362	SRX1424704	SRR3545579	ATAC-Seq early 2-cell alpha amanitin treatment rep1
PRJNA277362	SRX1680987	SRR3336394	ATAC-Seq early 2-cell alpha amanitin treatment rep2
PRJNA277362	SRX1680987	SRR3336395	ATAC-Seq early 2-cell alpha amanitin treatment rep2
PRJNA277362	SRX1424702	SRR2927010	ATAC-Seq early 2-cell rep1
PRJNA277362	SRX1424702	SRR3545576	ATAC-Seq early 2-cell rep1
PRJNA277362	SRX1424702	SRR3545577	ATAC-Seq early 2-cell rep1
PRJNA277362	SRX1424703	SRR2927012	ATAC-Seq early 2-cell rep2
PRJNA277362	SRX1424703	SRR2927013	ATAC-Seq early 2-cell rep2
PRJNA277362	SRX1424703	SRR3545578	ATAC-Seq early 2-cell rep2
PRJNA277362	SRX1424705	SRR2927015	ATAC-Seq 2-cell rep1
PRJNA277362	SRX1424705	SRR2927016	ATAC-Seq 2-cell rep1
PRJNA277362	SRX1424705	SRR3545580	ATAC-Seq 2-cell rep1
PRJNA277362	SRX1424706	SRR2927017	ATAC-Seq 2-cell rep2
PRJNA277362	SRX1424706	SRR2927018	ATAC-Seq 2-cell rep2
PRJNA277362	SRX902535	SRR1840426	ATAC-Seq 4-cell rep1
PRJNA277362	SRX902535	SRR3472347	ATAC-Seq 4-cell rep1
PRJNA277362	SRX902535	SRR3545573	ATAC-Seq 4-cell rep1
PRJNA277362	SRX1424708	SRR2927020	ATAC-Seq 4-cell rep2
PRJNA277362	SRX1424708	SRR3401496	ATAC-Seq 4-cell rep2
PRJNA277362	SRX1424708	SRR3545581	ATAC-Seq 4-cell rep2
PRJNA277362	SRX1424709	SRR2927021	ATAC-Seq 8-cell rep1
PRJNA277362	SRX1424709	SRR3401556	ATAC-Seq 8-cell rep1
PRJNA277362	SRX1424709	SRR3401562	ATAC-Seq 8-cell rep1
PRJNA277362	SRX1424710	SRR2927022	ATAC-Seq 8-cell rep2
PRJNA277362	SRX1424710	SRR3401567	ATAC-Seq 8-cell rep2
PRJNA277362	SRX1424710	SRR3401568	ATAC-Seq 8-cell rep2
PRJNA277362	SRX1424711	SRR2927023	ATAC-Seq ICM rep1
PRJNA277362	SRX1424711	SRR3545582	ATAC-Seq ICM rep1
PRJNA277362	SRX1770459	SRR3536933	ATAC-Seq ICM rep2
PRJNA277362	SRX1770459	SRR3536934	ATAC-Seq ICM rep2
PRJNA277362	SRX1770459	SRR3536935	ATAC-Seq ICM rep2
PRJNA277362	SRX902546	SRR1840439	ATAC-Seq 200 mESC
PRJNA277362	SRX902546	SRR1840440	ATAC-Seq 200 mESC
PRJNA277362	SRX902546	SRR1840441	ATAC-Seq 200 mESC
PRJNA277362	SRX902546	SRR4011728	ATAC-Seq 200 mESC

Table-5.2: mouse early embryo ATAC-seq dataset.
The ATAC-seq dataset from Wu *et al.* (J. Wu et al., 2016) is composed by a total of 37 samples, representing 7 embryonic stages, each represented by 2 biological replicated, except for the mESC sample.

The embryonic stages represented by the RNA-seq and ATAC-seq dataset mostly overlap as six out of 8 RNA-seq samples have a counterpart in the ATAC-seq dataset. The ATAC-seq dataset lacks the MII-oocyte ad zygote as it is technically challenging to extract enough chromatin to perform an ATAC-seq at these stages. To overcome this issue, the authors of the paper treated early zygotes (PN3 – pronucleus phase 3) in CZB medium supplemented with alpha-amanitin for about 14 h (J. Wu et al., 2016). Alpha-amanitin is an inhibitor of the RNA-pol-II, thus alpha-amanitin treated zygotes are transcriptionally inactive representing a pre-ZGA-like sample.

After having retrieved the raw RNA and ATAC-seq reads of both datasets from the ENA-EBI database, the technical replicate files corresponding to the same experiment have been merged. Next, the quality of the raw reads have been assessed by using FastQC (Andrews, 2010). Having detected the presence of both adapters and low-quality reads, the sequencing reads of both datasets have been trimmed using Trimmomatic (v0.38, parameters: ILLUMINACLIP:NexteraPE-PE.fa:2:30:10:2:keepBothReads, LEADING:5 TRAILING:5, SLIDINGWINDOW:4:15 MINLEN:30) (Bolger et al., 2014).

RNA-seq dataset analysis – gene expression

To quantify the gene expression values of the coding and non-coding genes annotated in the mouse genome, the RNA-seq reads have been mapped on the murine genome (mm10 version – gencode vM22 version) using STAR (v2.6.0c) (Dobin et al., 2013). Default parameters have been used except for the number of multimapping reads that have been set to 80 (--outFilterMultimapNmax 80). The expression of both coding and non-coding genes annotated in the murine genome have been then quantified using htseq-count (v0.11.2, parameters: --stranded no -m union --nonunique all) (Anders et al., 2015). Next, to identify the differentially expressed genes (DE) between the early 2-cell and the zygote stage edgeR has been used (Robinson et al., 2010). Normalisation of raw read counts has been applied using the TMM method whereas the common, trended and tagwise dispersions have been estimated by maximizing the negative binomial likelihood (default). Next, differentially expressed genes have been tested performing a quasi-likelihood F-tests (glmQLFit and glmQLTest). Genes have been considered as differentially expressed when showing FDR < 0.05 and log2FC < -1 or > 1 (2-fold in linear

scale). The same analysis has been repeated to detect differentially expressed between 2- and early 2-cell and 4- and 2-cell stages.

RNA-seq dataset analysis – TE expression

TE locus specific expression levels have been calculated using SQuIRE (Yang et al., 2019). First, the reference genome and the annotation datasets referring to the murine mm10 genome version have been downloaded and prepared for the subsequent analyses using the SQuIRE Fetch and Clean modules, then the trimmed reads have been mapped on the reference genome using the Map module and finally read counts have been estimated using the Count module (strandedness='0'). Elements annotated as DNA, LINE, SINE, LTR and RC have been selected and differentially expressed TE loci have been identified using edgeR as previously described. TE loci showing FDR < 0.05 and log2FC <-1 or >1 have been considered as differentially expressed.

In order to summarize the expression levels of specific TE consensus the TEspeX pipeline previously described has been used. Briefly, a reference transcriptome is built merging the RepBase TE sequences (Bao et al., 2015) and the Ensembl transcript sequences containing all the coding and non-coding annotated transcripts (Zerbino et al., 2018). Reads are then mapped on the reference transcriptome using STAR (v2.6.0c) (Dobin et al., 2013) and assigning primary alignment flag to all the alignments with the best score. All alignments flagged as primary (−F 0 × 100 parameter) are then selected using samtools (v1.3.1) (H. Li et al., 2009). To avoid counting reads mapping on TE fragments embedded in coding and/or long non-coding transcripts, reads mapping with best-scoring alignments on any Ensembl transcript are discarded using Python scripts and Picard FilterSamReads (v2.18.4) (http://broadinstitute.github.io/Picard). Selected reads mapping exclusively on TEs and in the proper orientation are finally counted in each sample. Differentially expressed TE consensus sequences have been performed using edgeR as previously described except for the library size of each sample that has been calculated providing the total number of reads mapped on the transcriptome (coding, non-coding and TE consensus sequences) instead of using the default values.

Gene ontology (GO) enrichment analysis

GO enrichment analysis has been performed by using topGO (Alexa & Rahnenfuhrer, 2019). GO enrichment analysis has been conducted on the GO terms associated to the early 2-cell/zygote upregulated genes, using as background the GO terms associated to the whole set of coding and non-coding annotated genes. First, the statistical significance of the enrichments has been tested with the Fisher's Exact Test (algorithm='weight'). Then, GO terms associated to less than 15 significant genes have been discarded prior to FDR calculation (Benjamini & Hochberg). Significant threshold has been imposed to FDR < 0.05.

Upregulated genes chromosomal and structural organisation

To define whether early 2-cell/zygote upregulated genes are significantly enriched in specific chromosomes, compared to the rest of the transcriptome, the number of upregulated genes on each murine chromosome (excluding scaffolds and mitochondrial chromosome) have been counted. The same analysis has been repeated selecting an equal number of randomly selected genes 100 times. Z-score has been consequently calculated subtracting to the number of upregulated genes on each chromosome the mean number of random genes on the same chromosome and dividing by the standard deviation of the number of random genes on that chromosome (z-score describes the position of a raw score in terms of its distance from the mean, when measured in standard deviation units). P-value has next been calculated from z-score and next corrected using the FDR Benjamini & Hochberg correction. FDR significant threshold has been set to 0.05.

To define whether the upregulated genes are significantly enriched in genic clusters compared to the rest of the transcriptome, first genic clusters composed by at least 5 genes have been identified using ClusterScan (parameters: -n 5 -d 500000 –singletons) (Volpe et al., 2018). Next, the number of upregulated genes belonging to one of the previously defined clusters has been calculated. The genes have been considered as part of specific gene clusters when overlapping with at least the 50% of their length the cluster. The same analysis has been performed on an equal number of random genes calculating z-score, P-value and FDR as previously described. The same workflow has been followed to calculate the enrichment of upregulated TEs and ATAC peaks in genic cluster. To further analyse the cluster enrichment of the early 2-cell/zygote upregulated

genes the same analysis has been repeated cluster by cluster. The number of upregulated genes in each cluster has been calculated, performing then the same analysis on an equal number of random genes as previously described.

