THE UNIVERSITY
of EDINBURGH

Activity and regulation of vertebrate LINE-1 retrotransposons

Marie-Jeanne Hubertine Caroline Kempen

THE UNIVERSITY *of* EDINBURGH

PhD

Genetics and Molecular Medicine

The University of Edinburgh

2021

Declaration

This thesis has been composed of original research and analysis undertaken entirely by me. Where the work of others is included, their contributions have been duly acknowledged. The work in this thesis has not been submitted for any other degree or professional qualification.

Marie-Jeanne Hubertine Caroline Kempen

November 2021

Abstract

Transposable elements (TEs) are repetitive DNA sequences that can mobilise in our genome by a process called transposition. Based on their mechanism of transposition, TEs can be subdivided into transposons and retrotransposons, which use DNA and RNA intermediates respectively. They are found in all branches of life and have been highly successful at colonising genomes along evolution, constituting up to 50% of the human genome. Although their profound influence on evolution is unquestionable, their function and influence within the organism's biology is still largely unknown.

In humans, only retrotransposons are currently active, of which exclusively the Long INterspersed Element 1 (LINE-1 or L1) can mobilise autonomously. Although there are many cellular mechanisms in place to control LINE-1, their activity has been observed during embryogenesis, in neuronal cells, and in pathological conditions (i.e. cancer). In several cases, their insertional mutagenesis has been identified as the direct cause of genetic disorders and cancers, as well as a major contributor to disease progression by driving genetic instability. Notably, a growing body of evidence suggests that L1 encoded proteins and L1 retrotransposition intermediates can also influence cellular functions, such as inducing inflammation and senescence. Understanding the mechanisms underlying L1 regulation can help us prevent their deleterious effects and improve the prognosis of patients suffering from disorders where their activity is deregulated.

Our lab has recently identified Ribonuclease H2 (RNase H2), a protein frequently mutated in Aicardi-Goutières Syndrome (AGS), as a positive regulator of L1 retrotransposition. It was proposed that the interaction of both RNase H2 and L1-encoded ORF2p protein with Proliferating cell nuclear antigen (PCNA , a DNA clamp essential for DNA replication) through their PIP (PCNA Interaction Protein) motifs, lies at the basis of their mutual coordination and allows RNase H2 to fulfil its function in L1 retrotransposition. However, further research is needed to prove this model. Also, whether changes in LINE-1 retrotransposition levels contribute to the symptomology of AGS and other disorders where L1s are found to be deregulated remains unconfirmed. An easy to manipulate animal model, such as zebrafish, would be instrumental for research into these types of questions. Zebrafish are genetically very accessible and permit many research technologies unavailable for murine model. Despite the

presence of polymorphic L1 insertions suggesting recent mobilization, to date no active L1 copy has been characterized in the zebrafish genome.

This thesis has two distinct aims. Firstly, to elucidate the role of the PIP domain in RNase H2 and L1-ORF2p during retrotransposition. My research into this question revealed that the PIP motif of RNase H2 does not mediate the processes that underlie its function in human LINE-1 retrotransposition. Additionally, I found that very low levels of RNase H2 (<15% of WT) are sufficient to support WT levels of LINE-1 retrotransposition. The second aim was to identify active LINE-1 copies in the zebrafish genome, in order to validate this animal model for future *in vivo* LINE-1 research. The specific copies from the different LINE-1 families investigated in this study did not show measurable levels of retrotransposition in our experimental setting. Nonetheless, this work cannot exclude zebrafish as a potential tool for research into the biology of LINE-1 and other TEs. Additionally, a potentially valuable LINE-1 reporter system was designed for this work, allowing the assessment of translation and retrotransposition separately. In conclusion, this work contributes to our understanding of LINE-1 biology, and can be used as a stepping stone for further research into the role of RNase H2 and PCNA in LINE-1 retrotransposition, as well the exploration into whether zebrafish could be used as an animal model for TE research.

Lay summary

The blueprint of any living entity is encoded in its DNA, the molecule that contains the genetic information shared by all the cells within an organism. However, throughout lifetime, mutations occur in the DNA, altering its message and contributing to various genetic diseases. One source of mutations are Transposable Elements (TEs) or 'jumping genes', which are extremely abundant sequences in our DNA. These elements can move from one place in the DNA to another, thus causing mutations. Evolution has provided mechanisms to keep TEs in check, although they can still 'jump' in early human embryonic development and under pathological conditions. Understanding the factors that activate TEs and how cells harness their activity can ultimately improve prognosis/treatment of human diseases involving TEs.

Our lab recently identified the cellular factor RNase H2 as an important regulator of a group of active TEs in the human genome: Long Interspersed Element Class 1 retrotransposons (LINE-1s). Notably, RNase H2 is commonly mutated in a genetic disorder Aicardi-Goutières Syndrome (AGS; characterised by brain inflammation in infants). How RNase H2 interacts with LINE-1, its role in mobilisation and its contribution to the symptomology observed in AGS patients is currently unknown. Animal models represent a valuable resource to answer these questions, allowing researchers a window into pre-symptomatic stages of a disorder, helping them to shed light on the effects across tissues and organs, and detangle disease cause and course.

The first aim of my thesis was to assess whether a particular domain in LINE-1 is involved in the interaction with RNase H2, and what role it has in the mobilisation of new LINE-1 copies. By using different cellular models and techniques, I found that, although both this domain and RNase H2 are essential for LINE-1 mobilisation, their roles are distinct from each other. The second aim of my thesis explored zebrafish as a potential animal model for LINE-1 research by attempting to identify active LINE-1 copies in the zebrafish DNA. However, in our experimental approach, none of the zebrafish LINE-1 copies analysed here showed a measurable level of activity. The results in this thesis do not exclude zebrafish as a potential animal model for LINE-1 research, as further investigation is needed. This work represents a stepping stone for further research into the mechanisms underlying the influence of RNase H2

on LINE-1 mobilisation, as well as the potential of zebrafish as an animal model for investigating the importance of TEs in biomedicine.

I dedicate this thesis to my mother.

She always led by example, teaching me to keep going no matter what life throws at you.

If I can be even half the woman you were, I will be proud.

Acknowledgments

There are many people to thank for helping me make this PhD thesis a reality. No one can do it alone, and I have been very lucky to have had many people who have supported me through this challenging process. Here I would like to highlight the most essential players.

I would like to thank my supervisor José-Luis Garcia-Perez for giving me the opportunity to commence this thesis. The other members of our small, but significant lab, Emily, Emma and Kiko have been significant factors in keeping me going, through moral boosts, technical support and scientific advice. I couldn't have done it without you. Also special thanks to Alejandro for all the bioinformatics support and dealing with my naïve questions.

Furthermore, I would like to thank all the members of the Andrew Jackson lab, for practically adopting me as one of your own. Besides the fact that you have helped create a lovely work atmosphere, you were always there to answer my many questions, give advice and even technical support without ever making me feel like I was bothering you. I want to specifically highlight Olga and Martin who have been crucial in the completion of several parts of this PhD thesis. Thank you both so much. And thank you to Andrea for keeping the lab organised and ensuring I always had the ridiculous amount of tissue culture plates I needed. Also a big thank you to all the members of the Liz Patton for helping and advising me on the zebrafish related part of my PhD. Having had the opportunity to join your lab meetings has been very stimulating.

I would also like to express my gratitude to all the technical staff at the MRC Human Genetics Unit for their great job. Specifically Lizzie for her eternal patience with my tardiness, fun chats and always friendly support; and Cameron for being an excellent fish 'guru', keeping things interesting with funny emails and facilitating endless philosophical discussions (Neuroscience!!).

The best part of the PhD has been the new friends I made in the process. I am so grateful to all of you. Thank you to my office mates Emily, Emma, Brianda and Nikki for making even ordinary workdays memorable with cakes and laughs, as well as being my friends outside of work. Thank you to all the others, particularly Blanca, Carol-Anne, Kiko, Lukas, Michael and Sam for everything from regular climbing sessions and dinners, to exciting adventures in and

around the highlands. Special thank you to my best friend Paula for all your kindness and general awesomeness. Without you, I would have lost my sanity years ago.

I would also like to acknowledge my family for their general encouragement and support. Heel veel dank aan mijn vader, die er voor heeft gezorgd dat ik alles had om zorgeloos mijn dromen te volgen; mijn moeder, die altijd in me geloofde en me aanmoedigde; en mijn twee lieve broers die me hebben gesteund, ieder op zijn eigen manier. Mijn bonus gezinnen, Marijke, papa Jean, Merel en Gode, en ome Harry en tante Maria, die me een extra portie liefde hebben gegeven in mijn leven; en oom Raymond die altijd klaar stond om me een beetje te verwennen en de levensgenieter in me te stimuleren. Extra dank aan mijn tante Marie-José, die altijd voor me klaar stond en zelfs mijn hele thesis heeft proef gelezen. Jou steun en correcties waren onmisbaar. Tenslotte, heel erg veel dank aan mijn nicht(je) Esther, voor al je steun en liefde, en me altijd te komen opzoeken waar ik ook ga. Meer dan een nicht ben je een van mijn beste vriendinnen.

Finally, Kiko, my partner and best friend, you have been vital in all aspects of this thesis. Thank you for all your explanations, scientific advice and technical support, as well as keeping me sane in the writing process (which I know was a real challenge). But most importantly thank you for being the amazing person you are in all aspects of our life. Wherever you are, I am home.

Table of Contents

List of figures

List of tables

Abbreviations

A	Adenine
Act-β	beta-Actin
ADAR1	Double-stranded RNA-specific Adenosine DeAminase 1
AGS	Aicardi-Goutieres Syndrome
AP	APurinic/APyrimidinic
APC	Adenomatous Polyposis Coli
APOBEC3A	ApoliPOprotein B mRNA Editing enzyme, Catalytic polypeptide-like protein
AT	Ataxia Telangiectasia
ATM	AT Mutated
bp	base pair
BSA	Bovine Serum Albumin
C	Cytosine
C domain	C-terminal Cysteine-rich domain
CC	Coiled-Coil
cDNA	DNA complementary to first strand RNA
CMVp	CytoMegaloVirus immediate-early promoter
Cns	Consensus
CTD	Carboxyl-Terminal Domain
CV	Coefficient of Variation
DMEM	Dulbecco's Modified Eagle Medium
DNA	DeoxyriboNucleic Acid
dpf	days post fertilization
dpi	days post injection
dsRNA	double stranded RNA
EF1- α	Elongation Factor 1-alpha
EGFP	Enhanced Green Fluorescent Protein
EN	EndoNuclease
ERV	Endogenous RetroVirus
ESC	Embryonic Stem Cell
EV	Empty Vector
FACS	Fluorescence-Activated Cell Sorting
Fc	Fold change
FCS	Foetal Calf Serum
FPKM	Fragments Per Kilobase per Million mapped reads

G	Guanine
gDNA	genomic DNA
GEO	Gene Expression Omnibus
GS	Glycine-Serine
HEPES	4-(2-HydroxyEthyl)-1-PiperazineEthaneSulfonic acid
hnRNA	heterogeneous nuclear RNA
hpf	hours post fertilization
hpi	hours post injection
HRP	HorseRadish Peroxidase
HUSH	HUman Silencing Hub
HygR	Hygromycin Resistant
IFIH1	InterFeron Induced with Helicase C domain 1
IFN	InterFeroN
IGC	Institute of Genetics and Cancer
IGR	InterGenic Region
iPSC	induced Pluripotent Stem Cell
IRES	Internal Ribosome Entry Site
IVT	*In Vitro* Transcribed
KD	Knock-Down
KDa	Kilo Dalton
KO	Knock-Out
KRAB-ZNF	Krüppel-Associated Box Zinc Finger
L1-AS	L1-AntiSense
LB	Luria-Bertani
LINE	Long INterspersed Element
LINE-1, L1	Long INterspersed Element class 1
LTR	Long Terminal Repeat
mAB	monoclonal AntiBody
mblastI	intron-interrupted blasticidin-resistance gene retrotransposition indicator cassette
MECP2	MEthyl CpG binding Protein 2
miRNA	microRNA
mneoI	intron-interrupted neomycin-phosphotransferase gene retrotransposition indicator cassette
MPP8	M-Phase PhosphoProtein 8
mRNA	messenger RNA

MSCV	Retroviral Murine Stem Cell Virus
NGS	Next Generation DNA Sequencing
NLS	Nuclear Localization Signal
NPC	Neuronal Progenitor Cell
ns	not significant
NSL	No Stem-Loop
OH	HydrOxyl
OR	Oestrogen Receptor
ORF	Open Reading Frame
PABP	Poly(A) Binding Protein
Pb	Polybrene
PBS	Phosphate-Buffered Saline
PCNA	Proliferating Cell Nuclear Antigen
PCR	Polymerase Chain Reaction
PEI	PolyEthylenImine
PIP	PCNA Interaction Protein
piRNA	PIWI interacting RNA
PIWI	P-element Induced WImpy testis
poly(A)	polyAdenine
PPHLN1	PeriPHiLiN
qPCR	quantitative PCR
RH, RNase H	RiboNuclease H
RNA	RiboNucleic Acid
RNP	RiboNucleoProtein
RPM	Reads per Million mapped reads
RRM	RNA Recognition Motif
RT	Reverse Transcriptase
RTT	ReTT syndrome
RUNX3	RUNt-related transcription factor 3
SA	Splice Acceptor
SAMHD1	SAM domain and HD domain-containing protein 1
SD	Splice Donor
SDS-PAGE	Sodium Dodecyl Sulphate-Polyacrylamide Gel Electrophoresis
sgRNA	synthetic guide RNA
SINE	Short Interspersed Element
SL	Stem-Loop

SoF	Separation of Function
SOX	Sex determining Region Y-box
Sp1	Specificity protein 1
SRY	Sex determining Region Y
ssDNA	single stranded DNA
STC1	STannioCalcin-1
SV40	Simian Virus 40
SVA	SINE-VNTR-*Alu*
T	Thymine
TASOR	Transgene Activation SuppressOR
TB	Transformation Buffer
TBE	Tris/Borate/EDTA
TBR	Tail Binding Region
TBS	Tris Buffered Saline
TE	Transposable Element
TE	Tris EDTA
TEX 19.1	Testis EXpressed 19.1
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
Tm	melting Temperature
TPRT	Target Primed Reverse Transcription
TREX1	Three prime Repair EXonuclease 1
TSD	Target Site Duplication
TU	Tübingen
TUTase	Terminal Uridyl Transferase
UTR	UnTranslated Region
v/v	volume/volume
w/v	weight/volume
WCE	Whole Cell Extract
WT	Wild-Type
YY1	Yin Yang 1
ZfL	Zebrafish LINE

Chapter 1

Introduction

1. Introduction

Life as we know it is encoded in Deoxyribonucleic acid, or DNA. These molecules contain all the information necessary to build, and allow the functioning, of a living organism. Cellular DNA is typically composed out of two polynucleotide chains or strands, made up out of monomeric units called nucleotides, which consist of a deoxyribose moiety bound to a phosphate group in its 5'carbon containing one of four nucleobases: Adenine (A), Thymine (T), Cytosine (C) or Guanine (G); bound to its carbon 1'. The phosphate group of a given nucleotide binds to the carbon 3' of the next one in the chain, constituting a sugar-phosphate backbone of each DNA strand. This strand has the structure of an aperiodic crystal, which allows it to serve as the hardware to encode the genetic information (Schrodinger, 1944). Additionally, the nucleotide pairs A and T, and C and G can establish two and three hydrogen bonds respectively, when on different strands, allowing the two strands to connect in an antiparallel fashion and coil around each other, forming a double helix. This complementarity is the base for each strand to be able to serve as the template for the synthesis of a copy of the other, allowing DNA replication and the transmission of genetic information (Crick & Watson, 1953). The collection of DNA sequences found in each species is known as "The Genome", and in vertebrates this is compartmentalised in the nucleus and mitochondria. DNA in genomes can be divided into functional units called genes, which is a sequence of nucleotides that encodes the synthesis of a functional product. When a gene is expressed, the DNA is transcribed into Ribonucleic acid, or RNA. These RNAs can be functional on their own or can be translated into a polypeptide, or protein, that performs a function. Sequences in DNA that give rise to proteins are known as "coding DNA". However, only about 5% of the human genome is coding, giving rise to 30,000 - 40,000 proteins, while the rest is non protein-coding (Lander et al., 2001). Not long ago, the non-coding part of the genome was thought to be non-functional, colloquially considered "junk-DNA", but we now know that it contains many RNA genes and regulatory sequences involved in gene expression (Dunham et al., 2012). The majority of human DNA (up to 66%) is made up out of repetitive DNA, consisting mainly of endogenous Transposable Elements (TEs) (de Koning et al., 2011; Lander et al., 2001). TEs are genomic entities that can mobilize throughout the genome and are present in all branches of life, from bacteria to plants and animals. Their activity has had, and continues to have, a significant contribution to the size,

structure and function of genomes along evolution (Böhne et al., 2008; Bourque, 2009; Doolittle & Sapienza, 1980; Feschotte, 2008; Kalendar et al., 2000; Tollis & Boissinot, 2012).

Being found in nearly any genome examined, TEs were long considered to be selfish, parasitic elements, whose mere function was to propagate their own sequence within that of the host cell (Adams et al., 1980; Doolittle & Sapienza, 1980). Although there are many examples where the activity of TEs has led or contributed to cancer development and other diseases (Bundo et al., 2014; Kazazian et al., 1988; Miki et al., 1992; Shukla et al., 2013), recent research is uncovering how TEs may also play a role in normal cell function (Hall et al., 2014; Jönsson et al., 2019; Percharde et al., 2018; Upton et al., 2015), changing the view that TEs are simply selfish DNA. The human genome contains at least three active types of TEs, the Long Interspersed Element class 1 (LINE-1 or L1), *Alu* and SINE-VNTR-*Alu* (SVA). Of these three TEs, only LINE-1s are autonomously active, encoding the machinery necessary for their own mobilisation within the genome of their hosts. This machinery also mediates the mobilisation of *Alu* and SVA. Thus, LINE-1s represent the only source of retrotransposition activity in humans. To better understand how their activity contributes to health and disease, and how we can alter it under pathological conditions, we must turn to this genomic resident and unravel its intriguing molecular relationship with their host. In this thesis, I will take a closer look at the interaction of LINE-1 and host cellular factors, as well as develop new useful tools to aid LINE-1 research.

Specifically, I will investigate the mechanisms underlying LINE-1 regulation by Ribonuclease H2 (RNase H2), a protein previously found to facilitate retrotransposition. Additionally, I will explore the zebrafish genome to identify the presence of presumably active LINE-1 subfamilies, to determine whether this animal species could be used as a model in LINE-1 related research.

1.1. General background

1.1.1. Transposable elements

TEs are a polyphyletic group of sequences that have spread throughout the genomes of all living organisms. Except for some isolated horizontal transfer events (Gilbert & Feschotte,

2018), the evolution of each element has been confined within, and occurred in parallel to, the genome of their hosts. The loss and acquisition of different regulatory sequences (Castro-Diaz et al., 2014; Jacobs et al., 2014; Khan, Smit, & Boissinot, 2006; Sookdeo, Hepp, McClure, & Boissinot, 2013), the use of different mobilization machinery by different TEs in different genomes (Malik, 2005; Malik, Burke, & Eickbush, 1999) and the recombination between elements of different families (Sookdeo et al., 2013), means that each sequence module within these elements may have different evolutionary relationships, complicating strict taxonomical studies. Nonetheless, there is a series of structural and functional criteria that allow the classification of TEs and facilitate their definition.

TEs can be subdivided into autonomous and non-autonomous elements, based on whether they do, or do not, encode their own mobilization machinery. Autonomous TEs contain one or more Open Reading Frames (ORFs) in their sequences and have the ability to generate mature messenger RNAs (mRNAs) to translate their own mobilization machinery. Although transcribed, non-autonomous TEs lack coding capacity and rely on autonomous TEs to mediate their mobilization in *trans*, literally hijacking the mobilization machinery of autonomous elements. This implies that there is a functional interplay between autonomous and non-autonomous TEs. Beside coding capabilities, TEs can also be classified according to the intermediate nucleic acid used during their mobilisation. Class I, or retrotransposons, move using an intermediate RNA, while Class II, or DNA transposons, move through a DNA intermediate (Finnegan, 1989). Retrotransposons use a “copy-and-paste” mechanism, replicative in nature, during which a new copy is generated at a new or target *locus* of the genome, different from the original copy (called ‘donor’) which remains at the initial *locus*. In contrast, DNA transposons mobilise using a "cut-and-paste" approach, which may or may not result in the replication of the element, during which the original copy excises from its location (analogically called donor *locus*) and inserts into a new target or receptor *locus*. The strictly replicative nature of the "copy-and-paste" mobilization favours an increase in copy number and therefore contributes more significantly to the generation of repetitive DNA. As a result, the copy number of retrotransposons is typically higher than that of DNA transposons. Retrotransposons can be further subdivided into Long Terminal Repeat (LTR) and non-LTR elements, based on the presence or absence of LTR sequences in their 5′ and 3′ ends. In mammals, LTR-retrotransposons are also known as Endogenous Retroviruses (ERVs), due to

similarities in structure and insertion mechanism with infectious retroviruses. The fact that certain retrotransposons seem to be related to retroviruses, with bidirectional evolutionary exchanges between the two compartments, hinder the complete resolution of systematic phylogenetic studies. However, ERVs are not currently active in the human germline, although they were until recently, with polymorphic ERV insertions segregating in the human population (Feschotte & Gilbert, 2012). In fact, all currently active TEs in humans belong to the non-LTR retrotransposon class.