Upregulated genes/annotated TEs overlap analysis

To define the TE occupancy nearby the TSS of the early 2-cell/zygote upregulated genes, the number of upregulated gene TSS overlapping annotated TEs has been calculated. As previously described, only TEs belonging to DNA, RC, LTR, LINE and SINE classes have been considered in the analysis. The upregulated genes/annotated TEs overlap analysis has been performed considering the gene TSS +/- 100 nt. The same analysis has been repeated on randomly selected genes following the previously described methods to define a statistically significant enrichment or depletion. This analysis has been performed on the total number of early 2-cell/zygote upregulated genes as well as on the upregulated genes divided according to the functional annotation of the genic clusters they belong to. In this second analysis, the number of randomisations has been increased to 10,000 given the high variability deriving from the low number of genes composing each of the analysed clusters. The minigene plots represented in Figure-5.3B have been generated by using custom scripts developed in python3. Briefly, each upregulated gene has been subdivided in an equal number of bins. Each bin of each gene has then been overlapped with the investigated element (i.e., ERVL or LINE L1). When at least the 50% of the nucleotides of the bin overlapped the investigated element, 1 was assigned to the bin otherwise 0 was assigned generating, for each gene, a string of 0 and 1. Finally, for each bin, the mean value between all the genes has been calculated, multiplied by 100 and plotted. On the y-axis, for each bin, is thus represented the percentage of analysed genes covered by the investigated element.

ATAC-seq dataset analysis

To identify the genomic regions marked by open chromatin domains, ATAC-seq reads have been retrieved and trimmed as previously described. The ATAC-seq reads have then been mapped on the murine genome (mm10 version – gencode vM22 version) using bowtie2 (v2.3.5.1) (Langmead & Salzberg, 2012). Default parameters have been used when mapping single-ended reads whereas paired-ended reads have been mapped avoiding the selection of both discordant pairs and singleton reads (--no-mixed, --no-

discordant parameters). Next, the Genrich tool has been used to perform the ATAC peak calling analysis merging the biological replicates (v0.6, parameters: -j -d 250 -r -a 0 -q 0.05 -e chrM and genomic scaffolds) (https://github.com/jsh58/genrich). The Genrich-generated BAM files have next been converted in bigWigs format using the deepTools tool bamCoverage (v3.5.0, parameters: --normalizeUsing RPKM --extendReads 250 as in (J. Wu et al., 2016)) (Ramírez et al., 2016). BigWigs files have been next used for the peak visualisation on the integrative genome viewer and for the generation of the matrix used to generate the minigene plots represented in Figure-5.5B-G. Briefly, the matrix has been generated providing the genomic coordinates of the upregulated genes (early 2-/zygote, 2-/early 2- and 4-/2-cell) and the corresponding bigWigs files to the deepTools tool computeMatrix (v3.5.0, parameters for protein coding/non-coding genes: --beforeRegionStartLength 50000, --afterRegionStartLength 50000, --regionBodyLength 20000, --binSize 250, --maxThreshold 100, --skipZeros. Parameters for TEs: --beforeRegionStartLength 25000, --afterRegionStartLength 25000, --regionBodyLength 2000, --binSize 50, --maxThreshold 10, --skipZeros) (Ramírez et al., 2016). The deepTools computeMatrix and the custom python tool described in the previous section rely on similar algorithm rationale based on the subdivision of the analysed genes in bins. However, while the custom python script computes the overlap between genomic coordinates of specified features (i.e., genes and TEs), the computeMatrix tool calculates, for each gene bin, the overall value of ATAC-seq reads mapped in the bin. Having generated the matrix, the mean value of ATAC-seq reads between all the genes has been calculated for each bin, standardised (z-score standardisation as in (J. Wu et al., 2016)) and plotted.

ATAC peaks annotation

To annotate the identified ATAC peaks the R/Bioconductor package ChIPseeker has been used (v1.24.0) (Yu et al., 2015). The TxDB object has been generated from the gtf annotation file used for previous analyses (gencode vM22 version, primary assembly) using the R/Bioconductor package GenomicFeatures (v1.40.1) (Lawrence et al., 2013). ATAC peaks have next been annotated taking advantage of the ChIPseeker annotatePeak function defining a promoter region of +/- 3kb (tssRegion=c(-3000, 3000)) and plotted by the plotAnnoBar function. In case of ATAC peaks overlapping more than one genomic feature, priorities have been assigned following the default parameters (promoter, 5'

UTR, 3' UTR, exon, intron, downstream, intergenic). The overlap between the intergenic ATAC peaks and the annotated TEs has been made by using bedtools intersect (v2.27.0, parameters: -f 0.5, -wao) (Quinlan & Hall, 2010). Positive intersections have been defined when at least half of the ATAC peak overlapped a specific TE. In case of peaks overlapping more than one annotated TEs, the TE showing the longest overlap has been selected.

Statistical analysis

All the statistical analyses performed externally to previously reported software (edgeR, topGO) have been conducted either in R (v3.6.2) (R Core Team, 2018) or in python (v3.7.6) (Rossum & Drake, 2001) taking advantage of the numpy (Harris et al., 2020) and scipy (SciPy 1.0 Contributors et al., 2020) libraries. PCA analyses has been performed using the ggbiplot R library. All the plots have been generated in R, using either generic R plotting functions or the ggplot2 library (Wickham, 2016).

Chapter 6
Concluding remarks and future perspectives

The research interests motivating my PhD were mainly centred on the investigation of the role that transposable element (TE) transcription plays during the Metazoan embryogenesis. Mounting evidence has indeed shown how TEs result remarkably transcribed during the initial phases of the embryonic development of several Metazoan species, including *Drosophila*, mouse and human (De Iaco et al., 2017; Hendrickson et al., 2017; Parkhurst & Corces, 1987). This phenomenon is probably linked with the intense epigenetic reorganizations undergoing in the zygotic genome during the earliest developmental phases that lead to the overall loss of heterochromatic regions and to the subsequent transcription of elements whose expression is normally silenced by such mechanisms in adult cells, TEs included (Eckersley-Maslin et al., 2018). However, at least in the mouse model, TE transcription is not just a passive process resulting from the loss of heterochromatic regions as it is instead deeply interconnected with crucial physiological events required for the activation of the zygotic genome (Macfarlan et al., 2012; Peaston et al., 2004; Torres-Padilla, 2020). From these observations, the interest in investigating i) the transcriptional dynamics characterising the TE expression in the Metazoan early embryos, ii) the grade of conservation of this phenomenon across different species and iii) whether the TE expression influences the transcriptional dynamics leading to the activation of the zygotic genome rose. However, no reliable bioinformatics tools capable to quantify the TE expression from RNA-seq data were available at the time when my PhD started. Thus, the first goal of my PhD project was to develop a bioinformatics pipeline, called TEspeX, capable to quantify the TE expression from RNA-seq datasets without being biased by the transcription of TE fragments embedded in coding/non-coding transcripts. Having developed and validated TEspeX on both simulated and real datasets, the TE transcriptional landscape of three Metazoan species (*C. elegans*, zebrafish and mouse) has been defined.

Given the lack of knowledge regarding the TE transcriptional landscape in the *C. elegans* embryo with no reports describing evidence of TE expression in this model, the first analysed species was the nematode *C. elegans*. Importantly, since the adult *C. elegans* is composed by approximately 1,000 cells, the gastrulation period begins when the embryo is still composed by a relatively small number of cells and as early as in the 16-cell stage the fate of all the embryo cells starts to be determined (Maduro, 2010; Sulston et al., 1983). Therefore, by quantifying the TE expression in each single-cell of the embryo from the 1- to the 16-cell stages the TE transcriptional landscape has been defined. Importantly, it has also been possible to investigate the role of TEs in driving the differentiation of particular sub-populations of cells toward specific phenotypes. A first overview of the TE transcriptional landscape of the *C. elegans* embryo highlighted how different TE classes present different expression profiles in different embryonic stages. In particular, while DNA transposons result constantly expressed among all the analysed cell types, LTR and non-LTR retrotransposons present opposite transcriptional profiles with the former resulting expressed in the initial stages whereas the latter in later ones. Interestingly, LTR retrotransposon expression resulted correlated with the expression of genes involved with the activation of the innate immune response. Intriguingly, in human embryonic stem cells it has been shown how the LTR retrotransposon HERVK, by encoding the viral-like *Rec* protein, activates the antiviral response protecting the embryo from exogenous infections (Grow et al., 2015). Although my results do not provide any direct evidence of full-length LTR retrotransposons maintaining their coding potential and thus encoding viral-like proteins, is nevertheless tempting to hypothesise that such immunoprotective function is similarly associated to both *C. elegans* and human LTR retrotransposons. To test this hypothesis, future studies could first identify *in silico* full-length LTR elements expressed at these stages and next, upon their silencing, measuring the embryo susceptibility to viral and bacterial infections. On the contrary, non-LTR retrotransposons resulted expressed in later developmental stages and especially in cell sub-types giving rise to neurons and tissues connected with the nervous system. Importantly, multiple lines of evidence have described a similar phenomenon in *Drosophila* and mammals where retrotransposons, and in particular non-LTR retrotransposons, result expressed and actively retrotransposed in neuronal precursor cells and in specific neuronal sub-populations (Coufal et al., 2009; Evrony et al., 2012; Muotri et al., 2005; Perrat et al., 2013). To validate whether also in *C. elegans* non-LTR

retrotransposons result retrotranspositionally active in neuronal precursor lineages, at first expressed full-length non-LTR retrotransposons should be identified and next retrotransposition events of such elements should be detected either by copy number variation TaqMan quantitative PCR assay or by generating and mining whole genome sequencing data.

Given the unicity of the *C. elegans* maternal to zygotic transition (MZT), characterised by a post-transcriptional and post-translational regulation rather than a transcriptional one and with the zygotic genome activation (ZGA) occurring with different timings in somatic and germline precursor cells (Robertson & Lin, 2015), the impact that TE expression might have on the transcriptional dynamics characterising ZGA is not easily assessable in this species and it has not been object of this study. On the contrary, the hypothetical transcriptional relationship between coding/non-coding genes and TEs at the ZGA onset is more easily definable in two vertebrate species, widely used in embryogenesis studies, such as zebrafish and mouse. Toward this end, the transcriptional dynamics characterising the zebrafish and mouse ZGA have been investigated. Consistent with the analysed biological context, characterised by intense waves of transcription, a remarkable fraction of coding/non-coding genes resulted transcriptionally activated upon ZGA, in both species. Moreover, the broad transcriptional activation occurring at these stages did not exclusively involved coding/non-coding genes as TEs resulted transcriptionally activated as well. However, while this activation resulted modest in zebrafish (99 TE loci), a remarkable number of TEs resulted transcribed at the ZGA onset in mouse (2,589 TE loci). The transcription of TEs upon murine ZGA, and in particular of ERVL LTR retrotransposons, is certainly not a surprise as several studies have previously described it, additionally showing how probably it is *Dux* itself, the main regulator of the mammals ZGA, to activate the ERVL element transcription (De Iaco et al., 2017; Hendrickson et al., 2017). On the other hand, these data were the first to describe a similar, even if less massive, TE transcriptional activation in the zebrafish embryo at ZGA. In particular, TEs belonging to the DNA transposon hAT and to the LTR retrotransposon Gypsy families resulted the most abundant TEs transcriptionally activated at ZGA highlighting, once again, the crucial role LTR retrotransposons play during the embryo development of Metazoans. Moreover, by the investigation of the genomic loci in which the transcriptionally activated genes and TEs reside on, I showed they are not randomly

distributed along the genome being instead located in transcriptionally dense genomic compartments. Although in both zebrafish and mouse a similar observation has already been described for specific set of genes (*i.e.*, zebrafish *miR-430* and *znf* gene clusters and murine *Zscan4* gene cluster) (Falco et al., 2007; Giraldez et al., 2006; Hadzhiev et al., 2019; Heyn et al., 2014; White et al., 2017), these data are the first to report it for a broad fraction of genes and to extend a similar observation to TEs. Based on these results and following previous speculations (Hadzhiev et al., 2019), it is tempting to hypothesise that there has been a selective pressure inducing the generation of this transcriptionally dense environments in order to generate genomic loci functioning as an aggregation of transcription factors and thus facilitating the activation of the zygotic genome. Moreover, given the intrinsic genomic structure of the genic clusters and the repetitive nature of TEs, it may be kinetically convenient for the embryo, especially in these transcriptional immature phases, to initiate the transcription from such loci. Indeed, considering the high sequence similarity characterising the gene copies of the same cluster and the different TE loci of the same family, just few activators are needed to activate many genes and TEs, at the same time and in a coordinated process. Given the repetitive nature of gene clusters and TEs, a further validation of these results is not an easy task, as it is not clearly feasible to silence and/or remove the entire set of transcribed gene clusters and TEs. It may be considered to genetically delete the gene clusters primarily involved with the zebrafish and mouse ZGA such as the zebrafish *miR-430* cluster and the murine *Duxf3* locus. However, recent studies have shown how the genetic deletion of such loci causes minor defects in zygotic genome activation (Chen & Zhang, 2019; Liu et al., 2020). Probably, this is exactly due to the highly genomic redundancy of the transcriptional activators characterising the embryonic genomes that leads other activators to function when the main ones result mutated. This observation is also evolutionary consistent with hypothetical evolutionary forces leading the embryo to evolve several and redundant transcriptional activators in order to still be capable to successfully activate the zygotic genome in case of misfunctioning of one of such activators.

Finally, my results showed how, in the mouse early embryo, LINE L1 may play crucial roles in the activation of the zygotic genome. Indeed, while being depleted from the TSS of the early expressed genes, LINE L1 elements overlapped an incredibly large fraction of the intergenic open chromatin domains identified in the murine genome at the ZGA onset.

Consistent with previous observations describing how LINE L1 transcripts act as chromatin remodellers in the mouse early embryo epigenetically repressing specific transcriptional programmes (Jachowicz et al., 2017; Percharde et al., 2018), it is likely that depending on their proximity with genic regions, LINE L1 might play different roles. However, further analyses aimed at a deeper investigation of the genomic, epigenomic and transcriptional properties of such LINE L1 elements are required to better understand these results.

In summary, these results strongly consolidate the idea that transposable elements actively shape the transcriptional dynamics underlying the Metazoans embryogenesis. Moreover, the functions transposable elements play within this context appear to be conserved across different Metazoan species thus suggesting the key role they play in such a crucial biological event as the embryogenesis is.

References

Abe, K.-I., Yamamoto, R., Franke, V., Cao, M., Suzuki, Y., Suzuki, M. G., Vlahovicek, K., Svoboda, P., Schultz, R. M., & Aoki, F. (2015). The first murine zygotic transcription is promiscuous and uncoupled from splicing and 3' processing. *The EMBO Journal*, *34*(11), 1523–1537. https://doi.org/10.15252/embj.201490648

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *Molecular Biology of the Cell* (4th edition). https://www.ncbi.nlm.nih.gov/books/NBK21054/

Alexa, A., & Rahnenfuhrer, J. (2019). *TopGO: Enrichment Analysis for Gene Ontology*.

Altun, Z.F., & Hall, D.H. (2009). Alimentary System, Pharynx. *WormBook*. http://www.wormatlas.org/hermaphrodite/pharynx/Phaframeset.html

Altun, Z.F., & Hall, D.H. (2011). Nervous system, general description. *WormBook*. http://www.wormatlas.org/hermaphrodite/nervous/Neuroframeset.html

Ancelin, K., Syx, L., Borensztein, M., Ranisavljevic, N., Vassilev, I., Briseño-Roa, L., Liu, T., Metzger, E., Servant, N., Barillot, E., Chen, C.-J., Schüle, R., & Heard, E. (2016). Maternal LSD1/KDM1A is an essential regulator of chromatin and transcription landscapes during zygotic genome activation. *ELife*, *5*, e08851. https://doi.org/10.7554/eLife.08851

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data.* http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Ansaloni, F., Scarpato, M., Di Schiavi, E., Gustincich, S., & Sanges, R. (2019). Exploratory analysis of transposable elements expression in the C. elegans early embryo. *BMC Bioinformatics*, *20*(Suppl 9), 484. https://doi.org/10.1186/s12859-019-3088-7

Aoki, F., Worrad, D. M., & Schultz, R. M. (1997). Regulation of transcriptional activity during the first and second cell cycles in the preimplantation mouse embryo. *Developmental Biology*, *181*(2), 296–307. https://doi.org/10.1006/dbio.1996.8466

Aravin, A. A., Hannon, G. J., & Brennecke, J. (2007). The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science (New York, N.Y.)*, *318*(5851), 761–764. https://doi.org/10.1126/science.1146484

Bachiller, S., del-Pozo-Martín, Y., & Carrión, Á. M. (2017). L1 retrotransposition alters the hippocampal genomic landscape enabling memory formation. *Brain, Behavior, and Immunity*, *64*, 65–70. https://doi.org/10.1016/j.bbi.2016.12.018

Baillie, J. K., Barnett, M. W., Upton, K. R., Gerhardt, D. J., Richmond, T. A., De Sapio, F., Brennan, P. M., Rizzu, P., Smith, S., Fell, M., Talbot, R. T., Gustincich, S., Freeman, T. C., Mattick, J. S., Hume, D. A., Heutink, P., Carninci, P., Jeddeloh, J. A., & Faulkner, G. J. (2011). Somatic retrotransposition alters the genetic landscape of the human brain. *Nature*, *479*(7374), 534–537. https://doi.org/10.1038/nature10531

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *6*, 11. https://doi.org/10.1186/s13100-015-0041-9

Barau, J., Teissandier, A., Zamudio, N., Roy, S., Nalesso, V., Hérault, Y., Guillou, F., & Bourc'his, D. (2016). The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. *Science*, *354*(6314), 909–912. https://doi.org/10.1126/science.aah5143

Baugh, L. R. (2003). Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. *Development*, *130*(5), 889–900. https://doi.org/10.1242/dev.00302

Bedrosian, T. A., Quayle, C., Novaresi, N., & Gage, F. H. (2018). Early life experience drives structural variation of neural genomes in mice. *Science (New York, N.Y.)*, *359*(6382), 1395–1399. https://doi.org/10.1126/science.aah3378

Bendall, M. L., de Mulder, M., Iñiguez, L. P., Lecanda-Sánchez, A., Pérez-Losada, M., Ostrowski, M. A., Jones, R. B., Mulder, L. C. F., Reyes-Terán, G., Crandall, K. A., Ormsby, C. E., & Nixon, D. F. (2019). Telescope: Characterization of the retrotranscriptome by accurate estimation of transposable element expression. *PLoS Computational Biology*, *15*(9), e1006453. https://doi.org/10.1371/journal.pcbi.1006453

Berh, J., Tymoczko, J., & Stryer, L. (2002). *Biochemistry* (5th ed.). https://www.ncbi.nlm.nih.gov/books/NBK21154/

Bessereau, J.-L. (2006). Transposons in C. elegans. *WormBook*. https://doi.org/10.1895/wormbook.1.70.1

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., & Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome Biology*, *19*(1), 199. https://doi.org/10.1186/s13059-018-1577-z

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

Bushati, N., Stark, A., Brennecke, J., & Cohen, S. M. (2008). Temporal reciprocity of miRNAs and their targets during the maternal-to-zygotic transition in Drosophila. *Current Biology: CB*, *18*(7), 501–506. https://doi.org/10.1016/j.cub.2008.02.081

Bushnell, B. (2014). *BBMap*. sourceforge.net/projects/bbmap/

Canapa, A., Barucca, M., Biscotti, M. A., Forconi, M., & Olmo, E. (2015). Transposons, Genome Size, and Evolutionary Insights in Animals. *Cytogenetic and Genome Research*, *147*(4), 217–239. https://doi.org/10.1159/000444429