Nonetheless, genomes of all currently living organisms typically contain examples from all the three main types described above, although only certain families might be active in each species. Sequences derived from currently inactive or dead elements are extremely abundant in genomes, providing a source of raw genetic material for evolution. Indeed, there are numerous examples where TEs have contributed to new regulatory and coding sequences, by providing Transcription Factor Binding Sites (TFBSs), promoters, enhancers, and silencers/insulators, in a process called exaptation (reviewed in de Souza, Franchini, & Rubinstein, 2013; Etchegaray et al., 2021). Additionally, TEs can be subjected to molecular domestication, when they are accommodated to fulfil cellular functions favouring the host. This way they continue to play a role in species adaptation and evolution.

1.1.2. Discovery of transposable elements

TEs were first described in 1950 by McClintock, who observed their activity in maize through changes in the coloration of the corn kernel due to the insertion/excision of a DNA sequence in the gene responsible for the aleurone coloration of the kernel, inactivating/rehabilitating its function (McClintock, 1950). McClintock referred to these sequences as “controlling elements”, which we now know were the first DNA Transposons characterised. At the time, her discoveries shook the foundation of current paradigms, which believed DNA to be a highly stable entity, apart from sporadic mutations, leading them to be met with strong criticism from the scientific community. However, in the late 1970s, following TEs’ discovery in other model organisms such as yeast (Farabaugh & Fink, 1980; Nevers & Saedler, 1977; Potter et al., 1979; Young et al., 1979), McClintock’s work started to be recognised, winning her a Nobel prize in Physiology or Medicine in 1983. Not long after that, the presence of TEs was also described in the human genome, although their expression

remained unconfirmed (Adams et al., 1980). Based on their length, 2 types of TEs were distinguished: Short Interspersed Elements (SINEs), less than 500 base pair (bp) long (such as *Alu* elements), and Long Interspersed Elements (LINEs), typically over 5 Kb (including the LINE-1) (Singer, 1982). Speculations about their origin and nature started to arise, such as whether they could provide binding sites for chromosomal proteins, signal chromosomal folding, or simply that they had no function for the organism and were a result of parasitism (Adams et al., 1980; Singer, 1982). In the mid 1980s, the Singer lab first confirmed the expression of LINE-1 derived RNAs in human cells, and identified the presence of several ORFs in their sequence (Skowronski & Singer, 1985). Due to structural similarities between characterised LINE-1 sequences and processed pseudogenes, it was already suggested that these elements might have expanded through a replicative process mediated by a Reverse Transcriptase (RT). Further analyses of the LINE-1 ORFs identified homologies with retroviral RTs, suggesting, for the first time, that LINE-1s could autonomously replicate by reverse transcription of their intermediary RNA (Fanning & Singer, 1987; Hattori et al., 1986; Skowronski & Singer, 1985).

What really changed LINE-1 research was the characterization of the first two LINE-1 mobilization events that lead to human disease, by Kazazian in the late 1980s. Kazazian and colleagues documented 2 independent cases of haemophilia A in unrelated patients, caused by LINE-1 insertions into exon 14 of the factor VIII gene, which disrupted its function (Kazazian et al., 1988). The absence of this mutagenic insertion in both parents indicated that the insertion was a *de novo* retrotransposition event, which might have occurred during early embryogenesis or in the parental germline (Kazazian et al., 1988). Indeed, a recent study using sensitive genotyping PCR found that this insertion was very likely maternal germline-restricted (Richardson et al., 2017). The demonstration of ongoing LINE-1 retrotransposition in the human lineage boosted this field of research. Since then, dozens of genetic disorders caused by new and heritable LINE-1 insertions have been described (Hancks & Kazazian, 2016). Additionally, numerous studies have made a clear connection between LINE-1 activity and cancer (Brégnard et al., 2016; Miki et al., 1992; Morse et al., 1988; Nguyen et al., 2018; Rodić et al., 2014; Scott et al., 2016; Shukla et al., 2013), autoimmune diseases (Mavragani et al., 2016; Steinhoff & Schulz, 2003; Yooyongsatit et al., 2015) and neurological disorders (Bundo et al., 2014; Coufal et al., 2011; Muotri et al., 2010). However, in many cases it remains unclear whether LINE-1 activity or *de novo* insertions directly contribute to the symptomology of

diseases. Additionally, the ongoing mobilisation of non-autonomous active retrotransposons in the human genome, *Alu* and SVA, has also been associated with inherited human disorders (Deininger & Batzer, 1999; reviewed in Hancks & Kazazian, 2012; Ostertag et al., 2003). As mentioned previously, these elements are *trans*-mobilised by the enzymatic machinery of active LINE-1s (Dewannieux, Esnault, & Heidmann, 2003; Hancks, Goodier et al., 2011; Raiz et al., 2012). Altogether, more than 100 cases of genetic disorders have been characterised in humans due to *de novo* retrotransposition of LINE-1s, *Alu*s and SVAs (reviewed in Hancks & Kazazian, 2016). In addition to their intrinsic mutagenic potential, the abundance of retrotransposons in the human genome, comprising at least a third of its genomic mass (de Koning et al., 2011; Lander et al., 2001), can also lead to genetic disorders driven by recombination processes (Beck et al., 2011). In sum, the combination of ongoing retrotransposon activity and their abundance in the human genome makes research into their activity, regulation and function of great value.

1.2. Non-LTR retrotransposons

As mentioned above, currently active TEs in humans are all from the Non-LTR class (1.1.1), and include LINE-1, *Alu* and SVA retrotransposons (1.1.2). Non-LTR elements have an uncertain evolutionary origin, without clear viral homologies beside the RT domain of LINE-1s. However, comparative genomic studies strongly suggest that non-LTR retrotransposons have undergone little horizontal transfer and have rather dispersed in the genome of different species by vertical transmission (Malik et al., 1999), with the exception of BovB LINE-1s (Ivancevic et al., 2018). This makes it possible to differentiate lineages of non-LTR retrotransposons and group them in distinct clades. The retrotransposition machinery of autonomous non-LTR elements cluster around a basic ORF containing an RT domain. Unlike retroviruses and LTR-retrotransposons, which reverse transcribe their intermediate RNAs into dsDNA in the cytoplasm before integration, non-LTR elements reverse transcribe the intermediate RNA in the nucleus, directly at the new genomic target *locus*, in a process known as Target Primed Reverse Transcription (TPRT). Throughout evolution, different lineages of non-LTR retrotransposons have acquired additional enzymatic activities/domains, such as Endonuclease (EN), Ribonuclease H (RNase H) or nucleic acid chaperone activities (Malik, 2005; Malik et al., 1999), which are involved in the TPRT process. Presumably, the acquisition of new

domains might allow non-LTR retrotransposons to be more and more independent from host cellular factors during retrotransposition. Thus, there is significant variability in the structure of different non-LTR clades, with reported differences in the promoter used to transcribe RNAs, the utilised retrotransposition machinery and the way this machinery is recruited, as well as the tropism of new insertions. Key biochemical mechanistic insights into the non-LTR retrotransposition process have been uncovered using a site-specific non-LTR element from insects, R2 retrotransposons, which can be found in *Bombyx mori*, *Drosophila spp.* and many other arthropods (reviewed in Eickbush & Eickbush, 2015). However, the best studied retrotransposition process is that of human and mouse LINE-1s, and knowledge obtained from these elements is often used to predict how other, less studied, non-LTR elements are regulated, retrotranspose and impact genomes.

1.2.1. LINE-1 structure and mobilisation

LINE-1s are the most widely distributed TE in vertebrates, although their copy number is highly variable among species (Tollis & Boissinot, 2012). In humans, ~17% of our genome is made up of LINE-1-derived sequences. If we also account for *Alu* and SVA, mobilised in *trans* by LINE-1s, an astonishing >30% of our genome has been generated by LINE-1 activity throughout evolution. The completion of the first human reference genome, together with recent long-read whole genome sequencing, revealed the presence of 1,2 million LINE-1 copies in the haploid human genome, belonging to different subfamilies (Lander et al., 2001; Nurk et al., 2021). Within these LINE-1 subfamilies that can be found in our genome, up to 16 are exclusive to primates, which would have spread in genomes since the divergence with rodents some 70 million years ago (Khan et al., 2006). LINE-1 evolution in mammals resulted in a ladder-shaped phylogeny, with typically only a few families producing novel insertions, and each subfamily being succeeded by a younger one (Boissinot & Sookdeo, 2016). This mode of evolution, reminiscent of the Red Queen hypothesis (Van Valen, 1973), is consistent with an arms race between the host repressing LINE-1 retrotransposition and LINE-1s evolving to escape this repression. Consistently, human LINE-1 subfamilies are of monophyletic origin (Badge, Alisch, & Moran, 2003; Jacobs et al., 2014), and only the most recent subfamily of elements (L1PA1, or L1Hs), homo sapiens specific, is currently active in the human population. Within this subfamily, only a subset of elements continues to generate inter- and intra-

individual variability in humans (Beck et al., 2010; Brouha et al., 2003; Sassaman et al., 1997). In fact, the majority of L1PA1 copies are not retrotransposition competent or active, due to the accumulation of mutations in their functional domains or alterations in their structure, generated during and/or after integration (Grimaldi, Skowronski, & Singer, 1984; Lander et al., 2001). Using the first human reference genome, ~100 L1Hs elements with intact ORFs were found in an average human, of which at least a third were polymorphic in the human population (Brouha et al., 2003). Notably, using an *in vitro* retrotransposition assay in cultured cells, it was established that an average human genome contains 80-100 active L1s (Moran et al., 1996). Remarkably, nearly 90% of the retrotransposition potential in an average human genome is generated by just 5-10 highly active L1Hs elements, which were termed "hot" L1s (Brouha et al., 2003). More recently, through the analysis of additional genomes, it was found that "hot" L1s are highly polymorphic, indicating they may be segregating in the human population, subjected to natural selection and other evolutionary forces (Beck et al., 2010).

A consensus active LINE-1 consists of a 6kb long single transcriptional unit containing two non-overlapping ORFs, ORF1 and ORF2, separated by a short Intergenic Region (IGR), flanked by 5' and 3'-Untranslated Regions (5´UTR and 3´UTR, of 900 and 250 bp respectively) and ending in a polyadenine (poly(A)) tract of variable length (Grimaldi et al., 1984; Scott et al., 1987; Singer et al., 1993) (***Figure 1.1***). The 5'UTR contains an internal RNA polymerase II promoter, which allows the generation of the full-length RNA used as a retrotransposition intermediate, and that encompass the totality of the LINE-1 sequence (Swergold, 1990). The core of the promoter activity appears to be contained within the first 150 bp of the 5'UTR sequence, where binding sites for important Transcription Factors (TF) such as Yin Yang 1 (YY1), Sex determining Region Y (SRY)-box 2 (SOX 2) and Runt-related transcription factor 3 (RUNX3) have been characterised (Athanikar, Badge, & Moran, 2004; Muotri et al., 2005; Swergold, 1990; Tchénio, Casella, & Heidmann, 2000; Yang et al., 2003). Besides regulating promoter activity, YY1 binding near the 5´end of the LINE-1 sequence appears to be essential for the positioning of the transcription machinery at the beginning of the LINE-1 sequence, ensuring that transcription starts at the +1 nucleotide (Athanikar et al., 2004). Intriguingly, the same 5'UTR has also been shown to possess conserved antisense promoter activity (Macia et al., 2011), which generates a chimeric transcript containing approximately 500 bp from the antisense LINE-1 strand and a variable length genomic sequence (Speek, 2001). Notably, a

fraction of L1-Antisense (L1-AS) derived transcripts could be translated in pluripotent cells, giving rise to a protein of variable sequence known as L1-ORF0p, which could enhance mobilization of LINE-1s by an uncharacterised mechanism (Denli et al., 2015).

Figure 1.1. Schematic of an active human LINE-1 – An active LINE-1 starts with a 900 bp long 5'UTR containing both sense and antisense promoter activities (indicated with black arrows). This is followed by two ORFs (*ORF1*, red box; *ORF2*; blue box), separated by a short (63 bp) IGR, and ends in a 250 bp 3'UTR with a poly(A) tract of variable length (A_n). Relevant domains are indicated within the ORFs: Coiled-Coil domain (CC), RNA Recognition Motif (RRM), Carboxyl-Terminal Domain (CTD), Endonuclease domain (EN), PCNA Interaction Protein motif (PIP), Z-domain (Z), Reverse Transcriptase domain (RT) and C-terminal Cysteine-rich domain (C).

The two proteins encoded by the main ORFs in the LINE-1 sequence, ORF1p and ORF2p respectively, are essential for retrotransposition. L1-ORF1p has a molecular weight of about 40 kDa and is synthesized in a cap-dependent manner (Dmitriev et al., 2007). It contains a Coiled-Coil (CC) domain at its N-terminal, that mediates the formation of protein trimers (Khazina et al., 2011); a non-canonical RNA Recognition Motif (RRM) is found in its central region, made up of two non-canonical Ribonucleoprotein domains (RNP1 and 2) and four conserved salt bridges; and a series of conserved nucleotides in the Carboxyl-Terminal Domain (CTD) that that have been described to be important during retrotransposition (Doucet et al., 2010; Gilbert & Moran, 2002; Khazina & Weichenrieder, 2009). L1-ORF1p possesses nucleic acid chaperone activity, that presumably facilitates certain aspects of TPRT (Martin & Bushman, 2001). The protein encoded by ORF2 is approximately 150 kDa (Ergün et al., 2004) and is synthesized in very low quantity, apparently by a re-initiation phenomenon from a ribosome finishing ORF1 translation (Alisch et al., 2006). ORF2p has EN and RT activities, which are critical for the initiation of reverse transcription of the LINE-1 RNA at site of integration (Feng et al., 1996; Mathias et al., 1991). Additionally, it contains a Proliferating Cell Nuclear Antigen (PCNA) Interaction Protein (PIP) motif, a Z-domain and a C-terminal Cysteine-rich (C) domain (Christian et al., 2016; Fanning & Singer, 1987; Taylor et al., 2013). The Z-domain contains a putative RNA-binding motif (Jamburuthugoda & Eickbush, 2014), while the C domain shows high affinity to RNA, which suggests they may play a role during cDNA synthesis and/or nucleic acid

interactions during retrotransposition (Piskareva et al., 2013). The PIP motif allows interaction of ORF2p with PCNA and is important during retrotransposition, although the underlying mechanism is unclear (Taylor et al., 2013).

At the start of a typical round of retrotransposition, the full-length bicistronic LINE-1 RNA is transcribed from the internal promoter located in its 5′UTR by host encoded RNA Polymerase II (Athanikar et al., 2004; Grimaldi et al., 1984; Swergold, 1990) (***Figure 1.2A***). This RNA is exported to the cytoplasm, where LINE-1 encoded proteins are synthesised, and ORF1p and ORF2p bind their encoding RNA with *cis* preference, forming the LINE-1 Ribonucleoprotein Particle (RNP) (Hohjoh & Singer, 1996; Kulpa & Moran, 2005; Wei et al., 2001). The preference of ORF2p to bind its encoding RNA in *cis* appears to be enhanced by slowing translation of the C-terminal region of this protein, allowing the nascent and partially functional ORF2p protein to bind the poly(A) tail of the LINE-1 RNA by simple proximity (Ahl et al., 2015). L1 RNPs can access the nucleus in dividing cells during mitosis, when the nuclear membrane breaks down (Mita et al., 2018), or in non-dividing cells, likely through mediation of interacting partners (Kubo et al., 2006; Macia et al., 2017). Once in the nucleus, L1 RNPs can retrotranspose by a mechanism termed TPRT, which is initiated by the endonucleolytic cleavage of the bottom strand of genomic DNA at a loose target consensus sequence (5'-TTTTT/AA-3' and variants), catalysed by the EN activity of ORF2p (Cost et al., 2002; Flasch et al., 2019; Jurka, 1997) (***Figure 1.2B***). The DNA segment flanking the cleavage site dehybridizes, and the single-stranded, T-rich 3'-Hydroxyl (OH) flap hybridizes to the 3' poly(A) region of the intermediate L1 RNA, in a "snap-velcro" manner (Viollet, Monot, & Cristofari, 2014). The 3'-OH of the genomic DNA acts as a primer and allows ORF2p-mediated reverse transcription of the LINE-1 RNA, generating the first-strand cDNA of the insertion, resulting in a LINE-1 RNA:cDNA hybrid covalently linked to the genome (Luan et al., 1993). It is unclear how this structure is resolved and how second strand synthesis occurs, but it is speculated that host cellular factors, including the DNA repair machinery, are involved in these steps (Benitez-Guijarro et al., 2018; Gasior etal., 2006; Suzuki et al., 2009; Taylor et al., 2013). Most new LINE-1 insertions are typically flanked with short duplications of the genomic insertion site, called Target Site Duplications (TSDs). It has been speculated that they arise as a consequence of asymmetric cleavage in the opposite genomic strand, which occurs to generate a 3′OH that is then used to prime second strand cDNA synthesis. The presence of 2-20bp long TSDs, a long poly(A) tail and integrations into a

consensus EN recognition site (5'TTTTT/AA) are considered hallmarks of standard TPRT (Jurka, 1997). However, most of the *de novo* LINE-1 insertions are 5' truncated to some degree, most likely due to inefficiency of the retrotransposition process and/or interference from host factors restricting LINE-1 mobilisation (Gilbert, et al. 2005) (1.2.2).

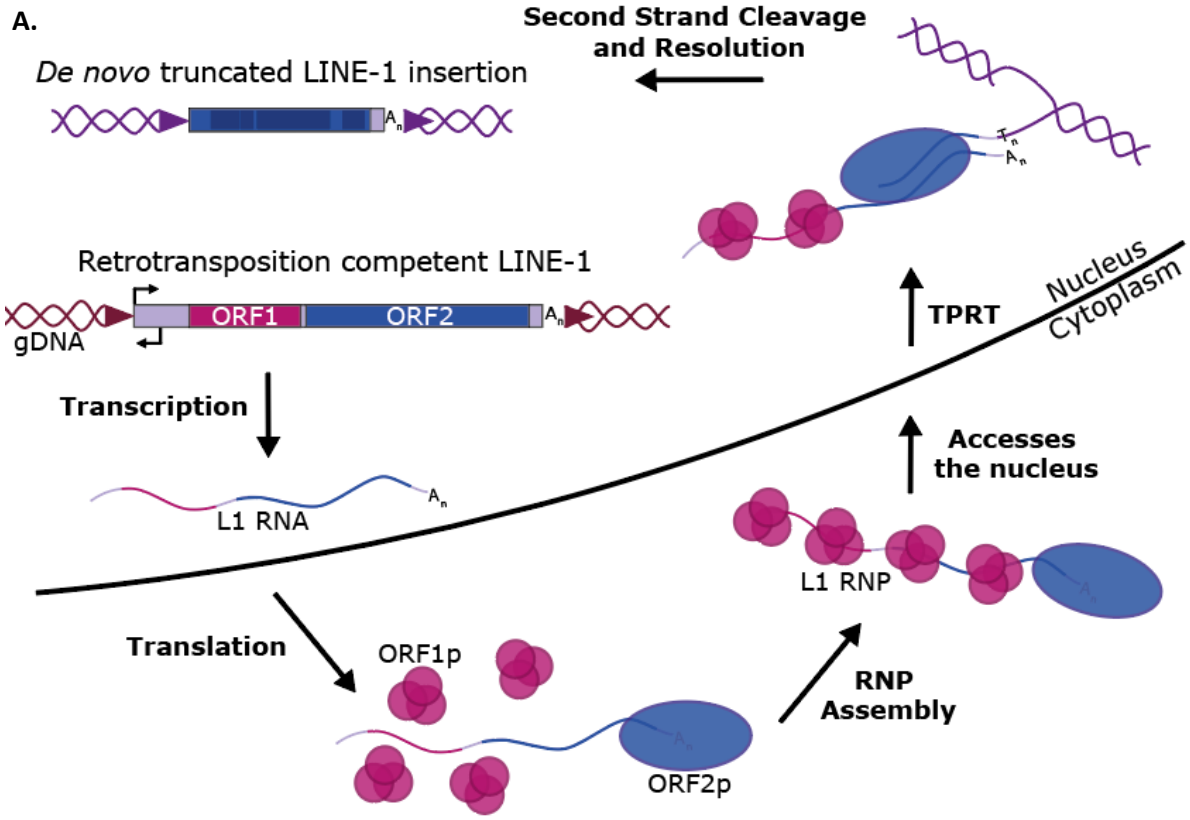


Figure 1.2. Schematic of the LINE-1 retrotransposition cycle and TPRT process – (A) A schematic overview of the LINE-1 retrotransposition cycle. A retrotransposition competent LINE-1 is transcribed, and the RNA is transported to the cytoplasm where its proteins are synthesised and ORF1p and ORF2p bind their encoding RNA with *cis* preference. The resulting L1 RNP accesses the nucleus where the EN activity of ORF2p catalyses single strand cleavage, initiating TPRT and finally resulting in a *de novo* LINE-1 insertion (frequently 5' truncated). Promoters are depicted as arrows; ORF1 and ORF2 are depicted using red and blue boxes, respectively, and their encoded proteins as circles, using matching colours; TSDs are depicted by triangles. (B) [in the next page] A detailed representation of the TPRT process. The EN recognition motif is cleaved by the EN activity of the ORF2p releasing a single-stranded, T-rich 3'-OH flap to which the 3' poly(A) region of the L1 RNA intermediate can prime. This initiates reverse transcription of the L1 RNA by the RT activity of the ORF2p. The ORF2p is believed to perform template switching from the L1 RNA to the genomic DNA, anchoring the new insertion. This structure is resolved through unknown processes, resulting in a de novo insertion in the genomic DNA, typically flanked by TSDs (containing the restored EN recognition motif in the 5' genomic flanking region.

B.

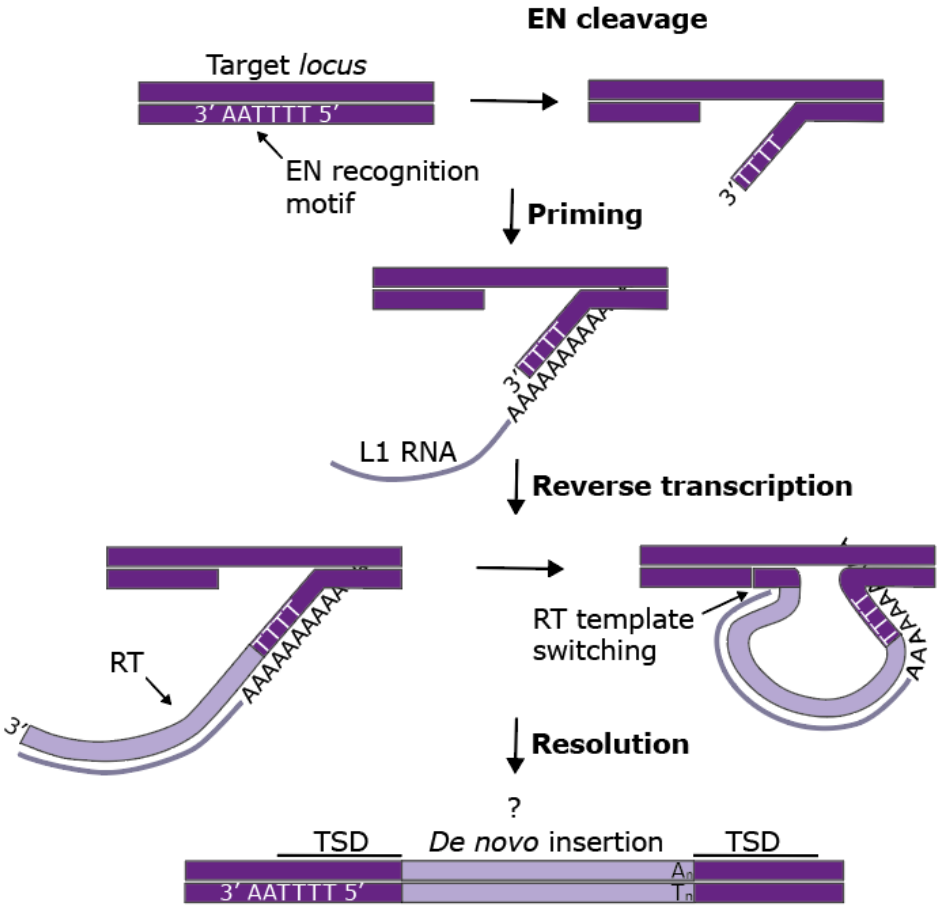


Figure 1.2. Schematic of the LINE-1 retrotransposition cycle and TPRT process – Legend in the previous page.

1.2.2. Activity and regulation of non-LTR elements

As mentioned above, due to the almost exclusive vertical transfer of non-LTR elements within species, their perpetuation is linked to that of their hosts. Additionally, TEs are a source of genetic diversity on which natural selection can act to drive innovations. However, despite these potential benefits as drivers of evolution, the host is at odds with the deleterious consequences that insertional mutagenesis can have on genome function. This has created a natural selection scenario that favours the accumulation of TE copies whose activity can be adjusted within margins that do not clash with the biological fitness of the host. Because of this forced co-evolution, a complex interplay between the TE life cycle and host cellular factors has emerged through evolution, tolerating certain levels of TE activity in particular developmental and cellular niches of the host. In humans, under physiological conditions, TE activity has exclusively been described in early embryonic development (Garcia-Perez et al., 2007; van den Hurk et al., 2007), and in somatic cells from the neuronal lineage (Macia et al.,

2017; Muotri et al., 2005; Upton et al., 2015). Although this level of TE activity may be acceptable at a species level, it does not preclude from deleterious mutations at the individual level, which can suffer diseases by *de novo* TE insertions (reviewed in Macia, Blanco-Jimenez, & García-Pérez, 2015). It is therefore not surprising that in humans, more than 100 cases of hereditary diseases have been linked to *de novo* insertions of TEs in the germline (reviewed in Hancks & Kazazian, 2016). Additionally, under certain pathological circumstances TEs can become deregulated, with detrimental consequences to the host (1.1.2). Thus, it is vital to human health that the correct levels of TE activity are maintained. Predicting under what circumstances TEs become deregulated, and how to regain control when this occurs, are two major areas of interest to the biomedical field. However, this requires an understanding of how cells regulate TEs, and a deeper understanding of their potential impacts.

TE regulation occurs at all the different stages of their life cycle, although much of what we know is derived from studies on mammalian LINE-1 retrotransposons (***Figure 1.3***). Previous studies have documented that one of the main regulatory mechanisms controlling mammalian TEs is by targeting LINE-1 transcription. Notably, targeting transcription of autonomous TEs will also control the activity of non-autonomous elements. Thus, epigenetics appears to be one of the primary control points of retrotransposon activity in mammals. Epigenetic marks at TE-sequences are primarily established during early embryonic development, and are maintained throughout the life of an organism, controlling their expression in somatic tissues (Bestor & Bourc'his, 2004; Yoder, Walsh, & Bestor, 1997). For most TEs, this is mediated through different proteins of the Krüppel-Associated Box Zinc Finger (KRAB-ZNF) protein family, which appear to have expanded and evolved to recognize specific sequences of the different families and subfamilies of TEs (Castro-Diaz et al., 2014). As an example, KRAB-ZNFP91 can selectively control the expression, and activity, of SVA retrotransposons as well as certain primate specific LINE-1s from the L1PA3 subfamily (Jacobs et al., 2014). Until the emergence of the L1PA2 subfamily, nearly 12 million years ago, this protein was able to repress LINE-1 activity in pluripotent cells. Although additional KRAB-ZNFPs might control L1PA2 elements, the exact protein/s targeting these relatively young LINE-1s have not been characterized. Interestingly, several of the more recent LINE-1 subfamilies (i.e., L1Hs) have arisen from an evasion event of this silencing mechanism, and appear to mainly be silenced by DNA methylation of a canonical CpG island present in their 5'UTR/promoter (Muotri et al., 2010; Yoder et al., 1997). In germ

cells, DNA methylation of LINE-1 promoters appears to be mediated by small cellular RNAs, likely from the PIWI-interacting RNA class (Houwing et al., 2007; Marchetto et al., 2013; Xu et al., 2008), while during early embryogenesis, other candidate factors including p53 and YY1 are involved in the establishment of DNA methylation at L1 promoters (Sanchez-Luque et al., 2019; Wylie et al., 2016). Additionally, the Human Silencing Hub (HUSH) complex, comprised of Transgene Activation Suppressor (TASOR), M-Phase Phosphoprotein 8 (MPP8) and Periphilin (PPHLN1, isoform 2), mediates chromatin remodelling of young LINE-1s in pluripotent cells, influencing their silencing (Liu et al., 2018), but whether there is a connection with DNA methylation remains to be determined. Finally, the Methyl CpG binding Protein 2 (MECP2) has also been demonstrated to be involved in the repression of LINE-1s, likely through the binding of the methylated LINE-1 promoter and recruiting other silencing factors (Muotri et al., 2010; Yu, 2001). If epigenetic suppression is not taking place, cellular TF including RUNX3 (Yang et al., 2003), Specificity protein 1 (Sp1) (Fedorov, Lukyanov, & Podgornaya, 2006), YY1 (Athanikar et al., 2004; Becker et al., 1993; Minakami et al., 1992) and SOX 11 (Muotri et al., 2010; Tchénio et al., 2000) can bind the L1 promoter and activate LINE-1 RNA synthesis. Notably, by modulating the action of these factors, such as by the occupation of the SOX sites by SOX2, or through the capturing of TFs by other cellular factors, like interleukin 16 which can sequester Sp1, host cells can control LINE-1 transcription levels (Hotter et al., 2019; Kuwabara et al., 2009).

Although transcription might be the most effective way to control retrotransposon activity in human cells, numerous post-transcriptional mechanisms that target LINE-1 activity have also been documented. In fact, cells have several mechanisms in place to prevent retrotransposition by specifically targeting the intermediary LINE-1 RNA. Cellular factors such as the Microprocessor complex (Drosha-DGCR8) (Heras et al., 2013), P-element Induced Wimpy testis (PIWI) and PIWI interacting RNAs (piRNAs) (Houwing et al., 2007; Marchetto et al., 2013), condensin II and GAIT complexes (Ward et al., 2017), as well as Terminal Uridyl Transferases (TUTases) in cooperation with MOV10 (Li et al., 2013; Warkocki et al., 2018), are some of the known host factors able to target LINE-1 RNAs and trigger their degradation. Furthermore, certain small RNAs from the microRNA (miRNAs) class have recently been described as regulators of LINE-1 retrotransposition, especially in the context of tumoral cells. Two miRNAs that repress L1 retrotransposition, miR-128 (Idica et al., 2017) and let-7 (Tristán-

Ramos et al., 2020), have been shown to interact with ORF2 sequences, suggesting that miRNA control of retrotransposition is conserved through evolution. Furthermore, MOV10, Testis Expressed 19.1 (Tex19.1) and Double-stranded RNA-specific Adenosine Deaminase 1 (ADAR1) have been shown to interact with L1 RNPs, preventing integration. While MOV10 and Tex19.1 induce L1 RNP degradation through various mechanisms (Goodier et al., 2012; MacLennan et al., 2017), ADAR1 is believed to merely physically interfere with the activity of L1 RNPs, acting in a manner that is independent of ADAR1's enzymatic activity (Orecchini et al., 2017). Finally, there are additional host factors known to interfere with TPRT. LINE-1 cDNA synthesis can be interfered with by the SAM domain and HD domain-containing protein 1 (SAMHD1), Three prime Repair Exonuclease 1 (TREX1) and Apolipoprotein B mRNA Editing enzyme, Catalytic polypeptide-like protein 3A (APOBEC3A). The SAMHD1 enzyme appears to inhibit L1-ORF2p reverse transcriptase activity by reducing the content of nucleotide triphosphate (Zhao et al., 2013), while TREX1 seems to be involved in the degradation of reverse transcribed L1 cDNAs (Stetson et al., 2008). APOBEC3A also appears to attack exposed cDNA or single strand segments during the TPRT process, where it catalyses the conversion of cytosine residues to uracil by deamination (Richardson et al., 2014). This will then be recognized by the DNA repair machinery in a process that involves cutting the DNA strand, resulting in truncation or total elimination of nascent L1 cDNAs.

While research has clearly uncovered restrictors of retrotransposition, complementary research has identified factors that promote LINE-1 retrotransposition (i.e., activators). Although autonomous, LINE-1 relies on certain cellular factors for the completion of the retrotransposition process. The Poly(A) Binding Proteins N1 and C1 (PABPN1 and PABPC1) can associate with L1 RNPs, likely through binding the L1 RNA, and are believed to play a role in L1 RNP translocation to the nucleus (Dai et al., 2012). Furthermore, PCNA and RNase H2 are required for efficient LINE-1 mobilization, acting at LINE-1 integration steps. Notably, mediated by the L1-ORF2p PIP motif the L1 RNP can interact with PCNA at chromatin, forming a complex with other components of the replication fork, although the exact function of PCNA during this process requires elucidation (Taylor et al., 2013). The role of RNase H2 appears more clear, as its ability to degrade the RNA from RNA:DNA hybrids seems to be mediating the elimination of the LINE-1 RNA after first strand cDNA synthesis, allowing second strand synthesis to occur (Benitez-Guijarro et al., 2018). However, the exact mechanism through which RNase H2 and

LINE-1 interact is so far unknown and will be further explored in this thesis. Once the LINE-1 RNA has been reversed transcribed at the target *locus*, the DNA repair machinery is believed to resolve the final steps of the integration, resulting in the completion of the LINE-1 retrotransposition cycle (Gasior et al., 2006; Suzuki et al., 2009).

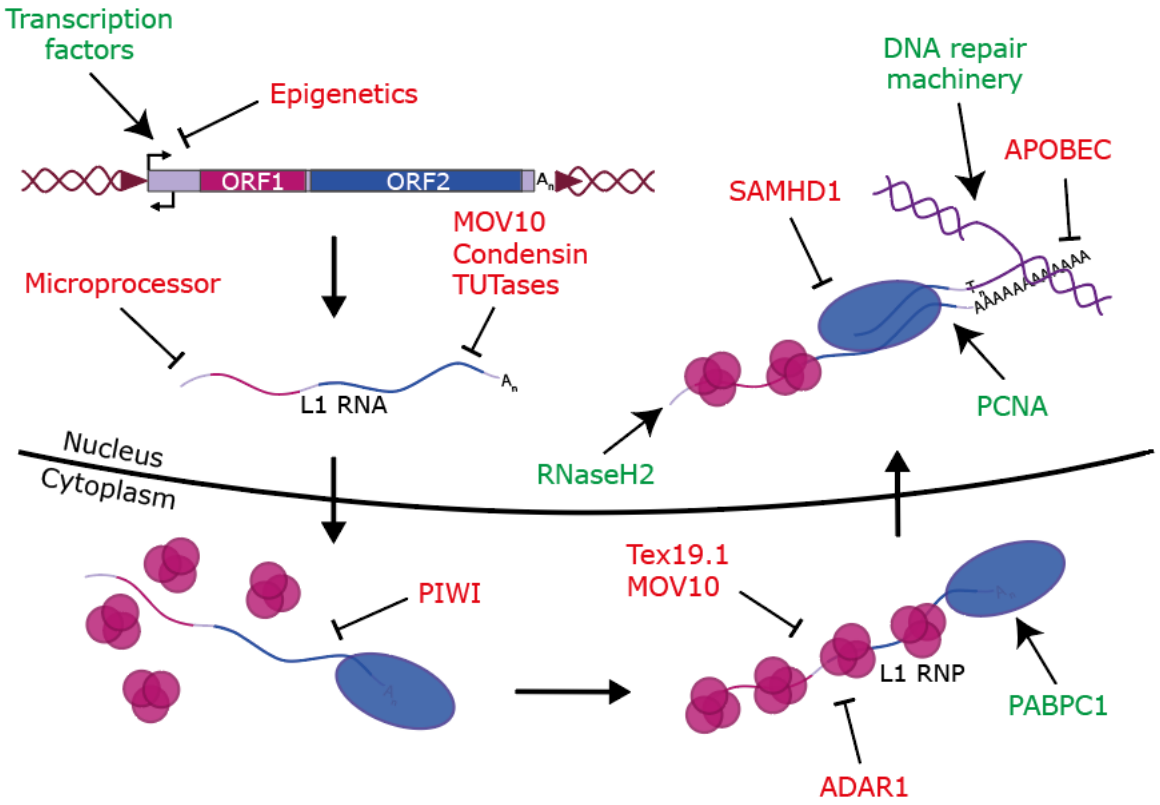


Figure 1.3. **Schematic overview of known LINE-1 regulators** – A schematic overview of the LINE-1 retrotransposition cycle (see also ***Figure 1.2A***), showing the identified cellular regulators and at which step they act to promote or interfere with LINE-1 retrotransposition (repressors = red; activators = green). See text for detailed explanation.

The absence or presence of these regulatory factors is believed to influence the cell-type specific activity of retrotransposons (reviewed in Goodier, 2016). For example, LINE-1 activity has been observed in cancers, where global hypomethylation of the genome and mutation of p53 often occur (reviewed in Burns, 2017 and Scott & Devine, 2017; Wylie et al., 2016). Indeed, reduced methylation of the LINE-1 promoter has been corroborated in various studies, not only in cancer cells, but also in embryonic development, the brain and other somatic tissues in response to factors such as stress, ageing, chronic inflammation and viral infections (Cecco et al., 2019; Coufal et al., 2009; Garcia-Perez et al., 2007; Sanchez-Luque et al., 2019; Sudhindar et al., 2021; Van Meter et al., 2014). Consistently, several LINE-1 donors of *de novo* insertions identified in cancer patients already showed hypomethylation of their

promoter in healthy tissue adjacent to the tumour, suggesting pre-tumour hypomethylation of certain copies, or failure to establish DNA methylation during embryonic development, which may facilitate mobilization and contribute to tumour development/progression (Nguyen et al., 2018; Schauer et al., 2018; Scott et al., 2016; Sudhindar et al., 2021).

1.2.3. Common structures found in Non-LTR retrotransposons

Most of the knowledge regarding the retrotransposition cycle, regulation and activity of non-LTR elements discussed above, has been obtained using mainly mammalian LINE-1 as a prototype non-LTR element. Although there will be a large overlap between what is known for LINE-1s and other elements, there are certain key differences (***Figure 1.4***). For instance, several strategies to initiate transcription in a way that ensures no genetic information is lost during retrotransposition have been described for different TEs (Deberardinis & Kazazian, 1999; Goodier et al., 2001; Mizrokhi, Georgieva & Ilyin, 1988; Swergold, 1990). Most non-LTRs use a canonical internal promoter in their 5'UTR (such as is the case for human and Zebrafish LINE-1s (ZfL1s)) (Boissinot & Sookdeo, 2016; Sookdeo et al., 2013), while others use different numbers of monomeric repeats arranged in tandem (as found in mouse LINE-1s and the Zebrafish LINE-2-2 (ZfL2-2)) (Deberardinis & Kazazian, 1999; Goodier et al., 2001; Sugano, Kajikawa, & Okada, 2006). The presence of a promoter in the 5'UTR provides the element with autonomy for the generation of its intermediate RNA. However, some non-LTR elements lack this feature and resort to using external transcription initiation sites provided by the host, combined with ribozymes (found amongst others in vertebrate R2 non-LTR retrotransposons and trypanosome LINE-like elements), which frequently conditions their tropism for target *loci* (de la Peña & Cervera, 2017; Eickbush & Eickbush, 2010; George et al., 2010; Sánchez-Luque et al., 2011). As the promoter region of non-LTR retrotransposons is highly regulated by the host organism (1.2.2), and there is a strong selective pressure for non-LTRs to evade these mechanisms of repression, promoters are one of the most variable regions found in these elements, and comparative genomic studies suggest that their sequences have changed significantly throughout evolution (Khan et al., 2006; Sookdeo et al., 2013).

The enzymatic activities of the protein(s) encoded by autonomous non-LTR elements also vary across different examples (***Figure 1.4***). Different lineages of autonomous non-LTR elements have acquired enzymatic activities additional to the core encoded RT domain along

evolution (Malik, 2005; Malik et al., 1999). From an evolutionary angle, a proposed model speculates that ancestral non-LTR retrotransposons, lacking EN activity, would have used the 3'-OH ends of genomic DNA generated by DNA breaks or DNA replication processes as a primer to initiate TPRT (Flasch et al., 2019; Zhong & Lambowitz, 2003). The acquisition of an EN domain would have relieved them from using this opportunistic mechanism of insertion. Different EN domains have been found across non-LTR elements, establishing different tropism preferences. For instance, elements whose EN domain is related to type IIS restriction enzymes (like arthropod R2 elements), can have extreme specificity for particular genomic sequences (Eickbush & Eickbush, 1995; Xiong & Eickbush, 1990). Because of this peculiarity, these elements are also known as site-specific LINEs. However, most lineages are derived from the acquisition of an Apurinic/Apyrimidinic (AP)-type EN in the N-terminal position, which confers them a wider degree of target sites, as is the case for mammalian and zebrafish LINE-1s (Feng et al., 1996; Feng, Schumann, & Boeke, 1998; Olivares et al., 1999; Sugano et al., 2006). More recently in LINE evolution, some lineages acquired an RNase H domain, normally located after their RT domain, which is thought to eliminate the intermediary RNA from the RNA:DNA hybrid after reverse transcription of the first cDNA. However, this domain is not widely distributed in non-LTR retrotransposons and has only been found in plant LINEs, a LINE element from *Trypanosoma cruzi*, L1Tc, and in *Drosophila silvestris* LOA elements (Boissinot & Sookdeo, 2016; Malik, 2005; Olivares et al., 2002; Smyshlyaev et al., 2013). Finally, some lineages acquired a second ORF positioned in 5' of the original RT-encoding ORF, which encodes a protein with a coil-coiled and RNA-binding domain, often possessing nucleic acid chaperone activity (Hohjoh & Singer, 1996; Kolosha & Martin, 1997; Martin & Bushman, 2001; Sugano et al., 2006). The acquisition of a second ORF resulted in a dicistronic RNA, leading to the development of unconventional mechanisms for the translation of the RT-encoding ORF, which can no longer take place in a canonical cap-dependent manner. These unconventional mechanisms of ORF2 translation are unknown for many elements, but in the case of human LINE-1s it is believed to occur through an inefficient re-initiation event by a ribosome that completed ORF1 translation (Alisch et al., 2006; Dmitriev et al., 2007). Instead, mouse LINE-1s appear to use an Internal Ribosome Entry Site (IRES) located near the 3' end of the ORF1 (Li et al., 2006), which can tolerate a significant amount of sequence change as revealed by codon optimization experiments (Han & Boeke, 2004). This set-up contributes to the low abundance of ORF2p compared to ORF1p.