Carmell, M. A., Girard, A., van de Kant, H. J. G., Bourc'his, D., Bestor, T. H., de Rooij, D. G., & Hannon, G. J. (2007). MIWI2 Is Essential for Spermatogenesis and Repression of Transposons in the Mouse Male Germline. *Developmental Cell*, *12*(4), 503–514. https://doi.org/10.1016/j.devcel.2007.03.001

Casale, A. M., Liguori, F., Ansaloni, F., Cappucci, U., Finaurini, S., Spirito, G., Persichetti, F., Sanges, R., Gustincich, S., & Piacentini, L. (2020). *Transposable Element activation promotes neurodegeneration in a* Drosophila *model of Huntington's disease* [Preprint]. Neuroscience. https://doi.org/10.1101/2020.11.19.389718

Chen, Z., & Zhang, Y. (2019). Loss of DUX causes minor defects in zygotic genome activation and is compatible with mouse development. *Nature Genetics*, *51*(6), 947–951. https://doi.org/10.1038/s41588-019-0418-7

Chisholm, A. D., & Xu, S. (2012). The Caenorhabditis elegans epidermis as a model skin. II: Differentiation and physiological roles. *Wiley Interdisciplinary Reviews. Developmental Biology*, *1*(6), 879–902. https://doi.org/10.1002/wdev.77

Chuong, E. B., Elde, N. C., & Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science*, *351*(6277), 1083–1087. https://doi.org/10.1126/science.aad5497

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

Clift, D., & Schuh, M. (2013). Restarting life: Fertilization and the transition from meiosis to mitosis. *Nature Reviews. Molecular Cell Biology*, *14*(9), 549–562. https://doi.org/10.1038/nrm3643

Conley, A. B., Piriyapongsa, J., & Jordan, I. K. (2008). Retroviral promoters in the human genome. *Bioinformatics (Oxford, England)*, *24*(14), 1563–1567. https://doi.org/10.1093/bioinformatics/btn243

Corsi, A. K., Wightman B., & Chalfie M. (2015). A Transparent window into biology: A primer on Caenorhabditis elegans. *WormBook*, 1–31. https://doi.org/10.1895/wormbook.1.177.1

Cosby, R. L., Chang, N.-C., & Feschotte, C. (2019). Host-transposon interactions: Conflict, cooperation, and cooption. *Genes & Development*, *33*(17–18), 1098–1116. https://doi.org/10.1101/gad.327312.119

Coufal, N. G., Garcia-Perez, J. L., Peng, G. E., Yeo, G. W., Mu, Y., Lovci, M. T., Morell, M., O'Shea, K. S., Moran, J. V., & Gage, F. H. (2009). L1 retrotransposition in human neural progenitor cells. *Nature*, *460*(7259), 1127–1131. https://doi.org/10.1038/nature08248

Coward, K., & Wells, D. (2013). Textbook of Clinical Embryology. In *Clinical Embryology*.

Dai, L., Taylor, M. S., O'Donnell, K. A., & Boeke, J. D. (2012). Poly(A) binding protein C1 is essential for efficient L1 retrotransposition and affects L1 RNP formation. *Molecular and Cellular Biology*, *32*(21), 4323–4336. https://doi.org/10.1128/MCB.06785-11

De Iaco, A., Planet, E., Coluccio, A., Verp, S., Duc, J., & Trono, D. (2017). DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nature Genetics*, *49*(6), 941–945. https://doi.org/10.1038/ng.3858

De Renzis, S., Elemento, O., Tavazoie, S., & Wieschaus, E. F. (2007). Unmasking Activation of the Zygotic Genome Using Chromosomal Deletions in the Drosophila Embryo. *PLoS Biology*, *5*(5), e117. https://doi.org/10.1371/journal.pbio.0050117

Deininger, P., Morales, M. E., White, T. B., Baddoo, M., Hedges, D. J., Servant, G., Srivastav, S., Smither, M. E., Concha, M., DeHaro, D. L., Flemington, E. K., & Belancio, V. P. (2017). A comprehensive approach to expression of L1 loci. *Nucleic Acids Research*, *45*(5), e31. https://doi.org/10.1093/nar/gkw1067

Dembny, P., Newman, A. G., Singh, M., Hinz, M., Szczepek, M., Krüger, C., Adalbert, R., Dzaye, O., Trimbuch, T., Wallach, T., Kleinau, G., Derkow, K., Richard, B. C., Schipke, C., Scheidereit, C., Stachelscheid, H., Golenbock, D., Peters, O., Coleman, M., … Lehnardt, S. (2020). Human endogenous retrovirus HERV-K(HML-2) RNA causes neurodegeneration through Toll-like receptors. *JCI Insight*, *5*(7), e131093. https://doi.org/10.1172/jci.insight.131093

Denli, A. M., Narvaiza, I., Kerman, B. E., Pena, M., Benner, C., Marchetto, M. C. N., Diedrich, J. K., Aslanian, A., Ma, J., Moresco, J. J., Moore, L., Hunter, T., Saghatelian, A., & Gage, F. H. (2015). Primate-specific ORF0 contributes to retrotransposon-mediated diversity. *Cell*, *163*(3), 583–593. https://doi.org/10.1016/j.cell.2015.09.025

Dewannieux, M., Esnault, C., & Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. *Nature Genetics*, *35*(1), 41–48. https://doi.org/10.1038/ng1223

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Eckersley-Maslin, M. A., Alda-Catalinas, C., & Reik, W. (2018). Dynamics of the epigenetic landscape during the maternal-to-zygotic transition. *Nature Reviews. Molecular Cell Biology*, *19*(7), 436–450. https://doi.org/10.1038/s41580-018-0008-z

Edgar, B. A., & Datar, S. A. (1996). Zygotic degradation of two maternal Cdc25 mRNAs terminates Drosophila's early cell cycle program. *Genes & Development*, *10*(15), 1966–1977. https://doi.org/10.1101/gad.10.15.1966

Engels, W. R. (1983). The P family of transposable elements in Drosophila. *Annual Review of Genetics*, *17*, 315–344. https://doi.org/10.1146/annurev.ge.17.120183.001531

Ermolaeva, M. A., & Schumacher, B. (2014). Insights from the worm: The C. elegans model for innate immunity. *Seminars in Immunology*, *26*(4), 303–309. https://doi.org/10.1016/j.smim.2014.04.005

Erwin, J. A., Marchetto, M. C., & Gage, F. H. (2014). Mobile DNA elements in the generation of diversity and complexity in the brain. *Nature Reviews. Neuroscience*, *15*(8), 497–506. https://doi.org/10.1038/nrn3730

Erwin, J. A., Paquola, A. C. M., Singer, T., Gallina, I., Novotny, M., Quayle, C., Bedrosian, T. A., Alves, F. I. A., Butcher, C. R., Herdy, J. R., Sarkar, A., Lasken, R. S., Muotri, A. R., & Gage, F. H. (2016). L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nature Neuroscience*, *19*(12), 1583–1591. https://doi.org/10.1038/nn.4388

Evrony, G. D., Cai, X., Lee, E., Hills, L. B., Elhosary, P. C., Lehmann, H. S., Parker, J. J., Atabay, K. D., Gilmore, E. C., Poduri, A., Park, P. J., & Walsh, C. A. (2012). Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell*, *151*(3), 483–496. https://doi.org/10.1016/j.cell.2012.09.035

Evrony, G. D., Lee, E., Park, P. J., & Walsh, C. A. (2016). Resolving rates of mutation in the brain using single-neuron genomics. *ELife*, *5*, e12966. https://doi.org/10.7554/eLife.12966

Fadloun, A., Le Gras, S., Jost, B., Ziegler-Birling, C., Takahashi, H., Gorab, E., Carninci, P., & Torres-Padilla, M.-E. (2013). Chromatin signatures and retrotransposon profiling in mouse embryos reveal regulation of LINE-1 by RNA. *Nature Structural & Molecular Biology*, *20*(3), 332–338. https://doi.org/10.1038/nsmb.2495

Falco, G., Lee, S.-L., Stanghellini, I., Bassey, U. C., Hamatani, T., & Ko, M. S. H. (2007). Zscan4: A novel gene expressed exclusively in late 2-cell embryos and embryonic stem cells. *Developmental Biology*, *307*(2), 539–550. https://doi.org/10.1016/j.ydbio.2007.05.003

Faulkner, G. J., Kimura, Y., Daub, C. O., Wani, S., Plessy, C., Irvine, K. M., Schroder, K., Cloonan, N., Steptoe, A. L., Lassmann, T., Waki, K., Hornig, N., Arakawa, T., Takahashi, H., Kawai, J., Forrest, A. R. R., Suzuki, H., Hayashizaki, Y., Hume, D. A., … Carninci, P. (2009). The regulated retrotransposon transcriptome of mammalian cells. *Nature Genetics*, *41*(5), 563–571. https://doi.org/10.1038/ng.368

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

Feschotte, C., & Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annual Review of Genetics*, *41*, 331–368. https://doi.org/10.1146/annurev.genet.40.110405.090448

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., & Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, *391*(6669), 806–811. https://doi.org/10.1038/35888

Flajnik, M. F. (2014). Re-evaluation of the immunological Big Bang. *Current Biology: CB*, *24*(21), R1060-1065. https://doi.org/10.1016/j.cub.2014.09.070

Fort, A., Hashimoto, K., Yamada, D., Salimullah, M., Keya, C. A., Saxena, A., Bonetti, A., Voineagu, I., Bertin, N., Kratz, A., Noro, Y., Wong, C.-H., de Hoon, M., Andersson, R., Sandelin, A., Suzuki, H., Wei, C.-L., Koseki, H., FANTOM Consortium, … Carninci, P. (2014). Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nature Genetics*, *46*(6), 558–566. https://doi.org/10.1038/ng.2965

Frank, J. A., & Feschotte, C. (2017). Co-option of endogenous viral sequences for host cell function. *Current Opinion in Virology*, *25*, 81–89. https://doi.org/10.1016/j.coviro.2017.07.021

Frazee, A. C., Jaffe, A. E., Langmead, B., & Leek, J. T. (2015). Polyester: Simulating RNA-seq datasets with differential transcript expression. *Bioinformatics (Oxford, England)*, *31*(17), 2778–2784. https://doi.org/10.1093/bioinformatics/btv272