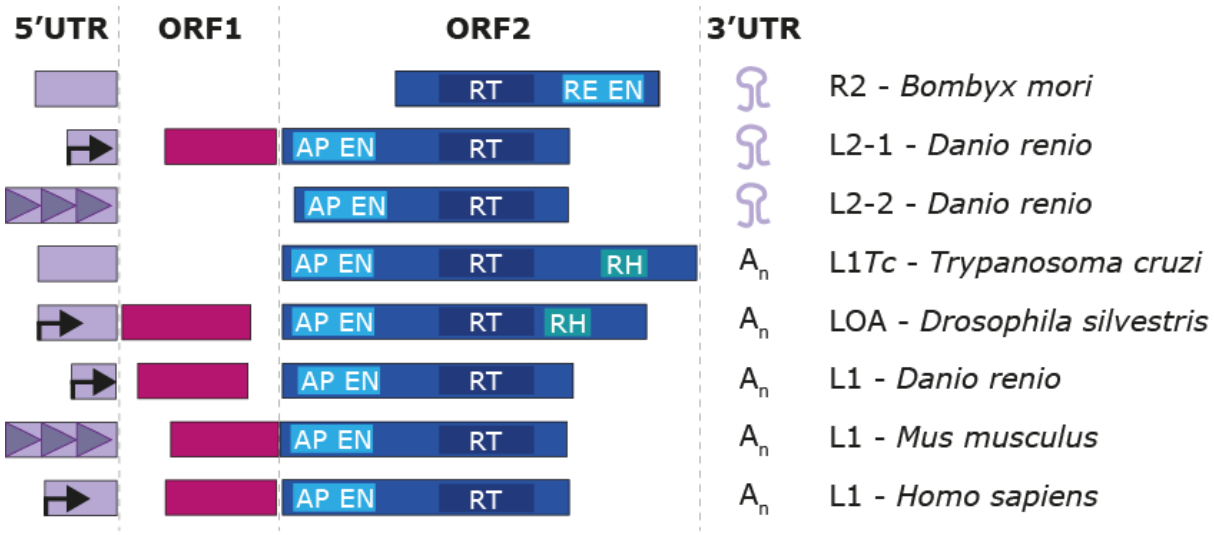


Figure 1.4. Schematic overview of the key differences found in selected lineages of non-LTR retrotransposons – A schematic overview of the 5' UTR promoter, ORFs and 3'UTR found in specific lineages of autonomous non-LTR elements in different species. The presence of a unitary (arrow) *vs* monomeric promoter (triangles), the number of ORFs, the different enzymatic domains in the ORFs, Apurinic/Apyrimidinic (AP)-type EN (AP EN), type IIS restriction enzyme EN (RE EN), RT and RNase H (RH), and the presence of a recognition motif (stem-loop) *vs* poly(A) tract (A_n) is indicated.

Autonomous non-LTRs can also be classified as stringent or relaxed types, depending on how their machinery recognizes their encoding RNA (reviewed in Okada et al., 1997), to ensure that their RNA is used as a template for the synthesis of the new copy (Wei et al., 2001) (***Figure 1.4***). The enzymatic machinery encoded by stringent elements recognizes a specific structural or sequence motif contained in the 3' RNA tail, as in ZfL2s and R2s from arthropods (Kajikawa & Okada, 2002; Luan & Eickbush, 1995; Otsu et al., 2017). The RT of the relaxed types generally shows avidity for the poly(A) tail of RNAs, and is able to bind it even before translation is completed, guaranteeing cis-binding due to proximity, as is the case for mammalian LINE-1s (Ahl et al., 2015; Moran et al., 1996). This allows non-autonomous elements, such as trypanosomatids NARTc and RIME (Bringaud et al., 2002) and human *Alu* and SVA, to mimic these features to hijack the RT of their driver elements (Ahl et al., 2015; Raiz et al., 2012). These two strategies may also lead to the presence of different features in the 3' end of the new insertions. Typically, stringent elements end in an exact recognition motif. On the other hand, the relaxed type frequently transduces downstream genomic sequences from the donor to the target *locus,* since the element's weak polyadenylation signal causes the transcription to continue until an opportunistic polyadenylation signal downstream in the genome is

encountered, incorporating extra sequences into the intermediary RNA (and the subsequent insertion). This process is known as 3′ transduction, and when it affects mammalian LINE-1s located in genes, coding exons can be retrotransposed to new genomic regions, in a process known as exon shuffling (Moran, DeBerardinis, & Kazazian, 1999).

1.3. LINE-1 in human disease

1.3.1. Impact of LINE-1 on the genome and cellular function

There are many ways in which retrotransposon activity and integration can impact the genome and disrupt or alter gene function, the most obvious one being through disruption of coding sequences (***Figure 1.5B***) (reviewed in Rebollo, Romanish, & Mager, 2012). Most of what we know about LINE-1 impact comes from research on mammalian LINE-1s, which will be the focus of this section. Because of their abundance in mammalian genomes, LINE-1 sequences can act as substrates for homologous and non-homologous recombination processes, which can lead to genomic alterations and disease (reviewed in Beck et al., 2011). Additionally, the LINE-1 sequence contains numerous suboptimal polyadenylation sites (Perepelitsa-Belancio & Deininger, 2003) as well as multiple potential Splice Donor (SD) and Acceptor (SA) sites, generating a variety of subgenomic and processed LINE-1 transcriptional side products (Belancio, Hedges, & Deininger, 2006) (***Figure 1.5A***). These sites have also been reported to contribute to the generation of hybrid transcripts between L1 elements and host genes, illustrating how a LINE-1 insertion can lead to alternative splicing and/or exonization (Belancio, Roy-Engel, & Deininger, 2008), as well as provide premature transcriptional stop signals that lead to truncated gene transcripts (Perepelitsa-Belancio & Deininger, 2003). Notably, several RNA binding proteins act by nucleating retrotransposon sequences, preventing genes from exonizing LINEs and SINEs during transcription (Attig & Ule, 2019). Furthermore, the conserved antisense promoter in the LINE-1 5'UTR can serve as an alternative promoter for protein-coding loci located in 5' of the insertion, altering gene expression and/or creating noncoding RNAs (Denli et al., 2015; Faulkner et al., 2009; Speek, 2001). When inserted inside a gene, this promoter activity can also result in split gene transcripts, with one half of the gene falling under control of the sense, and the other of the antisense promoter (Wheelan et al., 2005). The cellular systems in place to recognise and silence the LINE-1 promoter (1.2.2) may prevent this

deleterious spurious transcription, but the epigenetic silencing of this LINE-1 *loci* may also interfere with the regulatory landscape of nearby genes (Chow et al., 2010; Rebollo et al., 2011). Finally, the presence of LINE-1 sequences in a gene can attenuate its expression due to gradually pausing and/or dissociation of RNA polymerase from the template as it encounters longer adenosine-rich sequence stretches in the LINE-1 sequence (Han & Boeke, 2004; Han, Szak, & Boeke, 2004).

Considering the impact LINE-1 integration can have on the genome, it is no surprise that *de novo* LINE-1, *Alu* and SVA retrotransposition events in the germline or during embryonic development have been identified as the cause of up to 124 cases of hereditary single-gene diseases (reviewed in Hancks & Kazazian, 2016). Nonetheless, retrotransposition events are relatively rare in the human population, with an estimate LINE-1 retrotransposition rate of 1 heritable insertion in approximately 62 births (Ewing & Kazazian, 2010; Feusier et al., 2019). However, failure to regulate the expression of endogenous LINE-1 copies can also have negative impacts on cell function, even without retrotransposition taking place. Changes in gene expression and the production of non-coding RNAs, as well as excessive production of LINE-1 proteins and RNAs, can trigger undesirable cellular processes, including the activation of the innate immune response and initiation of apoptosis (Belgnaoui et al., 2006; Brégnard et al., 2016; Thomas et al., 2017; Wallace, Belancio, & Deininger, 2008). Additionally, L1-ORF2p can contribute to the formation of single/double strand breaks, leading to genome instability (Gasior et al., 2006). Thus, LINE-1 can have a major influence on the prognosis of disorders and diseases in which its regulation is affected.

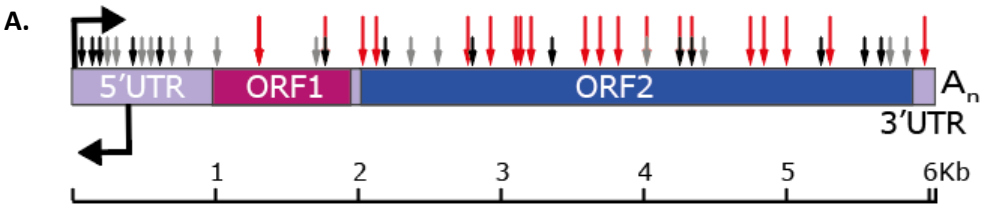


Figure 1.5. Schematic overview of the potential impact of LINE-1 integration on genomes – Legend in the next page.

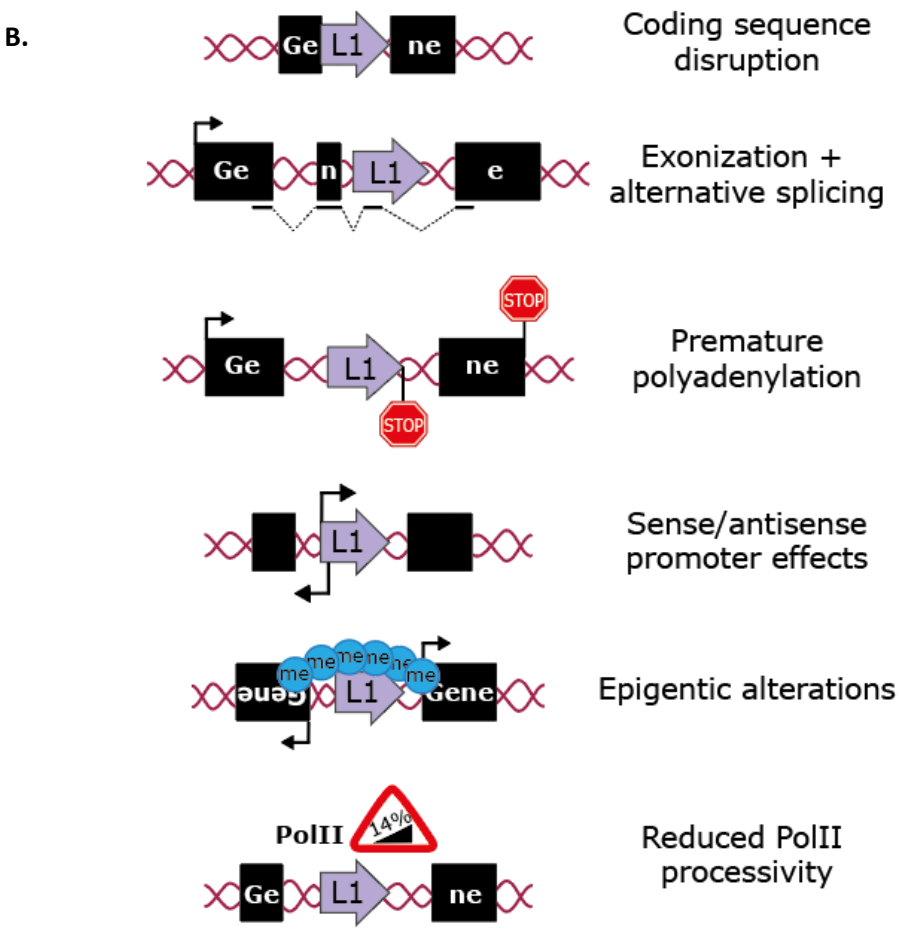


Figure 1.5. Schematic overview of the potential impact of LINE-1 integration on genomes – (A) [in the previous page] A schematic overview of predicted as well as experimentally validated SD (black arrows), SA (grey arrows) and polyadenylation signals (red arrows) present in the human LINE-1 sequence. Adapted from Belancio et al., 2006 (B) A schematic overview of the different ways in which LINE-1 integration can modify gene structure and expression. See text for detailed explanation.

1.3.2. LINE-1 in cancer

To date, there are four described cases in which a LINE-1 *de novo* insertion has been identified as the likely trigger for cancer development (reviewed in Scott & Devine, 2017): two cases of colon cancer, in which *de novo* insertions disrupted exons of the Adenomatous Polyposis Coli (*APC*) gene (Miki et al., 1992; Scott et al., 2016); a case of uterine cancer, by the interruption of an exon of the *PTEN* gene (Helman et al., 2014); and a case of liver cancer, caused by the affection of a regulatory sequence within an intron of the *ST18* gene (Shukla et al., 2013). The clearest example is one of the cases in colon cancer reported by Scott and colleagues, (i.e., an insertion in the *APC* gene). This tumour suppressor gene is mutated in 85% of colon cancer cases, with both alleles needing to be affected to trigger this type of cancer

(Bogearts & Prenen, 2014; Kinzler & Vogelstein, 1996). In the patient subject of this study, the insertion interrupting the exon 16 of one of the alleles of the *APC* gene was the second hit after a disruptive point mutation in the other allele of the same gene had already occurred (Scott et al., 2016). As discussed above (1.2.2), there are several reports documenting promoter hypomethylation of donor LINE-1s in healthy tissue adjacent to tumours (Nguyen et al., 2018; Schauer et al., 2018; Scott et al., 2016), suggesting that pre-tumour hypomethylation of certain L1 copies may pose a risk for tumour development.

While the above data demonstrate that LINE-1 retrotransposition has potential to induce cancer, a major conclusion from recent NGS studies is that the contribution of LINE-1s to cancer is most significant after tumour development (Bratthauer, Cardiff, & Fanning, 1994; Rodriguez-Martin et al., 2020). Most cancer types explored are characterised by genome-wide hypomethylation, including LINE-1 promoters, which correlates with their expression as well as retrotransposition levels (Alves, Tatro, & Fanning, 1996; Scott & Devine, 2017; Shukla et al., 2013; Tubio et al., 2014). So far, almost all somatic retrotransposition has been found in tumours of epithelial origin (Scott & Devine, 2017), with the highest activity in lung (Helman et al., 2014; Iskow et al., 2010) and colon (Ewing et al., 2015; Lee et al., 2012; Pitkänen et al., 2014; Scott et al., 2016; Solyom et al., 2012; Tubio et al., 2014), followed by tumours of the oesophagus (Doucet-O'Hare et al., 2015; Paterson et al., 2015), pancreas (Rodić et al., 2015), and ovary (Lee et al., 2012; Nguyen et al., 2018; Rodriguez-Martin et al., 2020; Tang et al., 2017). In these types of tumours, L1-ORF1p is typically also detectable by immunoassays, albeit at variable levels (De Luca et al., 2016; Rodić et al., 2014). This variation may be due to *p53* status, differences in promoter methylation, or expression of LINE-1 regulatory factors (1.2.2). In sum, the activation of LINE-1s in somatic tissues can have deleterious consequences for genome stability and therefore cancer progression (Daskalos et al., 2009). Consistently, LINE-1 promoter hypomethylation has been associated with poor prognosis (Iwagami et al., 2013; Ogino et al., 2008), as well as drug resistance, aggression and recurrence of certain cancers (Harada et al., 2015; Pattamadilok et al., 2008; van Hoesel et al., 2012). A noteworthy example is a case of ovarian cancer in which an insertion, potentially activating the chemoresistance related gene Stanniocalcin-1 (*STC1*), was found in greater copy number in the recurrent tumour after initial chemotherapy (Nguyen et al., 2018). This potentially reflects a selection for cells carrying the insertion during treatment. The fact that control mechanisms of LINE-1 are

attenuated in tumour cells, could provide a source of genetic diversity that promotes further tumour progression. Although more research is necessary, the detection of LINE-1 hypomethylation (Barchitta et al., 2014) and the presence of LINE-1 proteins (Burns, 2017) has been suggested as a potential biomarker for non-invasive screening and to assist in predicting clinical outcomes.

1.3.3. LINE-1 in other diseases

Activation of retrotransposons has also been observed in other human diseases, especially in disorders with a disposition towards autoimmune and neurological disorders. However, in most cases it is unclear to what extent these elements contribute to disease pathology and progression. Genetic disorders affecting retrotransposon regulatory factors, unsurprisingly, have been associated with an increase in LINE-1 activity and copy number. Here, I will discuss three examples: Rett syndrome (RTT), Ataxia Telangiectasia (AT) and Aicardi-Goutieres Syndrome (AGS); all neurological disorders characterised by neurodegeneration and containing an inflammatory component (Cortelazzo et al., 2014; Crow et al., 2014; Zaki-Dizaji et al., 2018).

RTT is caused by a mutation in the *MECP2* gene (Amir et al., 1999). This protein has been described to mediate epigenetic repression of LINE-1s under healthy conditions (Yu et al., 2001) (1.2.2). Indeed, human induced Pluripotent Stem Cells (iPSCs) derived from RTT patients were found to support increased LINE-1 retrotransposition levels (Muotri et al., 2010). Reciprocally, an increase in LINE-1 DNA copy numbers was found in post-mortem tissue from RTT patients compared to matched healthy controls (Zhao et al., 2019). The same was found for post-mortem samples from patients suffering from AT, a disease caused by autosomal recessive mutations inactivating the AT Mutated gene (*ATM*) (Coufal et al., 2011; Shiloh, 2001). Under healthy conditions, this protein is activated by the presence of double-strand DNA breaks, activating the DNA damage check-point and cell cycle arrest, either leading to DNA repair or p53-mediated apoptosis (Shiloh, 2001). Human Embryonic Stem Cells (ESCs) deficient for ATM, as well as ATM-KO mice, were found to facilitate more efficient LINE-1 retrotransposition, resulting in longer or perhaps more retrotransposition events per cell (Coufal et al., 2011). It has been proposed that more LINE-1 copies may be able to accumulate in the cell in the absence of functional ATM (Coufal et al., 2011), although further research is

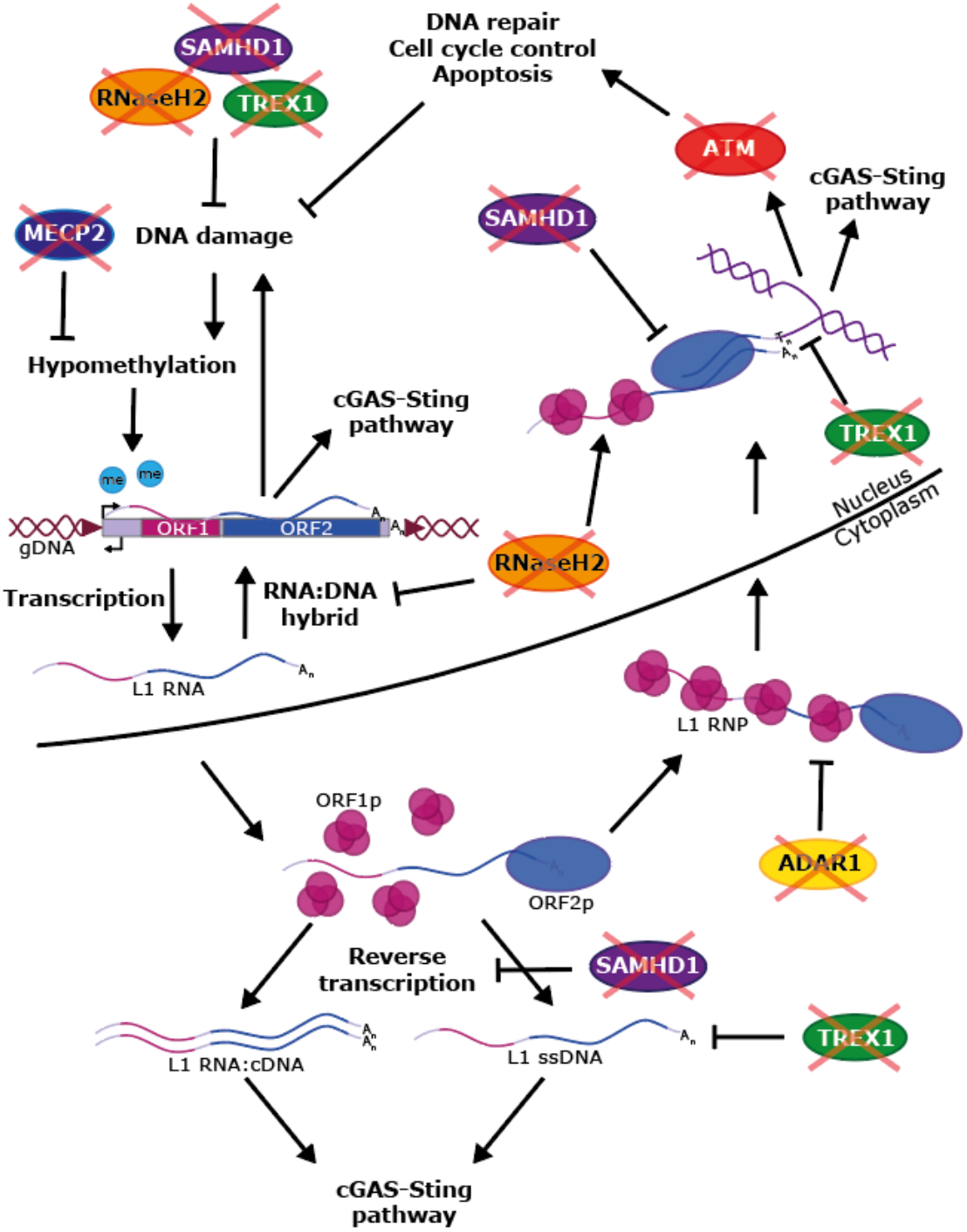


Figure 1.6. Schematic overview of how the LINE-1 lifecycle can be affected in RTT, AT and AGS – A schematic overview of how different proteins mutated in RTT (MECP2), AT (ATM) and AGS (RNase H2, SAMHD1, ADAR1 and TREX1) can affect the LINE-1 lifecycle, and how this can lead to activation of the cGAS-STING pathway, and consequentially the innate immune system. Based on the models proposed by Cecco, Lim, Thomas and Tunbak and colleagues (Cecco et al., 2019; Lim et al., 2015; Thomas et al., 2017; Tunbak et al., 2020).

needed to clarify the reciprocal influence of the failure of this DNA repair pathway and LINE-1 mobilisation. Finally, AGS can be traced to mutations in multiple genes of which at least 9 have been identified to date, *TREX1*, *ADAR1*, *SAMHD1*, the *RNASEH2* subunits A, B and C (forming the RNase H complex), Interferon Induced with Helicase C domain 1 (*IFIH1*), and most recently two members from the histone pre-mRNA processing machinery (*LSM11* and *RNU7-1*). Remarkably, of the proteins encoded in these genes, TREX1, ADAR1, SAMHD1 and the RNase H2 have been found to be involved in LINE-1 regulation (1.2.2). TREX1, ADAR1 and SAMHD1 have been shown to supress LINE-1 retrotransposition through targeting various LIINE-1 intermediates (Herrmann et al., 2018; Hu et al., 2015; Li et al., 2017; Orecchini et al., 2017; Zhao et al., 2013), while RNase H2 was found to facilitate LINE-1 integration (Benitez-Guijarro et al., 2018). Intriguingly, primary fibroblasts derived from AGS patients with various mutations showed a small but significant decrease in DNA methylation at the LINE-1 5'UTR, but no significant increase in LINE-1 transcripts compared to healthy cells (Lim et al., 2015). However, excessive loads of RNA:DNA hybrid R-loops were found with a particular enrichment in LINE-1 rich genomic regions. Additionally, pluripotent stem cells lacking TREX1 suffer from cytoplasmic accumulation of DNA species, most of which are single stranded DNAs derived from young LINE-1s (Thomas et al., 2017).