Friedli, M., Turelli, P., Kapopoulou, A., Rauwel, B., Castro-Díaz, N., Rowe, H. M., Ecco, G., Unzu, C., Planet, E., Lombardo, A., Mangeat, B., Wildhaber, B. E., Naldini, L., & Trono, D. (2014). Loss of transcriptional control over endogenous retroelements during reprogramming to pluripotency. *Genome Research*, *24*(8), 1251–1259. https://doi.org/10.1101/gr.172809.114

Garcia-Perez, J. L., Widmann, T. J., & Adams, I. R. (2016). The impact of transposable elements on mammalian development. *Development (Cambridge, England)*, *143*(22), 4101–4114. https://doi.org/10.1242/dev.132639

Gerdes, P., Richardson, S. R., Mager, D. L., & Faulkner, G. J. (2016). Transposable elements in the mammalian embryo: Pioneers surviving through stealth and service. *Genome Biology*, *17*(1), 100. https://doi.org/10.1186/s13059-016-0965-5

Giraldez, A. J., Mishima, Y., Rihel, J., Grocock, R. J., Van Dongen, S., Inoue, K., Enright, A. J., & Schier, A. F. (2006). Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science (New York, N.Y.)*, *312*(5770), 75–79. https://doi.org/10.1126/science.1122689

Göke, J., Lu, X., Chan, Y.-S., Ng, H.-H., Ly, L.-H., Sachs, F., & Szczerbinska, I. (2015). Dynamic Transcription of Distinct Classes of Endogenous Retroviral Elements Marks Specific Populations of Early Human Embryonic Cells. *Cell Stem Cell*, *16*(2), 135–141. https://doi.org/10.1016/j.stem.2015.01.005

Grabundzija, I., Messing, S. A., Thomas, J., Cosby, R. L., Bilic, I., Miskey, C., Gogol-Döring, A., Kapitonov, V., Diem, T., Dalda, A., Jurka, J., Pritham, E. J., Dyda, F., Izsvák, Z., & Ivics, Z. (2016). A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. *Nature Communications*, *7*(1), 10716. https://doi.org/10.1038/ncomms10716

Gregory, T. (2005). *The Evolution of the Genome*. Elsevier. https://doi.org/10.1016/B978-0-12-301463-4.X5000-1

Grishok, A., & Mello, C. C. (2002). RNAi (Nematodes: Caenorhabditis elegans). *Advances in Genetics*, *46*, 339–360.

Grow, E. J., Flynn, R. A., Chavez, S. L., Bayless, N. L., Wossidlo, M., Wesche, D. J., Martin, L., Ware, C. B., Blish, C. A., Chang, H. Y., Pera, R. A. R., & Wysocka, J. (2015). Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature*, *522*(7555), 221–225. https://doi.org/10.1038/nature14308

Guo, C., Jeong, H.-H., Hsieh, Y.-C., Klein, H.-U., Bennett, D. A., De Jager, P. L., Liu, Z., & Shulman, J. M. (2018). Tau Activates Transposable Elements in Alzheimer's Disease. *Cell Reports*, *23*(10), 2874–2880. https://doi.org/10.1016/j.celrep.2018.05.004

Haberle, V., Forrest, A. R. R., Hayashizaki, Y., Carninci, P., & Lenhard, B. (2015). CAGEr: Precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Research*, *43*(8), e51. https://doi.org/10.1093/nar/gkv054

Hackett, J. A., Kobayashi, T., Dietmann, S., & Surani, M. A. (2017). Activation of Lineage Regulators and Transposable Elements across a Pluripotent Spectrum. *Stem Cell Reports*, *8*(6), 1645–1658. https://doi.org/10.1016/j.stemcr.2017.05.014

Hadzhiev, Y., Qureshi, H. K., Wheatley, L., Cooper, L., Jasiulewicz, A., Van Nguyen, H., Wragg, J. W., Poovathumkadavil, D., Conic, S., Bajan, S., Sik, A., Hutvàgner, G., Tora, L., Gambus, A., Fossey, J. S., & Müller, F. (2019). A cell cycle-coordinated Polymerase II transcription compartment encompasses gene expression before global genome activation. *Nature Communications*, *10*(1), 691. https://doi.org/10.1038/s41467-019-08487-5

Haeussler, M., Zweig, A. S., Tyner, C., Speir, M. L., Rosenbloom, K. R., Raney, B. J., Lee, C. M., Lee, B. T., Hinrichs, A. S., Gonzalez, J. N., Gibson, D., Diekhans, M., Clawson, H., Casper, J., Barber, G. P., Haussler, D., Kuhn, R. M., & Kent, W. J. (2019). The UCSC Genome Browser database: 2019 update. *Nucleic Acids Research*, *47*(D1), D853–D858. https://doi.org/10.1093/nar/gky1095

Hamatani, T., Carter, M. G., Sharov, A. A., & Ko, M. S. H. (2004). Dynamics of global gene expression changes during mouse preimplantation development. *Developmental Cell*, *6*(1), 117–131. https://doi.org/10.1016/s1534-5807(03)00373-3

Hamm, D. C., & Harrison, M. M. (2018). Regulatory principles governing the maternal-to-zygotic transition: Insights from Drosophila melanogaster. *Open Biology*, *8*(12), 180183. https://doi.org/10.1098/rsob.180183

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hendrickson, P. G., Doráis, J. A., Grow, E. J., Whiddon, J. L., Lim, J.-W., Wike, C. L., Weaver, B. D., Pflueger, C., Emery, B. R., Wilcox, A. L., Nix, D. A., Peterson, C. M., Tapscott, S. J., Carrell, D. T., & Cairns, B. R. (2017). Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nature Genetics*, *49*(6), 925–934. https://doi.org/10.1038/ng.3844

Heyn, P., Kircher, M., Dahl, A., Kelso, J., Tomancak, P., Kalinka, A. T., & Neugebauer, K. M. (2014). The earliest transcribed zygotic genes are short, newly evolved, and different across species. *Cell Reports*, *6*(2), 285–292. https://doi.org/10.1016/j.celrep.2013.12.030

Hobert, O. (2010). Neurogenesis in the nematode Caenorhabditis elegans. *WormBook*. https://doi.org/10.1895/wormbook.1.12.2

Http://broadinstitute.github.io/picard. (n.d.). http://broadinstitute.github.io/picard

Https://github.com/jsh58/Genrich. (n.d.).

Imbeault, M., Helleboid, P.-Y., & Trono, D. (2017). KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*, *543*(7646), 550–554. https://doi.org/10.1038/nature21683

Jachowicz, J. W., Bing, X., Pontabry, J., Bošković, A., Rando, O. J., & Torres-Padilla, M.-E. (2017). LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nature Genetics*, *49*(10), 1502–1510. https://doi.org/10.1038/ng.3945

Jeong, H.-H., Yalamanchili, H. K., Guo, C., Shulman, J. M., & Liu, Z. (2018). An ultra-fast and scalable quantification pipeline for transposable elements from next generation sequencing data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, *23*, 168–179.

Jin, Y., Tam, O. H., Paniagua, E., & Hammell, M. (2015). TEtranscripts: A package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics (Oxford, England)*, *31*(22), 3593–3599. https://doi.org/10.1093/bioinformatics/btv422

Johnson, R., & Guigó, R. (2014). The RIDL hypothesis: Transposable elements as functional domains of long noncoding RNAs. *RNA (New York, N.Y.)*, *20*(7), 959–976. https://doi.org/10.1261/rna.044560.114

Jönsson, M. E., Garza, R., Johansson, P. A., & Jakobsson, J. (2020). Transposable Elements: A Common Feature of Neurodevelopmental and Neurodegenerative Disorders. *Trends in Genetics*, *36*(8), 610–623. https://doi.org/10.1016/j.tig.2020.05.004

Jönsson, M. E., Garza, R., Sharma, Y., Petri, R., Södersten, E., Johansson, J. G., Johansson, P. A., Atacho, D. A., Pircs, K., Madsen, S., Yudovich, D., Ramakrishnan, R., Holmberg, J., Larsson, J., Jern, P., & Jakobsson, J. (2020). *Activation of endogenous retroviruses during brain development causes neuroinflammation* [Preprint]. Neuroscience. https://doi.org/10.1101/2020.07.07.191668

Jönsson, M. E., Ludvik Brattås, P., Gustafsson, C., Petri, R., Yudovich, D., Pircs, K., Verschuere, S., Madsen, S., Hansson, J., Larsson, J., Månsson, R., Meissner, A., & Jakobsson, J. (2019). Activation of neuronal genes via LINE-1 elements upon global DNA demethylation in human neural progenitors. *Nature Communications*, *10*(1), 3182. https://doi.org/10.1038/s41467-019-11150-8

Jukam, D., Shariati, S. A. M., & Skotheim, J. M. (2017). Zygotic Genome Activation in Vertebrates. *Developmental Cell*, *42*(4), 316–332. https://doi.org/10.1016/j.devcel.2017.07.026

Kapitonov, V. V., & Jurka, J. (2006). Self-synthesizing DNA transposons in eukaryotes. *Proceedings of the National Academy of Sciences*, *103*(12), 4540–4545. https://doi.org/10.1073/pnas.0600833103

Kapitonov, Vladimir V., & Jurka, J. (2005). RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biology*, *3*(6), e181. https://doi.org/10.1371/journal.pbio.0030181

Kapitonov, Vladimir V., & Koonin, E. V. (2015). Evolution of the RAG1-RAG2 locus: Both proteins came from the same transposon. *Biology Direct*, *10*, 20. https://doi.org/10.1186/s13062-015-0055-8

Kapusta, A., Kronenberg, Z., Lynch, V. J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M., & Feschotte, C. (2013). Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genetics*, *9*(4), e1003470. https://doi.org/10.1371/journal.pgen.1003470

Kapusta, A., Suh, A., & Feschotte, C. (2017). Dynamics of genome size evolution in birds and mammals. *Proceedings of the National Academy of Sciences*, *114*(8), E1460–E1469. https://doi.org/10.1073/pnas.1616702114

Kassiotis, G., & Stoye, J. P. (2016). Immune responses to endogenous retroelements: Taking the bad with the good. *Nature Reviews. Immunology*, *16*(4), 207–219. https://doi.org/10.1038/nri.2016.27

Kazazian, H. H. Jr. (2011). *Mobile DNA: finding treasure in junk*.