Although the contribution of LINE-1s to the symptomology of these disorders is unclear, a common feature is increased presence of LINE-1 derived nucleic acids. There is increasing evidence indicating that endogenous TEs (including LINE-1s) can be a major source of nucleic acids, triggering the innate immune system in many autoimmune diseases (reviewed in Volkman & Stetson, 2014). While provocative, this working model has clear parallelisms with processes activated upon viral infections. Under healthy conditions, viral DNA in the cytoplasm triggers activation of the cGAS-STING pathway, subsequently producing Interferons (IFNs) to initiate an inflammatory response (Thomas et al., 2017). Cytosolic and nuclear RNA:DNA hybrids (in the form of R-loops), have also been shown to trigger this pathway (Mankan et al., 2014; Weinreb et al., 2021). Hypomethylation of the LINE-1 promoter can lead to expression of functional and non-functional LINE-1s, which in turn can result in RNA:DNA hybrids at the transcription site (Lim et al., 2015), and double stranded RNA (dsRNA) in the cytoplasm (Tunbak et al., 2020) (***Figure 1.6***). Aberrant reverse transcription of the LINE-1 RNA in the cytoplasm can also result in RNA:DNA hybrids and single stranded DNA (ssDNA) side products (Cecco et al.,

2019; Thomas et al., 2017). Furthermore, LINE-1 mediated double stranded breaks have been shown to contribute to DNA damage (Gasior et al., 2006). All these features have individually been shown to activate the innate immune response through the cGAS-STING pathway (Cecco et al., 2019; Härtlova et al., 2015; Mankan et al., 2014; Weinreb et al., 2021). Therefore, this could provide a way for LINE-1 to contribute to these symptomologies without having to take into account the random nature of LINE-1 retrotransposition in the human genome (***Figure 1.6***).

1.4. Models in retrotransposon research

LINE-1 research comes with several complications and limitations, mostly derived from its repetitive nature and its abundance in mammalian genomes, directly or indirectly. Since ~17% of our DNA is made of LINE-1 copies, finding new LINE-1 insertions in genomes is equivalent to "finding a needle in a haystack". Furthermore, because LINE-1 copies are so similar to each other, identifying the donor of a *de novo* insertion is equally challenging. Additionally, LINE-1 sequences are widely spread in introns and are therefore transcribed as 'read-through', thus coupled to the expression of many genes. There is also evidence that LINE-1 RNAs are long-lived components of structural heterogeneous nuclear RNAs (hnRNAs) (Hall et al., 2014). These relatively abundant species of LINE-1 RNAs can obscure the presence of true LINE-1 transcripts originating from active elements (Belancio et al., 2008; Belancio et al., 2006; Skowronski, Fanning, & Singer, 1988). These and other limitations motivated Moran and colleagues to adapt a reporter assay, originally developed by Boeke and colleagues to study yeast LTR-retrotransposons (Boeke et al., 1985), for the study of human LINE-1s: the engineered LINE-1 retrotransposition reporter gene construct (Moran et al., 1996). This construct contains a full-length LINE-1 that is tagged in its 3'UTR with a retrotransposition indicator cassette, which consists of a reporter gene with its own promoter and polyadenylation signal located in antisense orientation with respect to the LINE-1, but interrupted by an intron in sense orientation (i.e., the LINE-1 orientation). This configuration ensures that the expression of the reporter gene will only occur upon completion of a full cycle of *bona fide* retrotransposition (***Figure 1.7***). Initially, these constructs were used in cell culture to test the retrotransposition competence of different genomic LINE-1 copies (Brouha et al., 2003; Moran et al., 1996; Sassaman et al., 1997), using an antibiotic resistance cassette that

leads to an assay in which the number of surviving colonies acts as a read-out of retrotransposition activity. However, the need for antibiotic selection, fixing and staining of cells to score retrotransposition was a limitation for certain more sensitive cell types, such as ESC or Neuronal Progenitor Cells (NPCs), and for its use *in vivo*. This motivated researchers to apply the same rationale in order to develop an Enhanced Green Fluorescent Protein (*EGFP*)- and a luciferase reporter-based assay (Ostertag, 2000; Xie et al., 2011). Although these different reporters have been successfully used to study retrotransposition, the original antibiotic resistance-based assays are the most sensitive to detect low levels of retrotransposition.

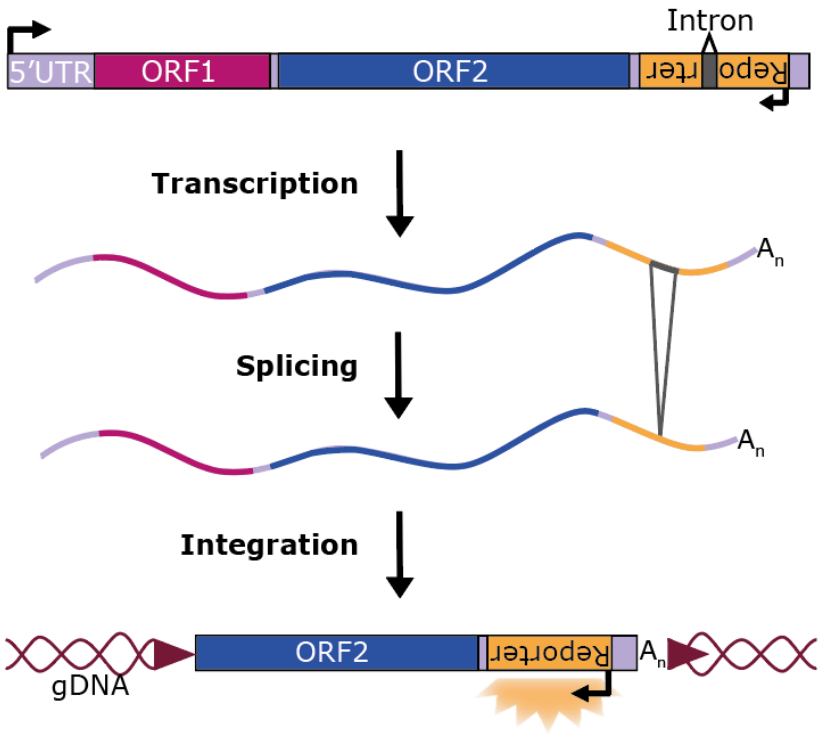


Figure 1.7. Rationale of the LINE-1 retrotransposition reporter construct assay – A schematic representation of the principle behind the LINE-1 retrotransposition reporter construct tagged with an undefined reporter, under control of its own promoter (depicted by black arrows), is shown. The reporter also contains its own polyadenylation sequence (not included in figure). See text for detailed explanation.

The development of the engineered LINE-1 mobilization assay has been instrumental to increase our knowledge of LINE-1 biology and their regulation. The assay was used to test all the different LINE-1 copies annotated in the first human genome reference draft, showing that only a few full-length copies were "hot" and could retrotranspose at a high level, and that punctual sequence changes can dramatically affect LINE-1 activity (Brouha et al., 2003;

Kimberland et al., 1999; Lutz et al., 2003; Ostertag, 2000; Seleme et al., 2006). Similarly, different cell types permissive to LINE-1 retrotransposition have been identified using this assay (i.e., cancer cells, NPCs and ESCs) (Coufal et al., 2009; Garcia-Perez et al., 2007; Moran et al., 1996). Combining engineered LINE-1 constructs with affinity proteomics, gene Knock-Down (KD) and Knock-Out (KO) strategies, also allowed the identification of different host cellular factors involved in LINE-1 regulation (Coufal et al., 2011; Hu et al., 2015; Taylor et al., 2013; Thomas et al., 2017). Even though this assay was established >25 years ago, it still remains one of the gold-standard tools used in this field of research.

Notably, in 2002 the Kazazian lab generated the first mouse model to study retrotransposition *in vivo*, using engineered L1 vectors. This mouse model was instrumental to uncover cellular niches with ongoing LINE-1 retrotransposition in higher organisms. In this way, it was demonstrated that an engineered human LINE-1 element could retrotranspose in the mouse male germline, before the onset of meiosis II (Ostertag et al., 2002; Prak et al., 2003). However, a limitation of this mouse model was the use of a sperm cell specific promoter (the acrosin promoter), to drive *EGFP* expression upon insertion. Consistently, analyses of brain and other tissues did not reveal any *EGFP* expression (Prak et al., 2003). A later transgenic mouse model, containing an engineered LINE-1 which used a pan-active CMV promoter to drive reporter (*EGFP*) expression, confirmed retrotransposition in the germ line, as well as identified retrotransposition in several areas of the rodent brain, enriched in NPCs (Muotri et al., 2005). This seminal study, which demonstrated that LINE-1s have the potential to drive genomic mosaicism in brain, was followed by other studies revealing ongoing LINE-1 activity in neurons, particularly in the hippocampus as well as the cerebellum (Muotri et al., 2009). Without a doubt, these models are key to uncover novel aspects of LINE-1 biology, but they are also limited in their capabilities. Indeed, the use of a sperm promoter to drive *EGFP* prevented Kazazian and colleagues from exploring retrotransposition in other tissues, such as brain, demonstrating how the chosen promoter has a large influence on the cell type in which the reporter can be expressed due to cell specific regulation.

Notably, the development and cost reduction of high throughput whole-genome sequencing, complements research with these artificial set-ups. High coverage deep sequencing of single cells (Evrony et al., 2012), sequencing coupled with enrichment techniques (Baillie et al., 2011; Erwin et al., 2016; Ewing & Kazazian, 2010) or a combination of

both (Evrony et al., 2015; Sanchez-Luque et al., 2019), now makes it possible to analyse the mobilisation of endogenous LINE-1s. Nonetheless, animal models are still a valuable resource to further uncover the role LINE-1 plays in disease development, as they can allow an *in vivo* window into the pre-symptomatic phase of a disease, something impossible in post-mortem human tissue and/or cell lines. Thus, having access to animal models containing LINE-1 elements comparable/analogous to human LINE-1s can be of tremendous value. Although there are certain key differences between LINEs across species, the mechanism of retrotransposition as well as other features are largely conserved, and the obtained knowledge is partially transferable (reviewed in Bodak, Yu, & Ciaudo, 2014). So far, significant research has been carried out in mice, specifically exploring endogenous mouse LINE-1s, providing insights into the effects of LINE-1 insertions on the genome (Rebollo et al., 2011), LINE-1 regulation in various cell types (Trelogan & Martin, 1995; Xu et al., 2008) and the frequency of heritable retrotransposition in the germline (Richardson et al., 2017). Active LINE-2 elements have also been described in zebrafish (Sugano et al., 2006), and our lab has been using this model for several years to get a better understanding of how the activity and regulation of these elements compares to mammalian LINE-1s (unpublished data). More recently, 17 different LINE-1 subfamilies were described to be present in the zebrafish genome (Boissinot & Sookdeo, 2016), likely to be active (Duvernell, Pryor, & Adams, 2004). Although it remains to be confirmed that any of these 17 subfamilies still contain retrotransposition competent copies, it is a very exciting discovery as it might expand the options of animal models in which the biology of endogenous LINE-1s can be investigated. In this thesis, I will explore these recently described zebrafish LINE-1 subfamilies to uncover their potential to serve as a model for human LINE-1 research.

1.5. Aims and objectives

This thesis has two distinct aims:

(1) get a better understanding of the role of RNase H2 and, indirectly, PCNA in the retrotransposition cycle of human LINE-1s; and
(2) determine whether there are active LINE-1 copies in the zebrafish genome.

To address the first aim (1), I particularly focused on:

- The involvement of the PIP motif of RNase H2 and L1-ORF2p in TPRT.
- The level of RNase H2 activity necessary to support efficient LINE-1 retrotransposition.
- Determining whether introducing an exogenous RNase H domain in the human L1-ORF2p eliminates its dependence on host-encoded RNase H2.

To address the second aim (2), I performed analyses of the 17 pre-described LINE-1 subfamilies in the zebrafish genome to determine:

- The presence of ZfL1 copies in the zebrafish reference genome with the potential to encode the machinery required for retrotransposition, and the analysis of their expression levels along zebrafish development (*in silico*).
- The recovery of these copies from DNA obtained from the zebrafish individuals from the TU strain; the engineering of expression and retrotransposition reporters; and the assessment of their retrotransposition potential in *in vitro* and *in vivo* experimental settings.

Chapter 2

Materials and methods

2. Materials and methods

2.1. General reagents

2.1.1. Sources of reagents (chemicals, enzymes, culture media, antibodies)

Chemicals were purchased from Amersham Biosciences (GE Healthcare), BDH Laboratory Supplies (AnalaR, VWR), Fisher Chemicals, and Sigma Aldrich. Enzymes were obtained from New England Biolabs, Promega and Roche. Cell culture material was purchased from Gibco (Invitrogen) or Sigma Aldrich unless stated otherwise. Antibodies were purchased from Abcam or Cell Signaling.

2.1.2. Preparation of buffer solutions

All commonly used buffers (***Table 2.1***) were made using molecular grade water such as Milli-Q water (purified by filtration and deionization to resistivity of 18.2 MΩ cm at 25C in a Millipore Corporation water filtration station) or Invitrogen Ultrapure Water (distilled and filtered by 0.1 µm membrane filters). Solutions were sterilised in house, by autoclaving at 121°C for 15 min. Solutions that could not be autoclaved were filtered using 0.22 µm filter (Millipore).

Buffer	Composition
1× Phosphate-Buffered Saline (PBS)	137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4
1× Lysis buffer	50 mM Tris (pH 8.8), 1 mM EDTA, 0.5% Tween
1× Urea lysis buffer	8 M urea, 50 mM Tris (pH 7.5), 150 mM β-mercaptoethanol, 1× complete mini EDTA-free protease inhibitor cocktail (Roche, Cat. No. 04693159001)
1× Whole Cell Extract (WCE) buffer	50 mM Tris (pH 8), 280 mM NaCl, 0.5% NP40, 0.2 mM EDTA, 0.2 mM EGTA, 10% glycerol, 0.1 mM Na_3VO_4
1× Cytoplasmic buffer	20 mM HEPES (pH 7.9), 10 mM KCl, 1 mM EDTA, 0.1 mM Na_3VO_4, 10% glycerol
10× Tris-glycine SDS-PAGE running buffer	250 mM Tris-base, 1.92 M glycine, 1% (w/v) SDS
1× Immunoblotting transfer buffer	1X Tris-glycine SDS-PAGE running buffer, 20% (v/v) methanol
4× SDS protein sample loading buffer	0.5 M Tris-HCl (pH 6.8), 50% (v/v) glycerol, 2% (w/v) SDS, 0.1% (w/v) bromophenol blue
1× Tris EDTA (TE)	10 mM Tris-HCl (pH 8), 1 mM EDTA
10× Tris Buffered Saline (TBS)	0.5 M Tris base, 1.5 M NaCl (pH adjusted to 7.5 with HCl)
1× TBS-T	1X TBS, 2% (v/v) Tween-20

Continues in the next page.

Buffer	Composition
Blocking buffer	1X TBS-T, 5% (w/v) powdered skimmed milk (Marvel)
20× Tris/Borate/EDTA (TBE)	1.78 M Tris-base, 1.78 M H_3BO_3, 40 mM EDTA
1× Transformation Buffer (TB)	10 mM PIPES-HCl pH 6.7, 15 mM $CaCl_2$, 0.25 M KCl, 55 mM $MnCl_2$
60× E3	5 mM NaCl, 0.17mM KCl, 0.33 mM CaCl2, 0.33 mM $MgSo_4$, 10-5% Methylene Blue

Table 2.1. Commonly used buffers.

2.1.3. Preparation of cell culture drug stock solutions

Drugs were added to tissue culture media immediately prior to use, and used at the working concentrations indicated in ***Table 2.2***. Stock solutions were purchased, or made from powder dissolved in 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) or autoclaved molecular grade H_2O, in tissue culture hoods, according to manufacturers' instructions. After aliquoting, stock solutions were stored at 4˚C or –20˚C as indicated by the manufacturers.

Drug	Solvent	Stock conc.	Work conc.	Manufacturer (Cat. No.)	Stored at
Puromycin Dihydrochloride	H_2O	10 mg/mL	2 µg/mL	Gibco (A1113803)	–20˚C
Geneticin™ (G418 Sulphate)	H_2O	50 mg/mL	600 µg/mL	Gibco (10131035)	4˚C
Blasticidin S HCL	HEPES	10 mg/mL	5 µg/mL	Gibco (A1113903)	–20˚C
Zeocin	H_2O	100 mg/mL	150 µg/mL	Gibco (R25001)	–20˚C
Hygromycin B	HEPES	50 mg/mL	200 µg/mL	Millipore (400050)	4˚C

Table 2.2. Cell culture drug stock solutions.

2.1.4. DNA vectors

DNA vectors used in this study are listed in ***Table 2.3*** and ***Table 2.4.*** Those that were created for this work (***Table 2.4***) were constructed using standard cloning techniques (2.4) and oligonucleotides described in Appendix (6.1).

Identifier	Description	Source
pCEP4-L1.3-*mneoI* (JM101)	pCEP4 (Life Technologies) containing a full-length human Wild-Rype (WT) L1.3 [accession number L19088.1; (Dombroski, Scott, & Kazazian, 1993)] downstream of the Cytomegalovirus immediate-early promoter (CMVp), tagged with an antisense, intron-interrupted neomycin-phosphotransferase gene retrotransposition indicator cassette (*mneoI*) driven by a Simian Virus 40 (SV40) promoter (Moran et al., 1996). A downstream SV40 late polyadenylation signal facilitates expression.	JV. Moran

Continues in the next page.