Keegan, R. M., Chang, Y.-H., Metzger, M. J., & Dubnau, J. (2020). *Intercellular viral spread and intracellular transposition of* Drosophila *gypsy* [Preprint]. Genetics. https://doi.org/10.1101/2020.05.28.121897

Kentepozidou, E., Aitken, S. J., Feig, C., Stefflova, K., Ibarra-Soria, X., Odom, D. T., Roller, M., & Flicek, P. (2020). Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. *Genome Biology*, *21*(1), 5. https://doi.org/10.1186/s13059-019-1894-x

Kim, A., Terzian, C., Santamaria, P., Pélisson, A., Purd'homme, N., & Bucheton, A. (1994). Retroviruses in invertebrates: The gypsy retrotransposon is apparently an infectious retrovirus of Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(4), 1285–1289. https://doi.org/10.1073/pnas.91.4.1285

Klenov, M. S., Lavrov, S. A., Korbut, A. P., Stolyarenko, A. D., Yakushev, E. Y., Reuter, M., Pillai, R. S., & Gvozdev, V. A. (2014). Impact of nuclear Piwi elimination on chromatin state in Drosophila melanogaster ovaries. *Nucleic Acids Research*, *42*(10), 6208–6218. https://doi.org/10.1093/nar/gku268

Kramerov, D. A., & Vassetzky, N. S. (2011). Origin and evolution of SINEs in eukaryotic genomes. *Heredity*, *107*(6), 487–495. https://doi.org/10.1038/hdy.2011.43

Krug, L., Chatterjee, N., Borges-Monroy, R., Hearn, S., Liao, W.-W., Morrill, K., Prazak, L., Rozhkov, N., Theodorou, D., Hammell, M., & Dubnau, J. (2017). Retrotransposon activation contributes to neurodegeneration in a Drosophila TDP-43 model of ALS. *PLOS Genetics*, *13*(3), e1006635. https://doi.org/10.1371/journal.pgen.1006635

Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., Ng, H.-H., & Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genetics*, *42*(7), 631–634. https://doi.org/10.1038/ng.600

Lanciano, S., & Cristofari, G. (2020). Measuring and interpreting transposable element expression. *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-020-0251-y

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., … International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25. https://doi.org/10.1186/gb-2009-10-3-r25

Laricchia, K. M., Zdraljevic, S., Cook, D. E., & Andersen, E. C. (2017). Natural Variation in the Distribution and Abundance of Transposable Elements Across the Caenorhabditis elegans Species. *Molecular Biology and Evolution*, *34*(9), 2187–2202. https://doi.org/10.1093/molbev/msx155

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., & Carey, V. J. (2013). Software for Computing and Annotating Genomic Ranges.

PLoS Computational Biology, *9*(8), e1003118. https://doi.org/10.1371/journal.pcbi.1003118

Lécuyer, E., Yoshida, H., Parthasarathy, N., Alm, C., Babak, T., Cerovina, T., Hughes, T. R., Tomancak, P., & Krause, H. M. (2007). Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell*, *131*(1), 174–187. https://doi.org/10.1016/j.cell.2007.08.003

Lee, M. T., Bonneau, A. R., & Giraldez, A. J. (2014). Zygotic genome activation during the maternal-to-zygotic transition. *Annual Review of Cell and Developmental Biology*, *30*, 581–613. https://doi.org/10.1146/annurev-cellbio-100913-013027

Lee, M. T., Bonneau, A. R., Takacs, C. M., Bazzini, A. A., DiVito, K. R., Fleming, E. S., & Giraldez, A. J. (2013). Nanog, Pou5f1 and SoxB1 activate zygotic gene expression during the maternal-to-zygotic transition. *Nature*, *503*(7476), 360–364. https://doi.org/10.1038/nature12632

Leichsenring, M., Maes, J., Mössner, R., Driever, W., & Onichtchouk, D. (2013). Pou5f1 transcription factor controls zygotic gene activation in vertebrates. *Science (New York, N.Y.)*, *341*(6149), 1005–1009. https://doi.org/10.1126/science.1242527

Lev-Maor, G., Ram, O., Kim, E., Sela, N., Goren, A., Levanon, E. Y., & Ast, G. (2008). Intronic Alus influence alternative splicing. *PLoS Genetics*, *4*(9), e1000204. https://doi.org/10.1371/journal.pgen.1000204

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, L., Lu, X., & Dean, J. (2013). The maternal to zygotic transition in mammals. *Molecular Aspects of Medicine*, *34*(5), 919–938. https://doi.org/10.1016/j.mam.2013.01.003

Li, W., Jin, Y., Prazak, L., Hammell, M., & Dubnau, J. (2012). Transposable Elements in TDP-43-Mediated Neurodegenerative Disorders. *PLoS ONE*, *7*(9), e44099. https://doi.org/10.1371/journal.pone.0044099

Liang, H.-L., Nien, C.-Y., Liu, H.-Y., Metzstein, M. M., Kirov, N., & Rushlow, C. (2008). The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. *Nature*, *456*(7220), 400–403. https://doi.org/10.1038/nature07388

Litman, G. W., Rast, J. P., & Fugmann, S. D. (2010). The origins of vertebrate adaptive immunity. *Nature Reviews. Immunology*, *10*(8), 543–553. https://doi.org/10.1038/nri2807

Liu, Y., Zhu, Z., Ho, I. H. T., Shi, Y., Li, J., Wang, X., Chan, M. T. V., & Cheng, C. H. K. (2020). Genetic Deletion of miR-430 Disrupts Maternal-Zygotic Transition and Embryonic Body Plan. *Frontiers in Genetics*, *11*, 853. https://doi.org/10.3389/fgene.2020.00853

Lu, X., Sachs, F., Ramsay, L., Jacques, P.-É., Göke, J., Bourque, G., & Ng, H.-H. (2014). The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nature Structural & Molecular Biology*, *21*(4), 423–425. https://doi.org/10.1038/nsmb.2799

Lund, E., Liu, M., Hartley, R. S., Sheets, M. D., & Dahlberg, J. E. (2009). Deadenylation of maternal mRNAs mediated by miR-427 in Xenopus laevis embryos. *RNA (New York, N.Y.)*, *15*(12), 2351–2363. https://doi.org/10.1261/rna.1882009

Macfarlan, T. S., Gifford, W. D., Driscoll, S., Lettieri, K., Rowe, H. M., Bonanomi, D., Firth, A., Singer, O., Trono, D., & Pfaff, S. L. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*, *487*(7405), 57–63. https://doi.org/10.1038/nature11244

Maduro, M. F. (2010). Cell fate specification in the *C. elegans* embryo. *Developmental Dynamics*, 1315–1329. https://doi.org/10.1002/dvdy.22233

Maksakova, I. A., Romanish, M. T., Gagnier, L., Dunn, C. A., van de Lagemaat, L. N., & Mager, D. L. (2006). Retroviral Elements and Their Hosts: Insertional Mutagenesis in the Mouse Germ Line. *PLoS Genetics*, *2*(1), e2. https://doi.org/10.1371/journal.pgen.0020002

Martin, E. C., Vicari, C., Tsakou-Ngouafo, L., Pontarotti, P., Petrescu, A. J., & Schatz, D. G. (2020). Identification of RAG-like transposons in protostomes suggests their ancient bilaterian origin. *Mobile DNA*, *11*(1), 17. https://doi.org/10.1186/s13100-020-00214-y

Mätlik, K., Redik, K., & Speek, M. (2006). L1 antisense promoter drives tissue-specific transcription of human genes. *Journal of Biomedicine & Biotechnology*, *2006*(1), 71753. https://doi.org/10.1155/JBB/2006/71753

Maupetit-Mehouas, S., & Vaury, C. (2020). Transposon Reactivation in the Germline May Be Useful for Both Transposons and Their Host Genomes. *Cells*, *9*(5). https://doi.org/10.3390/cells9051172

McClintock, B. (1956). Intranuclear systems controlling gene action and mutation. *Brookhaven Symposia in Biology*, *8*, 58–74.

McCullers, T. J., & Steiniger, M. (2017). Transposable elements in *Drosophila*. *Mobile Genetic Elements*, *7*(3), 1–18. https://doi.org/10.1080/2159256X.2017.1318201

McGhee, J. (2007). The C. elegans intestine. *WormBook*. https://doi.org/10.1895/wormbook.1.133.1

McGrath, J., & Solter, D. (1984). Completion of mouse embryogenesis requires both the maternal and paternal genomes. *Cell*, *37*(1), 179–183. https://doi.org/10.1016/0092-8674(84)90313-1

McKerrow, W., & Fenyö, D. (2019). L1EM: A tool for accurate locus specific LINE-1 RNA quantification. *Bioinformatics*, btz724. https://doi.org/10.1093/bioinformatics/btz724

Mills, R. E., Bennett, E. A., Iskow, R. C., & Devine, S. E. (2007). Which transposable elements are active in the human genome? *Trends in Genetics*, *23*(4), 183–191. https://doi.org/10.1016/j.tig.2007.02.006

Mouse Genome Sequencing Consortium, Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., … Lander, E. S. (2002). Initial sequencing and comparative

analysis of the mouse genome. *Nature*, *420*(6915), 520–562. https://doi.org/10.1038/nature01262

Muñoz-López, M., & García-Pérez, J. L. (2010). DNA transposons: Nature and applications in genomics. *Current Genomics*, *11*(2), 115–128. https://doi.org/10.2174/138920210790886871

Muotri, A. R., Chu, V. T., Marchetto, M. C. N., Deng, W., Moran, J. V., & Gage, F. H. (2005). Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature*, *435*(7044), 903–910. https://doi.org/10.1038/nature03663