Identifier	Description	Source
pCEP4-L1.3-RT⁻-*mneoI* (JM105)	pCEP4-L1.3-*mneoI* containing a missense mutation in the RT domain of the L1.3-ORF2 (D702A). Used as a negative control in Neomycin-based retrotransposition assays (Moran et al., 1996).	JV. Moran
pT2neo	Vector containing SV40p-driven neomycin phosphotransferase gene flanked by Sleeping Beauty TIRs (Mátés et al., 2009).	Z. Ivics
pMSCV-Zeo (pJKN255)	Retroviral Murine Stem Cell Virus (MSCV) vector, optimized for integration and stable expression of a gene of interest, containing the Zeocin resistance gene (Addgene number: 75088. (Kendall et al., 2007)).	MAM. Reijns
pMSCV-RNH2B-Zeo (pMAR761)	pMSCV-Zeo-DEST (pMAR751) containing the human *RNASEH2B* (CCDS9425.1) coding sequence in the MSCV vector with Zeocin resistance. Gateway compatible version of pMSCV-Zeo (pJKN255) optimized for integration and stable expression of a gene of interest.	MAM. Reijns
pMSCV-RNH2B-PIP-Zeo (pMAR762)	pMSCV-RNH2B-Zeo containing 2 missense mutations in the *RNASEH2B* PIP domain (F300A/F301A) abolishing its interaction with PCNA (Bubeck et al., 2011).	MAM. Reijns
pMSCV-puro-EGFP (pMAR445)	pMSCV-Puro vector containing retroviral MSCV vector with puromycin resistance and optimized for integration and stable expression of a gene of interest, with the coding sequence of the EGFP containing an upstream Kozak sequence from pEGFP-N1 (ClonTech), cloned within the MSCV.	MAM. Reijns
GAG-Pol	GAG-Pol expressing vector, optimized for y-Retroviral packaging system.	MAM. Reijns
pMD2.G (pJKN145)	VSV-G envelope expressing vector, optimized for y-Retroviral packaging system.	MAM. Reijns
pCEP4-L1.3-*mblastI* (JJ101)	pCEP4 (Life Technologies) containing a full-length human WT L1.3 (Dombroski et al., 1993) downstream of the CMVp, tagged with an intron-interrupted blasticidin-resistance gene retrotransposition indicator cassette (*mblastI*) (Morrish et al., 2002). A downstream SV40 late polyadenylation signal facilitates expression.	JV. Moran
pCEP4-L1.3-RT⁻-*mblastI* (JJ116)	pCEP4-L1.3-*mblastI* containing a missense mutation in the RT domain of the L1.3-ORF2 (D702A). Used as a negative control in Blasticidin-based retrotransposition assays (Morrish et al., 2002).	JV. Moran
pCEP4-L1.3-PIP6-*mblastI* (JJ107)	pCEP4-L1.3-*mblastI* containing two missense mutations in the PIP domain of the L1.3-ORF2 (Y414A/Y415A).	JL. Garcia-Perez
pcDNA6.1	Vector containing an expression cassette for blasticidin S deaminase.	Invitrogen
pENTRY-E.coli rnhA Kozak (pMAR673)	Gateway Donor vector pDONR221 with the gateway-cloned coding sequence of Stbl3 *E. coli RNASEHI* (rnhA) containing a Kozak sequence.	MAM. Reijns
pENTRY-hRNASE H1 nuclear Kozak (pMAR676)	Gateway Donor vector pDONR221 with the gateway-cloned coding sequence of human nuclear *RNASEH1* containing a Kozak sequence.	MAM. Reijns

Continues in the next page.

Identifier	Description	Source
pENTRY-E.coli rnhA D70N Kozak (pMAR683)	pENTRY-E.coli RNHA Kozak containing a missense mutation in one of the conserved catalytic sites of *RNASEHI* (D70N).	MAM. Reijns
pENTRY-hRNASE H1 nuclear D119N Kozak (pMAR684)	pENTRY-hRNASE H1 nuclear Kozak containing a missense mutation in one of the conserved catalytic sites of *RNASEH1* (D119N).	MAM. Reijns
pGEM-T-ORF2C-OR	pGEM-T containing the C-terminal region of L1.3-ORF2 fused in-frame to an Oestrogen Receptor (OR) at the C-terminus.	S. Morell
pTOL2-Zfl2-2-EGFP-SL (ZfL-2-2-WT)	Modified pTOL2 transgenesis vector (Bessa et al., 2014) containing an active copy of a monomer-lacking ZfL2-2 (Sugano et al., 2006) downstream of a T7 promoter, tagged with an antisense *EGFP* expression cassette driven by a CMVp, and the ZfL2-2 3'UTR (upstream of the conserved stem-loop (SL) and microsatellite-like repeat (TGTAAAx3) necessary for LINE-2 retrotransposition).	T. Widmann (***Figure 6.7***)
pTOL2-Zfl2-2-EGFP-NSL	pTOL2-Zfl2-2-EGFP-SL lacking the conserved stem-loop (NSL) and microsatellite-like repeat (TGTAAAx3) necessary for LINE-2 retrotransposition.	T. Widmann

Table 2.3. Vectors used in this study that were gifted or previously generated by our team.

Identifier	Description	Figure
pCEP4-L1.3-PIP6-*mneoI* (JM101)	pCEP4-L1.3-*mneoI* (***Table 2.3***) containing two missense mutations in the PIP domain of the *L1.3-ORF2* (Y414A/Y415A) (See JJ107).	N/A
pCEP4-L1.3-eRH-*mneoI* (pMK001)	pCEP4-L1.3-*mneoI* (***Table 2.3***) containing the coding sequence of *E. coli RNASEHI* (amplified from pENTRY-E.coli rnhA Kozak), without Kozak sequence, fused downstream and in-frame to the C-terminus of L1.3-ORF2p by a linker sequence composed of 5× Glycine-Serine (GS).	Figure 6.1
pCEP4-L1.3-eRHm-*mneoI* (pMK002)	pCEP4-L1.3-*mneoI* (***Table 2.3***) containing the coding sequence of the catalytic dead *E. coli RNASEHI* (amplified from pENTRY-E.coli rnhA D70N Kozak), without Kozak sequence, fused downstream and in-frame to the C-terminus of L1.3-ORF2p by a linker sequence composed of 5× GS.	Figure 6.1
pCEP4-L1.3-hRH-*mneoI* (pMK003)	pCEP4-L1.3-*mneoI* (***Table 2.3***) containing the catalytic core of human nuclear *RNASEH1* (amplified from pENTRY-hRNASE H1 nuclear Kozak) fused downstream and in-frame to the C-terminus of L1.3-ORF2p by a linker sequence composed of 5× GS.	Figure 6.1
pCEP4-L1.3-hRHm-*mneoI* (pMK004)	pCEP4-L1.3-*mneoI* (***Table 2.3***) containing the catalytic core of the catalytic dead human nuclear *RNASEH1* (amplified from pENTRY-hRNASE H1 nuclear D119N Kozak) fused downstream and in-frame to the C-terminus of L1.3-ORF2p by a linker sequence composed of 5× GS.	Figure 6.1
pCEP4-L1.3-RT^{-}-eRH-*mneoI* (pMK005)	pCEP4-L1.3-eRH-*mneoI* (pMK001) containing the L1.3 with a missense mutation in the RT domain of the *L1.3-ORF2* (D702A).	Figure 6.1

Continues in the next page.

Identifier	Description	Figure
pCEP4-L1.3-RT^{-}-eRHm-*mneoI* (pMK006)	pCEP4-L1.3-eRHm-*mneoI* (pMK002) containing the L1.3 with a missense mutation in the RT domain of the *L1.3-ORF2* (D702A).	Figure 6.1
pCEP4-L1.3-RT^{-}-hRH-*mneoI* (pMK007)	pCEP4-L1.3-hRH-*mneoI* (pMK003) containing the L1.3 with a missense mutation in the RT domain of the *L1.3-ORF2* (D702A).	Figure 6.1
pCEP4-L1.3-RT^{-}-hRHm-*mneoI* (pMK008)	pCEP4-L1.3-hRHm-*mneoI* (pMK004) containing the L1.3 with a missense mutation in the RT domain of the *L1.3-ORF2* (D702A).	Figure 6.1
pCEP4-L1.3-PIPm-eRH-*mneoI* (pMK009)	pCEP4-L1.3-eRH-*mneoI* (pMK001) containing the L1.3 with two missense mutations in the PIP domain of the *L1.3-ORF2* (Y414A/Y415A).	Figure 6.1
pCEP4-L1.3-PIPm-eRHm-*mneoI* (pMK010)	pCEP4-L1.3-eRHm-*mneoI* (pMK002) containing the L1.3 with two missense mutations in the PIP domain of the *L1.3-ORF2* (Y414A/Y415A).	Figure 6.1
pCEP4-L1.3-PIPm-hRH-*mneoI* (pMK011)	pCEP4-L1.3-hRH-*mneoI* (pMK003) containing the L1.3 with two missense mutations in the PIP domain of the *L1.3-ORF2* (Y414A/Y415A).	Figure 6.1
pCEP4-L1.3-PIPm-hRHm-*mneoI* (pMK012)	pCEP4-L1.3-hRHm-*mneoI* (pMK004) containing the L1.3 with two missense mutations in the PIP domain of the *L1.3-ORF2* (Y414A/Y415A).	Figure 6.1
pTOL2-Zfl2-2-RT^{-}-EGFP-SL (pMK013)	pTOL2-Zfl2-2-EGFP-SL (***Table 2.3***) containing a missense mutation in the RT domain of the *zfL2-2-orf* (D689Y), abolishing retrotransposition competency (Sugano et al., 2006). Used as a negative control in EGFP-based retrotransposition assays.	Figure 6.7
pTOL2-Zfl1-7B-Chr4-EGFP (pMK014)	pTOL2-Zfl2-2-EGFP-NSL (***Table 2.3***) containing the ZfL1-7B element at Chr4:45903129-45908934[1] (amplified from Tübingen (TU) zebrafish strain) instead of the ZfL2-2 element (constructed using PCR and cloning).	Figure 6.4
pTOL2-ZfL1-7B-Chr5-EGFP (pMK015)	pTOL2-Zfl2-2-EGFP-NSL (***Table 2.3***) containing the ZfL1-7B element located at Chr5:41057392-41063181[1] (amplified from TU zebrafish strain) instead of the ZfL2-2 element (constructed using PCR and cloning).	Figure 6.4
pTOL2-ZfL1-7B-Chr6-EGFP (pMK016)	pTOL2-Zfl2-2-EGFP-NSL (***Table 2.3***) containing the ZfL1-7B element at Chr6:18504003-18509793[1] (amplified from TU zebrafish strain) instead of the ZfL2-2 element (constructed using PCR and cloning).	Figure 6.4
pTOL2-ZfL1-7B-Chr24-EGFP (pMK017)	pTOL2-Zfl2-2-EGFP-NSL (***Table 2.3***) containing the ZfL1-7B element at Chr24:31415464-31421258[1] (amplified from TU zebrafish strain) instead of the ZfL2-2 element (constructed using PCR and cloning).	Figure 6.4
pTOL2-ZfL1-7B-Cns-EGFP (pMK018)	pTOL2-Zfl2-2-EGFP-NSL (***Table 2.3***) containing the ZfL1-7B functional consensus sequence (constructed using PCR and cloning) instead of the ZfL2-2 element (constructed using PCR and cloning).	Figure 6.4

Continues in the next page.

Identifier	Description	Figure
pTOL2-ZfL1-10B-Chr5-EGFP (pMK019)	pTOL2-Zfl2-2-EGFP-NSL (***Table 2.3***) containing the ZfL1-10B element at Chr5:42036670-42042519[1] (amplified from TU zebrafish strain) instead of the ZfL2-2 element (constructed using PCR and cloning).	Figure 6.4
pTOL2-ZfL1-10B-Chr8-EGFP (pMK020)	pTOL2-Zfl2-2-EGFP-NSL (***Table 2.3***) containing the ZfL1-10B element at Chr8:24020995-24026840[1] (amplified from TU zebrafish strain) instead of the ZfL2-2 element (constructed using PCR and cloning).	Figure 6.4
pTOL2-ZfL1-10B-Chr18-EGFP (pMK021)	pTOL2-Zfl2-2-EGFP-NSL (***Table 2.3***) containing the ZfL1-10B element at Chr18:32301158-32307002[1] (amplified from TU zebrafish strain) instead of the ZfL2-2 element (constructed using PCR and cloning).	Figure 6.4
pTOL2-ZfL1-10B-Cns-EGFP (pMK022)	pTOL2-Zfl2-2-EGFP-NSL (***Table 2.3***) containing the ZfL1-10B functional consensus sequence (constructed using PCR and cloning) instead of the ZfL2-2 element (constructed using PCR and cloning).	Figure 6.4
pTOL2-ZfL1-12B-Chr19-EGFP (pMK023)	pTOL2-Zfl2-2-EGFP-NSL (***Table 2.3***) containing the ZfL1-12B element at Chr19:12893436-12899204[1] (amplified from TU zebrafish strain) instead of the ZfL2-2 element (constructed using PCR and cloning).	Figure 6.4
pTOL2-ZfL1-12B-Cns-EGFP (pMK024)	pTOL2-Zfl2-2-EGFP-NSL (***Table 2.3***) containing the ZfL1-12B functional consensus sequence (constructed using PCR and cloning) instead of the ZfL2-2 element (constructed using PCR and cloning).	Figure 6.4
pTOL2-ZfL1-7B-Chr4-EGFP-ZfL1-7B-3'UTR (pMK025)	pTOL2-Zfl1-7B-Chr4-EGFP (pMK014) with the antisense EGFP expression cassette moved upstream of the ZfL1-7B 3'UTR (constructed using PCR and cloning).	Figure 6.5
pTOL2-ZfL1-7B-Chr5-EGFP-ZfL1-7B-3'UTR (pMK026)	pTOL2-ZfL1-7B-Chr5-EGFP (pMK015), with the antisense EGFP expression cassette moved upstream of the ZfL1-7B 3'UTR (constructed using PCR and cloning).	Figure 6.5
pTOL2-ZfL1-7B-Chr6-EGFP-ZfL1-7B-3'UTR (pMK027)	pTOL2-ZfL1-7B-Chr6-EGFP (pMK016), with the antisense EGFP expression cassette moved upstream of the ZfL1-7B 3'UTR (constructed using PCR and cloning).	Figure 6.5
pTOL2-ZfL1-7B-Chr24-EGFP-ZfL1-7B-3'UTR (pMK028)	pTOL2-ZfL1-7B-Chr24-EGFP (pMK017), with the antisense EGFP expression cassette moved upstream of the ZfL1-7B 3'UTR (constructed using PCR and cloning).	Figure 6.5
pTOL2-ZfL1-7B-Cns-EGFP-ZfL1-7B-3'UTR (pMK029)	pTOL2-ZfL1-7B-Cns-EGFP (pMK018), with the antisense EGFP expression cassette moved upstream of the ZfL1-7B 3'UTR (constructed using PCR and cloning).	Figure 6.5
pTOL2-ZfL1-10B-Chr5-EGFP-ZfL1-10B-3'UTR (pMK030)	pTOL2-ZfL1-10B-Chr5-EGFP (pMK019), with the antisense EGFP expression cassette moved upstream of the ZfL1-10B 3'UTR (constructed using PCR and cloning).	Figure 6.5
pTOL2-ZfL1-10B-Chr8-EGFP-ZfL1-10B-3'UTR (pMK031)	pTOL2-ZfL1-10B-Chr8-EGFP (pMK020), with the antisense EGFP expression cassette moved upstream of the ZfL1-10B 3'UTR (constructed using PCR and cloning).	Figure 6.5
pTOL2-ZfL1-10B-Chr18-EGFP-ZfL1-10B-3'UTR (pMK032)	pTOL2-ZfL1-10B-Chr18-EGFP (pMK021), with the antisense EGFP expression cassette moved upstream of the ZfL1-10B 3'UTR (constructed using PCR and cloning).	Figure 6.5

Continues in the next page.

Identifier	Description	Figure
pTOL2-ZfL1-10B-Cns-EGFP-ZfL1-10B-3'UTR (pMK033)	pTOL2-ZfL1-10B-Cns-EGFP (pMK022), with the antisense EGFP expression cassette moved upstream of the ZfL1-10B 3'UTR (constructed using PCR and cloning).	Figure 6.5
pTOL2-ZfL1-12B-Chr19-EGFP-ZfL1-12B-3'UTR (pMK034)	pTOL2-ZfL1-12B-Chr19-EGFP (pMK023), with the antisense EGFP expression cassette moved upstream of the ZfL1-12B 3'UTR (constructed using PCR and cloning).	Figure 6.5
pTOL2-ZfL1-12B-Cns-EGFP-ZfL1-12B-3'UTR (pMK035)	pTOL2-ZfL1-12B-Cns-EGFP (pMK024), with the antisense EGFP expression cassette moved upstream of the ZfL1-12B 3'UTR (constructed using PCR and cloning).	Figure 6.5
pCEP4-ZfL1-7B-Chr4-*mneoI* (pMK036)	pCEP4-L1.3-*mneoI* (***Table 2.3***) containing the ZfL1-7B element at Chr4:45903129-45908934[1] (amplified from TU zebrafish strain) instead of the L1.3 element (constructed using PCR and cloning).	Figure 6.2
pCEP4-ZfL1-7B-Chr5-*mneoI* (pMK037)	pCEP4-L1.3-*mneoI* (***Table 2.3***) containing the ZfL1-7B element located on Chr5:41057392-41063181[1] (amplified from a TU strain zebrafish) instead of the L1.3 element (constructed using PCR and cloning).	Figure 6.2
pCEP4-ZfL1-7B-Chr6-*mneoI* (pMK038)	pCEP4-L1.3-*mneoI* (***Table 2.3***) containing the ZfL1-7B element located on Chr6:18504003-18509793[1] (amplified from a TU strain zebrafish) instead of the L1.3 element (constructed using PCR and cloning).	Figure 6.2
pCEP4-ZfL1-7B-Chr24-*mneoI* (pMK039)	pCEP4-L1.3-*mneoI* (***Table 2.3***) containing the ZfL1-7B element located on Chr24:31415464-31421258[1] (amplified from a TU strain zebrafish) instead of the L1.3 element (constructed using PCR and cloning).	Figure 6.2
pCEP4-ZfL1-7B-Cns-*mneoI* (pMK040)	pCEP4-L1.3-*mneoI* (***Table 2.3***) containing a ZfL1-7B element with the functional consensus sequence of the ZfL1-7B family (constructed using PCR and cloning) instead of the L1.3 element (constructed using PCR and cloning).	Figure 6.2
pCEP4-ZfL1-10B-Chr5-*mneoI* (pMK041)	pCEP4-L1.3-*mneoI* (***Table 2.3***) containing the ZfL1-10B element located on Chr5:42036670-42042519[1] (amplified from a TU strain zebrafish) instead of the L1.3 element (constructed using PCR and cloning).	Figure 6.2
pCEP4-ZfL1-10B-Chr8-*mneoI* (pMK042)	pCEP4-L1.3-*mneoI* (***Table 2.3***) containing the ZfL1-10B element located on Chr8:24020995-24026840[1] (amplified from a TU strain zebrafish) instead of the L1.3 element (constructed using PCR and cloning).	Figure 6.2
pCEP4-ZfL1-10B-Chr18-*mneoI* (pMK043)	pCEP4-L1.3-*mneoI* (***Table 2.3***) containing the ZfL1-10B element located on Chr18:32301158-32307002[1] (amplified from a TU strain zebrafish) instead of the L1.3 element (constructed using PCR and cloning).	Figure 6.2
pCEP4-ZfL1-10B-Cns-*mneoI* (pMK044)	pCEP4-L1.3-*mneoI* (***Table 2.3***) containing a ZfL1-10B element with the functional consensus sequence of the ZfL1-10B family (constructed using PCR and cloning) instead of the L1.3 element (constructed using PCR and cloning).	Figure 6.2

Continues in the next page.