Muotri, A. R., Zhao, C., Marchetto, M. C. N., & Gage, F. H. (2009). Environmental Influence on L1 Retrotransposons in the Adult Hippocampus. *Hippocampus*, *19*(10), 1002–1007. https://doi.org/10.1002/hipo.20564

Nance, J., Lee, J.-Y., & Goldstein, B. (2005). Gastrulation in C. elegans. *WormBook*. https://doi.org/10.1895/wormbook.1.23.1

Navarro, F. C., Hoops, J., Bellfy, L., Cerveira, E., Zhu, Q., Zhang, C., Lee, C., & Gerstein, M. B. (2019). TeXP: Deconvolving the effects of pervasive and autonomous transcription of transposable elements. *PLOS Computational Biology*, *15*(8), e1007293. https://doi.org/10.1371/journal.pcbi.1007293

Naville, M., Warren, I. A., Haftek-Terreau, Z., Chalopin, D., Brunet, F., Levin, P., Galiana, D., & Volff, J.-N. (2016). Not so bad after all: Retroviruses and long terminal repeat retrotransposons as a source of new genes in vertebrates. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*, *22*(4), 312–323. https://doi.org/10.1016/j.cmi.2016.02.001

Nigumann, P., Redik, K., Mätlik, K., & Speek, M. (2002). Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics*, *79*(5), 628–634. https://doi.org/10.1006/geno.2002.6758

Orecchini, E., Frassinelli, L., & Michienzi, A. (2017). Restricting retrotransposons: ADAR1 is another guardian of the human genome. *RNA Biology*, *14*(11), 1485–1491. https://doi.org/10.1080/15476286.2017.1341033

Osborne Nishimura, E., Zhang, J. C., Werts, A. D., Goldstein, B., & Lieb, J. D. (2015). Asymmetric transcript discovery by RNA-seq in C. elegans blastomeres identifies neg-1, a gene important for anterior morphogenesis. *PLoS Genetics*, *11*(4), e1005117. https://doi.org/10.1371/journal.pgen.1005117

Ozata, D. M., Gainetdinov, I., Zoch, A., O'Carroll, D., & Zamore, P. D. (2019). PIWI-interacting RNAs: Small RNAs with big functions. *Nature Reviews Genetics*, *20*(2), 89–108. https://doi.org/10.1038/s41576-018-0073-3

Pálfy, M., Joseph, S. R., & Vastenhouw, N. L. (2017). The timing of zygotic genome activation. *Current Opinion in Genetics & Development*, *43*, 53–60. https://doi.org/10.1016/j.gde.2016.12.001

Park, C.-E., Shin, M.-R., Jeon, E.-H., Lee, S.-H., Cha, K.-Y., Kim, K., Kim, N.-H., & Lee, K.-A. (2004). Oocyte-selective expression of MT transposon-like element, clone MTi7 and its role in oocyte maturation and embryo development. *Molecular Reproduction and Development*, *69*(4), 365–374. https://doi.org/10.1002/mrd.20179

Parkhurst, S. M., & Corces, V. G. (1987). Developmental expression of Drosophila melanogaster retrovirus-like transposable elements. *The EMBO Journal*, *6*(2), 419–424. https://doi.org/10.1002/j.1460-2075.1987.tb04771.x

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

Peaston, A. E., Evsikov, A. V., Graber, J. H., de Vries, W. N., Holbrook, A. E., Solter, D., & Knowles, B. B. (2004). Retrotransposons Regulate Host Genes in Mouse Oocytes and Preimplantation Embryos. *Developmental Cell*, *7*(4), 597–606. https://doi.org/10.1016/j.devcel.2004.09.004

Percharde, M., Lin, C.-J., Yin, Y., Guan, J., Peixoto, G. A., Bulut-Karslioglu, A., Biechele, S., Huang, B., Shen, X., & Ramalho-Santos, M. (2018). A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. *Cell*, *174*(2), 391-405.e19. https://doi.org/10.1016/j.cell.2018.05.043

Perrat, P. N., DasGupta, S., Wang, J., Theurkauf, W., Weng, Z., Rosbash, M., & Waddell, S. (2013). Transposition-Driven Genomic Heterogeneity in the Drosophila Brain. *Science*, *340*(6128), 91–95. https://doi.org/10.1126/science.1231965

Petrov, D. A. (2002). Mutational equilibrium model of genome size evolution. *Theoretical Population Biology*, *61*(4), 531–544. https://doi.org/10.1006/tpbi.2002.1605

Piacentini, L., Fanti, L., Specchia, V., Bozzetti, M. P., Berloco, M., Palumbo, G., & Pimpinelli, S. (2014). Transposons, environmental changes, and heritable induced phenotypic variability. *Chromosoma*, *123*(4), 345–354. https://doi.org/10.1007/s00412-014-0464-y

Pikó, L., & Clegg, K. B. (1982). Quantitative changes in total RNA, total poly(A), and ribosomes in early mouse embryos. *Developmental Biology*, *89*(2), 362–378. https://doi.org/10.1016/0012-1606(82)90325-6

Pontis, J., Planet, E., Offner, S., Turelli, P., Duc, J., Coudray, A., Theunissen, T. W., Jaenisch, R., & Trono, D. (2019). Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. *Cell Stem Cell*, *24*(5), 724-735.e5. https://doi.org/10.1016/j.stem.2019.03.012

Prudencio, M., Gonzales, P. K., Cook, C. N., Gendron, T. F., Daughrity, L. M., Song, Y., Ebbert, M. T. W., van Blitterswijk, M., Zhang, Y.-J., Jansen-West, K., Baker, M. C., DeTure, M., Rademakers, R., Boylan, K. B., Dickson, D. W., Petrucelli, L., & Link, C. D. (2017). Repetitive element transcripts are elevated in the brain of C9orf72 ALS/FTLD patients. *Human Molecular Genetics*, *26*(17), 3421–3431. https://doi.org/10.1093/hmg/ddx233

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

R Core Team. (2018). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.*

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., & Manke, T. (2016). deepTools2: A next generation web server for

deep-sequencing data analysis. *Nucleic Acids Research*, *44*(W1), W160-165. https://doi.org/10.1093/nar/gkw257

Richardson, S. R., Morell, S., & Faulkner, G. J. (2014). L1 retrotransposons and somatic mosaicism in the brain. *Annual Review of Genetics*, *48*, 1–27. https://doi.org/10.1146/annurev-genet-120213-092412

Robertson, S., & Lin, R. (2015). The Maternal-to-Zygotic Transition in C. elegans. In *Current Topics in Developmental Biology* (Vol. 113, pp. 1–42). Elsevier. https://doi.org/10.1016/bs.ctdb.2015.06.001

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Rodriguez-Terrones, D., & Torres-Padilla, M.-E. (2018). Nimble and Ready to Mingle: Transposon Outbursts of Early Development. *Trends in Genetics: TIG*. https://doi.org/10.1016/j.tig.2018.06.006

Rossum, G. van, & Drake, F. (2001). *Pyhton Reference Manual*.

Russell, S. J., Stalker, L., & LaMarre, J. (2017). PIWIs, piRNAs and Retrotransposons: Complex battles during reprogramming in gametes and early embryos. *Reproduction in Domestic Animals = Zuchthygiene*, *52 Suppl 4*, 28–38. https://doi.org/10.1111/rda.13053

Saleh, A., Macia, A., & Muotri, A. R. (2019). Transposable Elements, Inflammation, and Neurological Disease. *Frontiers in Neurology*, *10*, 894. https://doi.org/10.3389/fneur.2019.00894

Santoni, F. A., Guerra, J., & Luban, J. (2012). HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology*, *9*, 111. https://doi.org/10.1186/1742-4690-9-111

Schmitz, J., & Brosius, J. (2011). Exonization of transposed elements: A challenge and opportunity for evolution. *Biochimie*, *93*(11), 1928–1934. https://doi.org/10.1016/j.biochi.2011.07.014

Schubert, I., & Vu, G. T. H. (2016). Genome Stability and Evolution: Attempting a Holistic View. *Trends in Plant Science*, *21*(9), 749–757. https://doi.org/10.1016/j.tplants.2016.06.003

Schulz, K. N., & Harrison, M. M. (2019). Mechanisms regulating zygotic genome activation. *Nature Reviews. Genetics*, *20*(4), 221–234. https://doi.org/10.1038/s41576-018-0087-x

SciPy 1.0 Contributors, Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., … van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Sha, Q.-Q., Zhu, Y.-Z., Li, S., Jiang, Y., Chen, L., Sun, X.-H., Shen, L., Ou, X.-H., & Fan, H.-Y. (2020). Characterization of zygotic genome activation-dependent maternal mRNA

clearance in mouse. *Nucleic Acids Research*, *48*(2), 879–894. https://doi.org/10.1093/nar/gkz1111

Sorek, R. (2007). The birth of new exons: Mechanisms and evolutionary consequences. *RNA (New York, N.Y.)*, *13*(10), 1603–1608. https://doi.org/10.1261/rna.682507

Sripathy, S. P., Stevens, J., & Schultz, D. C. (2006). The KAP1 corepressor functions to coordinate the assembly of de novo HP1-demarcated microenvironments of heterochromatin required for KRAB zinc finger protein-mediated transcriptional repression. *Molecular and Cellular Biology*, *26*(22), 8623–8638. https://doi.org/10.1128/MCB.00487-06

Stitzel, M. L., & Seydoux, G. (2007). Regulation of the oocyte-to-zygote transition. *Science (New York, N.Y.)*, *316*(5823), 407–408. https://doi.org/10.1126/science.1138236

Subramanian, R. P., Wildschutte, J. H., Russo, C., & Coffin, J. M. (2011). Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology*, *8*, 90. https://doi.org/10.1186/1742-4690-8-90

Sulston, J. E., Schierenberg, E., White, J. G., & Thomson, J. N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. *Developmental Biology*, *100*(1), 64–119.