Identifier	Description	Figure
pCEP4-ZfL1-12B-Chr19-*mneoI* (pMK045)	pCEP4-L1.3-*mneoI* (***Table 2.3***) containing the ZfL1-12B element located on Chr19:12893436-12899204[1] (amplified from a TU strain zebrafish) instead of the L1.3 element (constructed using PCR and cloning).	Figure 6.2
pCEP4-ZfL1-12B-Cns-*mneoI* (pMK046)	pCEP4-L1.3-*mneoI* (***Table 2.3***) containing a ZfL1-12B element with the functional consensus sequence of the ZfL1-12B family (constructed using PCR and cloning instead of the L1.3 element (constructed using PCR and cloning).	Figure 6.2
pCEP4-ZfL1-7B-Chr4-*mneoI*-3'UTR (pMK047)	pCEP4-ZfL1-7B-Chr4-*mneoI* (pMK036), with the antisense *mneoI* moved upstream of the ZfL1-7B 3'UTR (constructed using PCR and cloning).	Figure 6.3
pCEP4-ZfL1-7B-Chr5-*mneoI*-3'UTR (pMK048)	pCEP4-ZfL1-7B-Chr5-*mneoI* (pMK037), with the antisense *mneoI* moved upstream of the ZfL1-7B 3'UTR (constructed using PCR and cloning).	Figure 6.3
pCEP4-ZfL1-7B-Chr6-*mneoI*-3'UTR (pMK049)	pCEP4-ZfL1-7B-Chr6-*mneoI* (pMK038), with the antisense *mneoI* moved upstream of the ZfL1-7B 3'UTR (constructed using PCR and cloning).	Figure 6.3
pCEP4-ZfL1-7B-Chr24-*mneoI*-3'UTR (pMK050)	pCEP4-ZfL1-7B-Chr24-*mneoI* (pMK039), with the antisense *mneoI* moved upstream of the ZfL1-7B 3'UTR (constructed using PCR and cloning).	Figure 6.3
pCEP4-ZfL1-7B-Cns-*mneoI*-3'UTR (pMK051)	pCEP4-ZfL1-7B-Cns-*mneoI* (pMK040), with the antisense *mneoI* moved upstream of the ZfL1-7B 3'UTR (constructed using PCR and cloning).	Figure 6.3
pCEP4-ZfL1-10B-Chr5-*mneoI*-3'UTR (pMK052)	pCEP4-ZfL1-10B-Chr5-*mneoI* (pMK041), with the antisense *mneoI* moved upstream of the ZfL1-10B 3'UTR (constructed using PCR and cloning).	Figure 6.3
pCEP4-ZfL1-10B-Chr8-*mneoI*-3'UTR (pMK053)	pCEP4-ZfL1-10B-Chr8-*mneoI* (pMK042), with the antisense *mneoI* moved upstream of the ZfL1-10B 3'UTR (constructed using PCR and cloning).	Figure 6.3
pCEP4-ZfL1-10B-Chr18-*mneoI*-3'UTR (pMK054)	pCEP4-ZfL1-10B-Chr18-*mneoI* (pMK043), with the antisense *mneoI* moved upstream of the ZfL1-10B 3'UTR (constructed using PCR and cloning).	Figure 6.3
pCEP4-ZfL1-10B-Cns-mneoI-3'UTR (pMK055)	pCEP4-ZfL1-10B-Cns-*mneoI* (pMK044), with the antisense *mneoI* moved upstream of the ZfL1-10B 3'UTR (constructed using PCR and cloning).	Figure 6.3
pCEP4-ZfL1-12B-Chr19-*mneoI*-3'UTR (pMK056)	pCEP4-ZfL1-12B-Chr19-*mneoI* (pMK045), with the antisense *mneoI* moved upstream of the ZfL1-12B 3'UTR (constructed using PCR and cloning).	Figure 6.3
pCEP4-ZfL1-12B-Cns-*mneoI*-3'UTR (pMK057)	pCEP4-ZfL1-12B-Cns-*mneoI* (pMK046), with the antisense *mneoI* moved upstream of the ZfL1-12B 3'UTR (constructed using PCR and cloning).	Figure 6.3
pTOL2-ZfL1-7B-Cns-2A-mCherry-EGFP-ZfL1-7B-3'UTR (pMK058)	pTOL2-ZfL1-7B-Cns-EGFP-ZfL1-7B-3'UTR (pMK029), with the mCherry coding sequence fused in-frame to the N-terminus of the ZfL1-7B-ORF2p by a self-processing 2A sequence (Kim et al., 2011) (constructed using PCR and cloning).	Figure 6.6
pTOL2-ZfL1-10B-Cns-2A-mCherry-EGFP-ZfL1-10B-3'UTR (pMK059)	pTOL2-ZfL1-10B-Cns-EGFP-ZfL1-10B-3'UTR (pMK033), with the mCherry coding sequence fused in-frame to the N-terminus of the ZfL1-7B-ORF2p by a self-processing 2A sequence (Kim et al., 2011) (constructed using PCR and cloning).	Figure 6.6

Continues in the next page.

Identifier	Description	Figure
pTOL2-ZfL1-12B-Cns-2A-mCherry-EGFP-ZfL1-12B-3'UTR (pMK060)	pTOL2-ZfL1-12B-Cns-EGFP-ZfL1-12B-3'UTR (pMK035), with the mCherry coding sequence fused in-frame to the N-terminus of the ZfL1-7B-ORF2p by a self-processing 2A sequence (Kim et al., 2011) (constructed using PCR and cloning).	Figure 6.6
pTOL2-ZfL2-2-2A-mCherry-EGFP-SL (pMK061)	pTOL2-Zfl2-2-EGFP-SL (***Table 2.3***), with the mCherry coding sequence fused in-frame to the N-terminus of the ZfL2-2-ORFp by a self-processing 2A sequence (Kim et al., 2011) (constructed using PCR and cloning).	Figure 6.8
pTOL2-Zfl2-2-RT^{-}-2A-mCherry-EGFP-SL (pMK062)	pTOL2-Zfl2-2-RT^{-}-EGFP-SL (pMK013), with the mCherry coding sequence fused in-frame to the N-terminus of the ZfL2-2-ORFp by a self-processing 2A sequence (Kim et al., 2011) (constructed using PCR and cloning).	Figure 6.8
pTOL2-ZfL2-2-EN^{-}-2A-mCherry-EGFP-SL (pMK063)	pTOL2-ZfL2-2-2A-mCherry-EGFP-SL (pMK062) containing a missense mutation in the EN domain of the *zfL2-2-orf* (E72A), abolishing retrotransposition competency (Honda et al., 2007) (constructed using PCR and cloning).	Figure 6.8
pTOL2-ZfL2-2-TBR2m-2A-mCherry-EGFP-SL (pMK064)	pTOL2-ZfL2-2-2A-mCherry-EGFP-SL (pMK062) containing a missense mutation in 2 conserved Tail Binding Regions (TBRs) of the *zfL2-2-orf* (W325A/R334A), abolishing its interaction with the stem-loop (Hayashi et al., 2014) (constructed using PCR and cloning).	Figure 6.8
pTOL2-ZfL2-2-2A-mCherry-EGFP-NSL (pMK065)	pTOL2-Zfl2-2-EGFP-NSL (***Table 2.3***), with the mCherry coding sequence fused in-frame to the N-terminus of the ZfL2-2-ORFp by a self-processing 2A sequence (Kim et al., 2011) (constructed using PCR and cloning).	N/A
pTOL2-ZfL2-2-2A-mCherry-STOP-EGFP-SL (pMK066)	pTOL2-ZfL2-2-2A-mCherry-EGFP-SL (pMK062) containing a stop codon 3' of the 2A sequence, abolishing ORFp synthesis (constructed using PCR and cloning).	Figure 6.8

[1]Chromosome annotation for the ZfL1 are according to the assembly 10/danRer10.

Table 2.4. Vectors that were generated as part of this thesis.

2.2. General methods

2.2.1. Manipulation of nucleic acids

2.2.1.1. Vector DNA preparation from *E. coli*

Small-scale vector DNA was purified from bacteria cultures using the QIAprep Spin Miniprep Kit (Qiagen, Cat. No. 27104) following manufacturer's protocol. DNA was extracted/purified from 5 mL of stationary phase *E. coli* culture and eluted in 50 µL of the provided elution buffer (10mM Tris-Cl, pH 8.5).

Larger scale vector DNA was purified from bacteria cultures using the ZymoPURE II Vector Midiprep Kit (Zymo Research, Cat. No. D4200). DNA was extracted/purified from 50-100 mL of stationary phase *E. coli* following manufacturer's protocol and eluted in 200 µL of the provided elution buffer (10 mM Tris-HCl, pH 8.5, 0.1 mM EDTA).

2.2.1.2. Polymerase Chain Reaction (PCR) methods

2.2.1.2.1. Primer design

Oligonucleotide primers for DNA amplification were designed using the Primer3 primer design program (http://primer3.sourceforge.net/). Genotyping primers were designed with melting temperature (T_m) of ~55°C (52-58°C), while primers for amplifying DNA fragments for cloning were designed with T_m of ~60°C (58-62°C). Primer pairs' T_m did not differ >4°C. The absence/interference of off-target amplification of each primer pair to be used on genomic DNA was checked using the UCSC In-Silico PCR tool (https://genome.ucsc.edu/index.html).

2.2.1.2.2. PCR amplification

Specific regions of DNA were amplified using PCR. The designed oligonucleotide primers were annealed to denatured template DNA and extended by a thermostable DNA polymerase. Genotyping and colony-PCR were performed using the DreamTaq Green PCR Master Mix (Thermo Scientific, Cat. No. K1081). A typical genotyping PCR reaction contained: 1-100 ng template DNA (vector or genomic DNA), 1× DreamTaq Green PCR Master Mix, 0.1 µM forward primer, 0.1 µM reverse primer, and molecular grade H_2O up to a total reaction volume of 10 µL. In colony PCRs, the template DNA was replaced by a single bacterial colony picked with a sterile pipette tip. The DreamTaq Green PCR amplification was performed on a DNA Engine Tetrad 2 thermal cycler (Bio-Rad) using the following program:

Step	Temperature (°C)	Time (min:sec)	Repeats
Initial denaturation	95	1:00	1
Denaturation	95	0:30	
Annealing	50-55[1]	0:30	30
Extension	72	1:00	
Final extension	72	5:00	1

[1]Annealing temperature is based on the melting temperature of the primer set used in each PCR (T_m –5°C).

Long PCR amplicons (>5 Kb in length) were obtained from genomic DNA by using the Roche Expand™ Long Range dNTPack (Sigma-Aldrich, Cat. No. 4829042001). A typical long amplicon PCR reaction contained: 50-100 ng genomic DNA, 1× Expand Long Range Buffer with 12.5 mM $MgCl_2$, 500 µM PCR Nucleotide Mix (dATP, dCTP, dGTP, dTTP at 10 mM each), 0.3 µM forward primer, 0.3 µM reverse primer, 3% DMSO, 3.5 U Expand Long Range Enzyme mix (5 U/µL) and molecular grade H_2O up to a total reaction volume of 50 µL. The Expand™ Long Range dNTPack PCR amplification was performed on a DNA Engine Tetrad 2 thermal cycler (Bio-Rad) using the following program:

Step	Temperature (°C)	Time (min:sec)	Repeats
Initial denaturation	92	2:00	1
Denaturation	92	0:10	
Annealing	58-62[1]	0:15	10
Extension	68	0:60/Kb[2]	
Denaturation	92	0:10	
Annealing	58-62[1]	0:15	25
Extension	68	0:60/kb + 0:20/cycle[2]	
Final extension	72	5:00	1

[1]Annealing temperature is based on the melting temperature of the primer set used in each PCR.
[2]Extension time depends on amplicon length.

When a high level of polymerase accuracy was required and the input DNA was a homogeneous template (i.e. vector DNA), Q5® High-Fidelity DNA Polymerase (New England Biolabs, Cat. No. M0491) was used. A typical High-Fidelity PCR reaction contained: 1-100 ng homogeneous DNA template, 1× Q5 reaction buffer, 200 µM dNTPs (Invitrogen, Cat. No. 10297117), 0.5 U Q5 High-Fidelity DNA Polymerase, 0.5 µM forward primer, 0.5 µM reverse primer, and molecular grade H_2O up to a total reaction volume of 25 µL. The Q5® High-Fidelity amplification was performed on a DNA Engine Tetrad 2 thermal cycler (Bio-Rad) with the following program:

Step	Temperature (°C)	Time (min:sec)	Repeats
Initial denaturation	98	0:30	1
Denaturation	98	0:10	
Annealing	58-63[1]	0:20	30
Extension	72	0:20/kb[2]	
Final extension	72	2:00	1

[1]Annealing temperature is based on the melting temperature of the primer set used in each PCR and was calculated using the NEB T_m calculator (http://tmcalculator.neb.com/#!/main).
[2]Extension time depends on amplicon length.

2.2.1.3. Digestion/restriction of nucleic acids

Purified vector DNA or PCR amplicons were digested with the appropriate restriction endonuclease in the buffer supplied by the manufacturer (New England Biolabs). A typical digestion reaction contained: 1× reaction buffer, 1-3 µg of DNA, 20 U of the appropriate restriction endonuclease(s) and molecular grade H_2O up to a total reaction volume of 20 µL. Reactions were incubated at the manufacturer recommended temperature (typically 37°C) for 1-3 h. For double digestions, the optimal buffer condition for both enzymes was selected using the manufacturer's guidelines. Digested vectors, from which the backbone was to be used for subsequent cloning steps, were treated with Antarctic Phosphatase (New England Biolabs, Cat. No. M0289) to prevent religation. A typical dephosphatase reaction contained: 20 µL digestion reaction, 1× Antarctic Phosphatase Reaction Buffer, 5 U Antarctic Phosphatase (5,000 U/mL) and molecular grade H_2O up to a total reaction volume of 30 µL. If the digestion-produced DNA fragments were to be used for subsequent cloning steps, they were resolved on agarose gel by electrophoresis, excised (using a sterile scalpel) and purified (2.2.1.5; 2.2.1.6).

2.2.1.4. Spectrophotometric quantification of nucleic acids

The concentration of nucleic acids was determined using a NanoDrop 1000 UV-Vis Spectrophotometer (Thermo Fisher Scientific) to measure the optical density at 260 nm of 2 µL of each sample. The purity of the nucleic acid sample was determined by measuring its absorbance at 230 nm, 260 nm and 280 nm and using the 260/280 ratio to assess protein contamination, and the 230/260 ratio to assess carbohydrate and lipid contamination. A sample free of protein, carbohydrate and lipid contamination should have a 260/280 ratio of 1.8-2.2 and a 230/260 ratio of ≥1.7 respectively.

2.2.1.5. Agarose gel electrophoresis

Nucleic acid samples were resolved by electrophoresis on agarose gels ranging from 0.8% to 2% agarose in TBE (w/v). Gels were prepared by dissolving a Hi-Pure Low EEO agarose (Biogene, Cat. No. 300-300) in 0.5× TBE buffer (***Table 2.1***), boiling the mixture in a microwave oven, cooling it for a few minutes, and adding 0.01% of SYBR™ Safe DNA Gel Stain (Invitrogen, Cat. No. S33102) following manufacturer's protocol. Nucleic acid samples were mixed with 6× Purple Gel Loading Dye (New England Biolabs, Cat. No. B7024S. Final concentration 1×), loaded

onto the gel, and resolved by size using 5 volts/cm voltage. For reference, a 1kb Plus DNA Ladder (Invitrogen, Cat. No. 10787018) containing DNA fragments of known sizes was included (typically 500ng lane). Nucleic acids were visualised using a UV transilluminator (BioDoc-It System, UVP), a Blue light LED transilluminator (Syngene) or a FujiFilm FLA-5100 Fluorescent Image Analyser (Raytek Scientific).

2.2.1.6. DNA purification

2.2.1.6.1. Gel extraction

DNA fragments produced by PCR (2.2.1.2) or digestion reactions (2.2.1.3) were resolved by agarose gel electrophoresis (2.2.1.5) and the DNA fragments of interest were excised from the gel using a sterile scalpel. The excised gel piece was initially frozen at –80°C and, once thawed, smashed using a pestle. Purification was done by phenol-chloroform extraction. The sample was thoroughly mixed with Phenol Solution (10 mM Tris-HCl, pH 8.0, 1 mM EDTA equilibrated) (Sigma-Aldrich, Cat. No. P4557) in a ratio 1:1 (w/v), followed by a 1 min incubation at –80°C. Phase separation was performed through centrifugation (10 min, 17,000 ×*g*) and the aqueous phase was transferred to a fresh tube. Chloroform/Isoamyl alcohol 24:1 (ACROS Organics, Cat. No. 10103971) was added to the collected aqueous phase in a ratio of 1:1 (v/v) and thoroughly mixed. Phase separation through centrifugation (10 min, 17,000 ×*g*) was repeated and the aqueous phase transferred to a fresh tube. Precipitation was done by adding absolute ethanol (Fisher Bioreagents, Cat. No. 10644795) in ratio 3:1 (v/v), 3M NaAc pH 5.2 (Millipore, Cat. No. 567422) in a ratio 1:10 (v/v) and 1 µL Glycogen 5 mg/mL (Invitrogen, Cat. No. AM9510), followed by a 30 min incubation at –80°C. The DNA was pelleted by centrifugation (30 min, 17,000 ×*g*, 4°C) and the supernatant was discarded by inversion. The DNA pellet was washed using 1mL of 70% absolute ethanol (Fisher Bioreagents) in molecular grade H_2O, followed by centrifugation (10 min, 17,000 ×*g*, 4°C) and the supernatant was discarded by inversion. The DNA pellet was air-dried and resuspended in an appropriate volume (typically 10 µL) of molecular grade H_2O.

2.2.1.6.2. Amplicon/Digested DNA purification

PCR (2.2.1.2) or digestion reactions (2.2.1.3) that were not resolved by agarose gel electrophoresis were also purified by phenol-chloroform extraction. The volume of the PCR/digestion reaction was brought up to 150 µL by adding molecular grade H_2O. Purification was done by first thoroughly mixing the sample with Phenol Solution (10 mM Tris-HCl, pH 8.0, 1 mM EDTA equilibrated) (Sigma-Aldrich) in a ratio 1:1 (v/v), followed by phase separation through centrifugation (10 min, 17,000 ×*g*) and transfer of the aqueous phase to a fresh tube. This process was then repeated using Chloroform/Isoamyl alcohol 24:1 (ACROS Organics). Precipitation was done by adding absolute ethanol (Fisher Bioreagents) in ratio 3:1 (v/v), 3M NaAc pH 5.2 (Millipore) in ratio 1:10 (v/v) and 1 µL Glycogen 5 mg/mL (Invitrogen), followed by a 30 min incubation at –80°C. The DNA was pelleted by centrifugation (30 min, 17,000 ×*g*, 4°C) and the supernatant was discarded by inversion. The DNA pellet was washed using 1mL of 70% absolute ethanol (Fisher Bioreagents) in molecular grade H_2O, followed by centrifugation (10 min, 17,000 ×*g*, 4°C) and the supernatant was discarded by inversion. Finally, the DNA pellet was air-dried and resuspended in 10 µL molecular grade H_2O.

2.2.1.6.3. Genomic DNA extraction

Tissues/cells were collected and lysed using lysis buffer (***Table 2.1***). 100 µL 1× lysis buffer and 0.8 U of Molecular Biology Grade Proteinase K (New England Biolabs, Cat. No. P8107S) was added to each sample and incubated at 60°C for 16 h, followed by inactivation of Proteinase K by incubation at 95°C for 10 min.

2.2.1.7. DNA ligation into vectors

Purified PCR products (2.2.1.6) were cloned into one of three commercially available vectors, depending on size and nature of their ends. Amplicons <3 Kb in length and with A-overhang ends, generated by the DreamTaq Green PCR Master Mix, were ligated into pGEM-T Easy Vector using the pGEM®-T Easy Vector Systems (Promega, Cat. No. A1360). Roche Expand™ Long Range dNTPack amplicons >3 Kb in length were ligated into TOPO-TA Cloning Vector using the TOPO™ TA Cloning™ Kit for Subcloning (Invitrogen, Cat. No. 450641), and those generated by Q5® High-Fidelity DNA Polymerase were ligated into TOPO-Blunt DNA Cloning Vector using the Zero Blunt™ TOPO™ PCR Cloning Kit (Invitrogen, Cat. No. 450245).

The ligation reactions were performed following manufacturer's instructions, using the maximum volume of PCR product.

Restriction digested (2.2.1.3) and purified (2.2.1.6) DNA fragments were cloned in specific custom vectors, digested with compatible restriction endonucleases and dephosphorylated, using T4 DNA ligase (New England Biolabs, Cat. No. M0202). A typical ligation reaction contained: 1 µL vector, 5 µL insert(s), 1× T4 DNA Ligase Buffer, 400 U T4 DNA Ligase (400,000 U/mL) and molecular grade H_2O up to a total reaction volume of 10 µL. The ligation reaction was incubated overnight at 16°C. The next day the appropriate amount of ligation reaction was transformed into ultracompetent *E. coli* bacteria (2.2.2.3).

2.2.1.8. DNA sequencing

Capillary DNA sequencing was performed at the Institute of Genetics and Cancer (IGC) sequencing service on an AB3130/3730 genetic analyser (Applied Biosystems). For vector sequencing, 1µL of 100-150 ng/µL concentrated vector was used. For PCR sequencing, 1µL of the PCR reaction volume was run on an agarose gel to confirm the presence of the appropriate product. The rest of the PCR reaction volume was provided to the sequencing service for clean-up and sequencing. For all sequencing reactions, 1 µL of a 5 µM solution of the appropriate primer was used. DNA sequencing data was analysed using Sequencher 5.4.6 DNA sequence analysis software (Gene Codes Corporation).

2.2.2. Microbial methods

2.2.2.1. Growth of bacteria

E. coli DH5α and *E. coli* Stbl4 cells were both grown in/on Luria-Bertani (LB) medium (10 g/L tryptone, 5 g/L yeast extract, 10 g/L NaCl, 1 g/L glucose) at 37°C and 30°C respectively. The relevant antibiotic(s) were added to the LB media at the required concentration to maintain selection for vector DNA in transformed bacteria as specified in (***Table 2.5***).

Antibiotic	Stock conc.	Working conc.	Solvent
Ampicillin	50 mg/mL	100 µg/mL	Molecular grade H_2O
Kanamycin	10 mg/mL	50 µg/mL	Molecular grade H_2O

Table 2.5. Antibiotics used for bacterial selection.

2.2.2.2. Preparation of chemically-competent bacterial cells

E. coli DH5α cells were grown overnight on LB agar at 37°C. A single colony was picked the next day, inoculated into 5 mL of LB medium with 20 mM $MgSO_4$, and grown overnight to stationary phase. Following this, 250 mL of LB with 20 mM $MgSO_4$ were inoculated with 2 mL of the stationary phase culture and incubated at 23°C in a shaking incubator at 200 rpm until reaching an OD600 of 0.4-0.6 (usually 8-10 h). The culture was then cooled on ice for ~15 min and cells were kept on ice for all subsequent steps. Cells were pelleted (10 min, 4,000 ×*g*, 4°C) and gently resuspended in 80 mL of ice-cold sterile TB buffer (***Table 2.1***). Cells were incubated on ice for 10 min, centrifuged (10 min, 4,000 ×*g*, 4°C), and gently resuspended in 20 mL of ice-cold TB buffer. After adding 1.5 mL of DMSO followed by a final 10 min incubation on ice, the bacterial cell suspension was split into 200 µL aliquots in cold, sterile tubes and snap-frozen on dry ice. Aliquots were stored at –80˚C. The competence of prepared *E. coli* cells was analysed using transformation of serially diluted pUC19 (or pBSKS).