Sun, W., Samimi, H., Gamez, M., Zare, H., & Frost, B. (2018). Pathogenic tau-induced piRNA depletion promotes neuronal death through transposable element dysregulation in neurodegenerative tauopathies. *Nature Neuroscience*, *21*(8), 1038–1048. https://doi.org/10.1038/s41593-018-0194-1

Sundaram, V., Choudhary, M. N. K., Pehrsson, E., Xing, X., Fiore, C., Pandey, M., Maricque, B., Udawatta, M., Ngo, D., Chen, Y., Paguntalan, A., Ray, T., Hughes, A., Cohen, B. A., & Wang, T. (2017). Functional cis-regulatory modules encoded by mouse-specific endogenous retrovirus. *Nature Communications*, *8*(1), 14550. https://doi.org/10.1038/ncomms14550

Sundaram, V., & Wysocka, J. (2020). Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *375*(1795), 20190347. https://doi.org/10.1098/rstb.2019.0347

Tabara, H., Sarkissian, M., Kelly, W. G., Fleenor, J., Grishok, A., Timmons, L., Fire, A., & Mello, C. C. (1999). The rde-1 gene, RNA interference, and transposon silencing in C. elegans. *Cell*, *99*(2), 123–132.

Tadros, W., Goldman, A. L., Babak, T., Menzies, F., Vardy, L., Orr-Weaver, T., Hughes, T. R., Westwood, J. T., Smibert, C. A., & Lipshitz, H. D. (2007). SMAUG is a major regulator of maternal mRNA destabilization in Drosophila and its translation is activated by the PAN GU kinase. *Developmental Cell*, *12*(1), 143–155. https://doi.org/10.1016/j.devcel.2006.10.005

Tadros, W., & Lipshitz, H. D. (2009). The maternal-to-zygotic transition: A play in two acts. *Development (Cambridge, England)*, *136*(18), 3033–3042. https://doi.org/10.1242/dev.033183

Tan, H., Wu, C., & Jin, L. (2018). A Possible Role for Long Interspersed Nuclear Elements-1 (LINE-1) in Huntington's Disease Progression. *Medical Science Monitor*, *24*, 3644–3652. https://doi.org/10.12659/MSM.907328

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., & Narechania, A. (2003). PANTHER: A library of protein families and subfamilies indexed by function. *Genome Research*, *13*(9), 2129–2141. https://doi.org/10.1101/gr.772403

Thomsen, S., Anders, S., Janga, S. C., Huber, W., & Alonso, C. R. (2010). Genome-wide analysis of mRNA decay patterns during early Drosophila development. *Genome Biology*, *11*(9), R93. https://doi.org/10.1186/gb-2010-11-9-r93

Tintori, S. C., Osborne Nishimura, E., Golden, P., Lieb, J. D., & Goldstein, B. (2016). A Transcriptional Lineage of the Early C. elegans Embryo. *Developmental Cell*, *38*(4), 430–444. https://doi.org/10.1016/j.devcel.2016.07.025

Tokuyama, M., Kong, Y., Song, E., Jayewickreme, T., Kang, I., & Iwasaki, A. (2018). ERVmap analysis reveals genome-wide transcription of human endogenous retroviruses. *Proceedings of the National Academy of Sciences*, *115*(50), 12565–12572. https://doi.org/10.1073/pnas.1814589115

Torres-Padilla, M.-E. (2020). On transposons and totipotency. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1795), 20190339. https://doi.org/10.1098/rstb.2019.0339

Ulitsky, I., Shkumatava, A., Jan, C. H., Subtelny, A. O., Koppstein, D., Bell, G. W., Sive, H., & Bartel, D. P. (2012). Extensive alternative polyadenylation during zebrafish development. *Genome Research*, *22*(10), 2054–2066. https://doi.org/10.1101/gr.139733.112

Upton, K. R., Gerhardt, D. J., Jesuadian, J. S., Richardson, S. R., Sánchez-Luque, F. J., Bodea, G. O., Ewing, A. D., Salvador-Palomeque, C., van der Knaap, M. S., Brennan, P. M., Vanderver, A., & Faulkner, G. J. (2015). Ubiquitous L1 Mosaicism in Hippocampal Neurons. *Cell*, *161*(2), 228–239. https://doi.org/10.1016/j.cell.2015.03.026

Volpe, M., Miralto, M., Gustincich, S., & Sanges, R. (2018). ClusterScan: Simple and generalistic identification of genomic clusters. *Bioinformatics (Oxford, England)*, *34*(22), 3921–3923. https://doi.org/10.1093/bioinformatics/bty486

Walser, C. B., & Lipshitz, H. D. (2011). Transcript clearance during the maternal-to-zygotic transition. *Current Opinion in Genetics & Development*, *21*(4), 431–443. https://doi.org/10.1016/j.gde.2011.03.003

Walsh, C. P., Chaillet, J. R., & Bestor, T. H. (1998). Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nature Genetics*, *20*(2), 116–117. https://doi.org/10.1038/2413

Wang, J., Xie, G., Singh, M., Ghanbarian, A. T., Raskó, T., Szvetnik, A., Cai, H., Besser, D., Prigione, A., Fuchs, N. V., Schumann, G. G., Chen, W., Lorincz, M. C., Ivics, Z., Hurst, L. D., & Izsvák, Z. (2014). Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature*, *516*(7531), 405–409. https://doi.org/10.1038/nature13804

Wang, Q. T., Piotrowska, K., Ciemerych, M. A., Milenkovic, L., Scott, M. P., Davis, R. W., & Zernicka-Goetz, M. (2004). A genome-wide study of gene activity reveals

developmental signaling pathways in the preimplantation mouse embryo. *Developmental Cell*, *6*(1), 133–144. https://doi.org/10.1016/s1534-5807(03)00404-0

White, R. J., Collins, J. E., Sealy, I. M., Wali, N., Dooley, C. M., Digby, Z., Stemple, D. L., Murphy, D. N., Billis, K., Hourlier, T., Füllgrabe, A., Davis, M. P., Enright, A. J., & Busch-Nentwich, E. M. (2017). A high-resolution mRNA expression time course of embryonic development in zebrafish. *ELife*, *6*, e30860. https://doi.org/10.7554/eLife.30860

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews. Genetics*, *8*(12), 973–982. https://doi.org/10.1038/nrg2165

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*.

Wolf, G., de Iaco, A., Sun, M.-A., Bruno, M., Tinkham, M., Hoang, D., Mitra, A., Ralls, S., Trono, D., & Macfarlan, T. S. (2020). KRAB-zinc finger protein gene expansion in response to active retrotransposons in the murine lineage. *ELife*, *9*, e56337. https://doi.org/10.7554/eLife.56337

Wragg, J., & Müller, F. (2016). Transcriptional Regulation During Zygotic Genome Activation in Zebrafish and Other Anamniote Embryos. *Advances in Genetics*, *95*, 161–194. https://doi.org/10.1016/bs.adgen.2016.05.001

Wu, J., Huang, B., Chen, H., Yin, Q., Liu, Y., Xiang, Y., Zhang, B., Liu, B., Wang, Q., Xia, W., Li, W., Li, Y., Ma, J., Peng, X., Zheng, H., Ming, J., Zhang, W., Zhang, J., Tian, G., … Xie, W. (2016). The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature*, *534*(7609), 652–657. https://doi.org/10.1038/nature18606

Wu, Y., Liu, W., Chen, J., Liu, S., Wang, M., Yang, L., Chen, C., Qi, M., Xu, Y., Qiao, Z., Yan, R., Kou, X., Zhao, Y., Shen, B., Yin, J., Wang, H., Gao, Y., & Gao, S. (2019). Nuclear Exosome Targeting Complex Core Factor Zcchc8 Regulates the Degradation of LINE1 RNA in Early Embryos and Embryonic Stem Cells. *Cell Reports*, *29*(8), 2461-2472.e6. https://doi.org/10.1016/j.celrep.2019.10.055

Yang, W. R., Ardeljan, D., Pacyna, C. N., Payer, L. M., & Burns, K. H. (2019). SQuIRE reveals locus-specific regulation of interspersed repeat expression. *Nucleic Acids Research*, *47*(5), e27. https://doi.org/10.1093/nar/gky1301

Yu, G., Wang, L.-G., & He, Q.-Y. (2015). ChIPseeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics (Oxford, England)*, *31*(14), 2382–2383. https://doi.org/10.1093/bioinformatics/btv145

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., To, J. K., … Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, *46*(D1), D754–D761. https://doi.org/10.1093/nar/gkx1098

Zhang, B., Zheng, H., Huang, B., Li, W., Xiang, Y., Peng, X., Ming, J., Wu, X., Zhang, Y., Xu, Q., Liu, W., Kou, X., Zhao, Y., He, W., Li, C., Chen, B., Li, Y., Wang, Q., Ma, J., … Xie, W. (2016). Allelic reprogramming of the histone modification H3K4me3 in early

mammalian development. *Nature*, *537*(7621), 553–557. https://doi.org/10.1038/nature19361

Zhang, W., Chen, F., Chen, R., Xie, D., Yang, J., Zhao, X., Guo, R., Zhang, Y., Shen, Y., Göke, J., Liu, L., & Lu, X. (2019). Zscan4c activates endogenous retrovirus MERVL and cleavage embryo genes. *Nucleic Acids Research*, *47*(16), 8485–8501. https://doi.org/10.1093/nar/gkz594

Zhang, Y.-J., Guo, L., Gonzales, P. K., Gendron, T. F., Wu, Y., Jansen-West, K., O'Raw, A. D., Pickles, S. R., Prudencio, M., Carlomagno, Y., Gachechiladze, M. A., Ludwig, C., Tian, R., Chew, J., DeTure, M., Lin, W.-L., Tong, J., Daughrity, L. M., Yue, M., … Petrucelli, L. (2019). Heterochromatin anomalies and double-stranded RNA accumulation underlie C9orf72 poly(PR) toxicity. *Science (New York, N.Y.)*, *363*(6428). https://doi.org/10.1126/science.aav2606

Zhou, L., & Dean, J. (2015). Reprogramming the genome to totipotency in mouse embryos. *Trends in Cell Biology*, *25*(2), 82–91. https://doi.org/10.1016/j.tcb.2014.09.006