2.2.2.3. Transformation of E. coli

2.2.2.3.1. Transformation of chemically competent bacterial cells

For transformation of chemically-competent *E. coli* DH5α, approximately 1 ng of vector DNA or 1 µL of a ligation reaction was added to 50 µL of competent cells thawed in ice. Bacterial cells and DNAs were incubated on ice for 30 min in pre-chilled tubes, followed by a 45 sec heat-shock at 42°C, and a 2 min incubation on ice. The bacterial cells were resuspended in 1 mL of pre-heated LB medium and incubated at 37°C for 30 min with shaking at 225 rpm. 50 µL of cells were plated onto half of an LB-agar plate containing the appropriate antibiotic(s). The rest of the cells were pelleted using a tabletop centrifuge (10 sec, 17,000 ×*g*). Part of the supernatant was removed until leaving ~50 µL, which was used to resuspend the pellet and plate it on the other half of the LB-agar plate. These two plating densities increased the chances of obtaining colonies for suboptimal ligation/transformation. The plates were incubated overnight at 37°C to obtain colonies.

2.2.2.3.2. Transformation of electro-competent bacterial cells

For transformation of electro-competent *E. coli* Stbl4, ElectroMAX™ Stbl4™ Competent Cells (Invitrogen, Cat. No. 11635018), approximately 1ng of vector DNA or 1 µL of a ligation reaction was added to 20 µL of competent cells thawed in ice. Cells and DNA were gently mixed and added to a pre-chilled MicroPulser Electroporation Cuvette, 0.1 cm gap (Bio-Rad, Cat. No. 1652089), and kept on ice until electroporation took place using a GenePulser (1.2 kV, 25 µF and 200 Ω). The cells were resuspended in 1 mL of SOC medium (Invitrogen, Cat. No. 15544034) and incubated for 1.5 h at 225 rpm, 30°C. 50 µL of cells were plated onto half of an LB-agar dish containing the appropriate antibiotic. The rest of the cells were pelleted using a tabletop centrifuge (10 sec, 17,000 ×*g*). Part of the supernatant was removed until leaving ~50 µL, which was used to resuspended the pellet and plate it on the other half of the LB-agar plate. These two plating densities increase the chances for obtaining colonies for suboptimal ligation/transformation. The plates were incubated overnight at 30°C to obtain colonies.

2.2.3. Cell culture methods

2.2.3.1. Maintenance of cell lines

Cell culture conditions used for the growth and maintenance of human cell lines are summarised in ***Table 2.6***. Cells were routinely subjected to mycoplasma testing. For cell passage, cells were washed with 1× PBS (***Table 2.1***), and lifted using 37°C pre-warmed trypsin:versene (1:1, v/v) or 1× TrypLE™ Express Enzyme (Gibco, Cat. No. 12605010) for 5 min at 37°C. The dissociation reaction was then neutralized with at least 2 volumes of normal growth medium, and cells were passaged 1:8-1:10, as required. The appropriate antibiotic(s) (***Table 2.2***) was/were added to the media at the required concentration to maintain selection of vector DNA in transfected and/or transduced cells.

Cell line	Cell culture medium	Supplements of cell culture medium	Cell culture conditions	Source
HeLa	Dulbecco's Modified Eagle Medium (DMEM), high glucose (Gibco, Cat. No. 41965039)	10% Foetal Calf Serum (FCS), 100 U/mL penicillin, 100 µg/mL streptomycin, 1% L-Glutamine (30.0 g/L)	37°C, 5% CO_2, normoxia	G. Stewart (University of Birmingham); originally from ATCC
HEK293ET				MAM. Reijns; originally from ATCC

Table 2.6. Cell culture conditions used for maintaining human cell lines.

2.2.3.2. Cell counting

When cells had to be seeded at a set density, cell concentration was determined by manually or automated cell counting. Manual counting was done using the FastRead 102, Disposable Cell Counting Slides (Biomedical, Cat. No. 3HBVS100) in combination with an Olympus CKX53 inverted microscope (Olympus). 10 µL of single cell suspension was loaded into the counting chamber, and 5 large squares were counted. The concentration was then calculated using the following formula: Cells/ml = (total cell counted in 5 squares / 5) × dilution factor × 10^4. Automated counting was done using the Countess II FL automated cell counter (Invitrogen, Life technologies) in combination with Countess™ Cell Counting Chamber Slides (Invitrogen, Cat. No. C10228) and Trypan Blue Stain (0.4%) (Invitrogen, Cat. No. T10282). 10 µL of 1:1 mix of single cell suspension and 0.4% Trypan blue stain was loaded into a counting chamber. The mixture was allowed to settle for 30 seconds before the slide was analysed using the Countess II FL. The concentration of life cells was taken as the cell density.

2.2.3.3. Cryopreservation of cell lines

For storage of cultured cells, 2-7 ×10^6 adherent cells were harvested by trypsinisation (2.2.3.2) and resuspended in 1 mL of freezing media (FCS containing 10% DMSO). Cells were stored in 2 mL cryostat tubes, frozen at –80˚C in Styrofoam containers overnight, and transferred for long-term storage in liquid nitrogen.

2.2.3.4. Flow cytometry

2.2.3.4.1. Sample preparation

In preparation for analysis by flow cytometry, cells grown in a 6-well tissue culture plate were trypsinised using 500 µL trypsin:versene per well, followed by inactivation using 1 mL of 10% FCS in 1× PBS (2.2.3.1). Cells were completely detached from the well by pipetting up and down after which the cell suspension was transferred to a fresh tube. Cells were pelleted by centrifugation (3 min, 200 ×*g*) and the supernatant was discarded by aspiration. Cells pellets were resuspended in 120 µL ice-cold 10% FCS in 1× PBS, and transferred to a 96-well round (U) bottom plate (Thermo Scientific, Cat. No. 10344311).

2.2.3.4.2. Flow cytometry analysis

The BD LSR Fortessa Cell Analyzer (BD Biosciences) was used to analyse cell fluorescence. The FlowJo v10.8.1 software (BD Bioscience) was used to analyse the data files. A first gate was set around the events representing singlets, using the plot Forward Scatter (FSC) Area (A) against FSC Height (H). Within this population, a second gate was made to exclude dead cells and debris, using the plot FSC-A against Side Scatter (SSC)-A. Finally, gates including the specific population(s) of interest were set based on a negative control (untransfected cells). To quantify EGFP expressing cells, 488-525/50-A (filter used to detect EGFP signal) was plotted against 561-586/15-A. To quantify EGFP positive and mCherry expressing cells, 488-525/50-A (filter used to detect EGFP signal) was plotted against 561-610/20-A. Percentage fluorescent cells, as well as their intensity, were calculated.

2.3. RNase H involvement in LINE-1 retrotransposition

2.3.1. RNASEH2B knock-out and complementation

2.3.1.1. RNASEH2B knock-out

To abrogate expression of RNASEH2B in a HeLa cell line, *RNASEH2B* gene KO was performed using the Synthego *Streptococcus pyogenes* Cas9 with 2 Nuclear Localization Signals (Cas9 2NLS Nuclease) in combination with 3 synthetic guide RNAs (sgRNA). The 3 sgRNAs were designed using the online Synthego CRISPR Design Tool (https://design.synthego.com/) and ordered from Synthego (**Table 2.7**). A Cas9:sgRNA RNP complex was formed by combining 20 pmol Synthego Cas9-2NLS, 180 pmol diluted sgRNA (30 µM) and Neon Resuspension R Buffer (Synthego) up to a total volume of 30 µL, and incubating the mix for 10 min at room temperature. The formed Cas9:sgRNA RNP complex was kept on ice until electroporation took place.

Genome	Target gene	Target exon	Sequence (5'-3')	Nuclease	Fig.
Homo sapiens Gencode Release 21 (GRCh38)	RNASEH2B:ENST00000336617 (primary)	3	AGAUUUGUCUUAACAGGAGA	SpCas9	3.2
			ACACAUAUUGAACAAGUAAA		
			AUUGAUUUAUAAACCAAGAA		

Table 2.7. List of sgRNAs used in this thesis.

70 - 80% confluent, cultured HeLa cells were collected from a T75 tissue culture flask using 3 mL TrypLE Express and neutralised in 6 mL culture media (*Table 2.6*; 2.2.3.1). The concentration of the HeLa single cell suspension was determined using the countess method (2.2.3.2). 5 $\times 10^5$ cells were transferred to a sterile 1.5 mL tube, pelleted by centrifugation (5 min, 500 $\times g$) and supernatant removed. Cells were washed with 1× PBS, pelleted by centrifugation (5 min, 500 $\times g$), supernatant removed and resuspended in 100 µL Resuspension Buffer R. Within 15 min, 70 µL of this cell suspension was added to the formed Cas9:sgRNA RNP complex, and the RNP/cell mixture was aspirated into a 100µL Viability Tip. According to manufacturer's instructions the following electroporation conditions were used: 1005 V pulse voltage, 35 ms pulse width, 2 pulses. Cells were single cell sorted 24 hours after electroporation and expanded (*Table 2.6*; 2.3.1.3).

2.3.1.2. RNASEH2B complementation (retroviral transduction)

RNASEH2B complementation was achieved using the retroviral Murine Stem Cell Virus (MSCV) system. For retroviral production, HEK293ET cells were seeded at a density of 1.0 $\times 10^6$ cells/well of a 6-well tissue culture plate. After 24 hours, the media was changed and cells were transfected with vectors for viral particle production using Polyethylenimine (PEI) and Opti-MEM reduced serum medium (Gibco, Cat. No. 31985070). The transfection mix contained: 8 µg PEI (1 µg/µL), 90 µL Opti-MEM reduced serum medium, 2 µg of the pMD2.G vector, 2 µg of the GAG-Pol vector and 2 µg of either i) pMSCV-Zeo (Empty Vector (EV)), ii) pMSCV-RNH2B-Zeo (*RNASEH2B* wildtype) or ii) pMSCV-RNH2B-PIP-Zeo (*RNASEH2B* PIP mutant (F300A/F301A)) vectors (Table 2.3). The transfection mixes were incubated for 15 min at room temperature before adding them to the plated HEK293ET cells. Cells were incubated with the transfection mixture overnight, after which the culture media was replaced with fresh media. Retrovirus-containing supernatants were collected 48 h post-transfection and filtered through a 33 mm Ezee™ Syringe Filters, 0.45 µm, PVDF (Elkay Laboratory Products, Cat. No. E25-PV45-50S). The filtrated supernatants containing retrovirus were immediately used for infection or aliquoted into 1.5 mL cryovials and stored at –80°C.

CRISPR RNASEH2B KO cells were complemented with RNASEH2B WT, RNASEH2B PIP mutant (F300A/F301A), or EV by retroviral transduction. HeLa CRISPR RNASEH2B KO cells were seeded at a density of 1 $\times 10^5$ cells/well of a 6-well tissue culture plate and grown for 24 h.

Then, the cells were transduced using Polybrene (Pb) and DMEM (***Table 2.6***), by replacing their media with a mix containing 2 µg Pb (5 mg/mL), 1800 µL DMEM (***Table 2.6***) and 200 µL of the viral supernatant with either *RNASEH2B* WT, *RNASEH2B* PIP mutant (F300A/F301A), or EV particles. After 48 h, the cells were split into 2 wells of a 6-well tissue culture plate with media containing the selective antibiotic Zeocin (***Table 2.2***). Transduced cells were selected for 14 days, changing the Zeocin containing media every other day. Untransduced cells were used as a negative control, confirming that the selection process was completed when all untransduced cells had died. After selection, the transduced cells were expanded.

2.3.1.3. Single cell sorting

Cells were single cell sorted into 96-well plates containing regular growth media (***Table 2.6***), by Fluorescence-Activated Cell Sorting (FACS) using the BD FACSJazz (BD Biosciences), and expanded. Once confluency was reached, clones were split and re-plated into 24-well plates, as well as collected for rapid DNA and/or whole cell extract preparation.

2.3.1.4. RNASEH2B model validation

2.3.1.4.1. CRISPR KO validation by PCR

Cells collected for rapid DNA preparation were lysed using DirectPCR Lysis Reagent (Viagen, Cat. No. 301-C). Briefly, 75 µL DirectPCR Lysis Reagent and 30 µg Proteinase K Recombinant PCR Grade Lyophilizate (Roche, Cat. No. 3115879001) was added to each sample and incubated at 55°C for 16 h, followed by inactivation of Proteinase K by incubation at 85°C for 45 min. 1 µL of the lysate was used for PCR amplification of the genomic region around the expected nucleotide change using DreamTaq Green PCR Master Mix (2.2.1.2; ***Table 6.1***). 1 µL of PCR product was resolved on a 1.5% Agarose gel to confirm deletion had taken place in within the amplified region (2.2.1.5). The rest of the PCR product was sent for capillary sequencing (2.2.1.8). Sequencing traces were analysed using Sequencher 4.8 DNA Sequence Analysis Software (Gene Codes Corporation).

2.3.1.4.2. Whole-cell extract preparation and protein quantification

For Whole-Cell Extract (WCE) preparation, cells were harvested using trypsine, washed with 1× PBS and resuspended in an appropriate volume (50-80 µL per 10^6 cells) of WCE buffer

(*Table 2.1*) containing 1 mM DTT and 1mM PMSF, followed by an incubation of 10 min on ice. Then, an equal volume of Cytoplasmic buffer with 1 mM DTT and 1 mM PMSF was added and incubated 10 min on ice. WCEs were cleared by centrifugation (17,000 ×*g*, 15 min, 4°C) and supernatants were transferred to a fresh tube and stored at –80°C or immediately used for protein quantification.

Total protein concentration in WCEs was determined using the Quick Start Bradford Protein Assay (Bio-Rad, Cat. No. 5000202). The quick Start Bovine Serum Albumin Standard Set (Bio-Rad, Cat. No. 5000206) (concentrations 0.125, 0.25, 0.5, 0.75, 1.0, 1.5 and 2.0 mg/mL) was used to generate a standard curve for reference. The protein concentration of a 1:10 dilution of WCE in molecular grade H_2O was quantified as follows: 10 µL of each solution was mixed with 190 µL of 1× Bradford dye reagent and the absorbance at 595 nm (A595) was measured after 5 min using a Multiskan Spectrum (Thermo Scientific) with the SkanIt Software (Thermo Scientific). Absorbance readings of the standard curve were plotted against protein concentration, and a line of best fit was used. Absorbance readings of the WCE samples were compared to the BSA standard curve to calculate protein concentration. Quantified samples were aliquoted and frozen at –80°C or immediately used for further analyses.

2.3.1.4.3. RNase H2 activity assay

RNase H2 activity was assessed using a FRET-based fluorescent substrate release assay, performed as previously described (Crow et al, 2006 and Reijns et al, 2011). Briefly, cellular RNase H2-specific activity was determined by measuring cleavage of a single-embedded ribonucleotide-containing short double-stranded DNA substrate (DRD:DNA) in WCEs. To correct for non-RNase H2 background nuclease activity against the DRD:DNA substrate present in WCEs, a DNA:DNA substrate of the same sequence was tested in parallel as a control. DRD:DNA and DNA:DNA substrates were formed by annealing a 3′-fluorescein-labelled oligonucleotide (5′GATCTGAGCCTGGGaGCT or 5′GATCTGAGCCTGGGAGCT; uppercase DNA, lowercase RNA) to a complementary 5′ DABCYL-labelled DNA oligonucleotide (Eurogentec). Reactions were performed in 100 µL final volume containing 60 mM KCl, 50 mM Tris-HCl pH 8.0, 10 mM $MgCl_2$, 0.01% BSA, 0.01% Triton X-100, 250 nM substrate and 100 ng/µL final protein concentration using previously prepared WCEs (2.3.1.3.1), in 96-well flat-bottomed plates at 24°C. A standard curve ranging from 0.78, 3.125, 12.5, 25 and 50 pmol cleaved

product was used as a reference. Fluorescence was read (100 ms) every 5 min for up to 90 min using a VICTOR2 1420 multilabel counter (Perkin Elmer), with a 480-nm excitation filter and a 535-nm emission filter.

For each time point, emission readings of the standard curve were plotted against the known quantity, and a line of best fit was used. Emission readings of the WCE samples were compared to the standard curve to calculate the amount of product formed in each well at each time point. The slope of the linear part of the line best fitting the amount of product formed over time was used to calculate RNase H2 activity. The activity was then normalised to that of the WT cell line (set to 100%).

2.3.1.4.4. RNase H2 protein levels

2.3.1.4.4.1. Sodium Dodecyl Sulphate-Polyacrylamide Gel Electrophoresis (SDS-PAGE)

Proteins in WCEs were separated according to their molecular weight by SDS-PAGE using either 4-12% Bis-Tris NuPAGE® Mini protein gel (Invitrogen, Cat. No. NP0321) in combination with the XCell SureLock Mini-Cell Electrophoresis System (Thermo Fisher), or 10% home-made Tris-glycine gels in combination with the Mini-PROTEAN® Tetra Handcast system (Bio-Rad). The Tris-glycine gels were cast using the Mini-PROTEAN® 0.75 mm Spacer plates (Bio-Rad). The resolving gel contained: 375 mM Tris (pH 8.8), 0.1% (w/v) SDS, 0.1% (w/v) ammonium persulphate, 0.1% TEMED (N'N'N'N'-tetramethylethylenediamine) (Thermo Scientific, Cat. No. 17919) and 10% acrylamide/bis-acrylamide (Sigma-Aldrich, Cat. No. A3574). The stacking gel contained: 125 mM Tris (pH 6.8), 0.1% SDS, 0.1% ammonium persulphate, 5% acrylamide/bisacrylamide (Sigma-Aldrich) and 0.1% TEMED (Thermo Scientific).

35-50 µg of protein sample in 1× SDS protein sample loading buffer (*Table 2.1*) was denatured by heating at 95°C for 5 min and loaded onto the appropriate gel pre-installed in corresponding system according to manufacturer's instructions. Protein samples were electrophoresed alongside the Precision Plus Protein™ Dual Xtra Prestained Protein Standard (Bio-Rad, Cat. No. 1610377). NuPAGE MOPS SDS running buffer (20×) (Invitrogen, Cat. No. NP000102) was used for Bis-Tris NuPAGE® gels, and 1× Tris-glycine running buffer (*Table 2.1*) for Tris-glycine gels. A constant voltage of 150-200 volts was applied until the desired separation was reached.

2.3.1.4.4.2. RNase H2 Immunoblotting

Proteins resolved by SDS-PAGE were transferred to Amersham Hybond P 0.2 PVDF membranes (Cytiva, Cat. No. 10600021) by wet transfer using a Mini-Trans-Blot Cell system (Bio-Rad). PVDF membranes (Cytiva) were first activated by immersion in 100% methanol after which the Mini Gel Holder Cassette (Bio-Rad) was loaded according to manufacturer's instructions and electrophoretic blotting was performed in 1× immunoblotting transfer buffer (*Table 2.1*) at 100 V for 1-1.5 h.

After electrophoretic transfer, the protein containing PVDF membranes (Cytiva) were blocked by 1 h incubation at room temperature in 5% milk solution (prepared using Marvel milk powder, Premier Foods) in TBS-T (*Table 2.1*), under constant agitation. The membranes were cut in two pieces between 37 and 50 kD, using the Precision Plus Protein Standard (Bio-Rad) as a guide, separating Tubulin 1 (50 kD) from the RNASEH2 subunits (35 - 18 kD). The Primary antibodies used were α-Tubulin (DM1A) Mouse mAb (Cell Signaling, Cat. No. 3873; used at 1:4,000) and sheep-α-RNASEH2 (Reijns et al., 2012; used at 1:750). Each part of the membrane was incubated overnight at 4°C with their corresponding antibody diluted in blocking solution under constant agitation. Next, the PDF membranes were washed 3 times for 15 min using 1× TBS-T (*Table 2.1*), after which each part was incubated for 1 h at room temperature with its corresponding Horseradish Peroxidase (HRP)-labelled secondary antibody (Goat anti-Mouse IgG H+L (Invitrogen, Cat. No. 62-6520; used at 1:10,000) and Rabbit anti-Sheep IgG H&L (Abcam, Cat. No. ad6747; used at 1:2,000) diluted in blocking solution) under constant agitation. Finally, the PVDF membranes were washed 3 times for 15 min using 1× TBS-T (*Table 2.1*).

To visualise immobilised HRP on the membranes, Amersham ECL Prime Western Blotting Reagent (Cytiva, Cat. No. GERPN2232) was used following manufacturer's instructions. For each 20 cm^2 PVDF membrane, 2 mL of ECL solution (1:1 mixture of solutions A and B) was added to the protein side of the PVDF membranes and incubated for 5 min at room temperature. Excess liquid was removed, and the membranes were placed between two acetate sheets and exposed to Kodak® BioMax® MS film (Sigma-Aldrich, Cat. No. Z363022), which was developed using a SRX-101A Developer (Konika). The chemiluminescence signal was also imaged using the ImageQuant LAS 4000 camera system (GE Healthcare).

Quantification of relative band intensities on not saturated ImageQuant LAS 4000 immunoblots images was performed using Fiji-ImageJ software (US National Institutes of Health) (Abràmoff et al., 2004; Schneider et al., 2012). An area encompassing a protein band was defined manually and kept constant for each measurement within the same immunoblot. The mean pixel intensity was calculated for each area. The signal intensity (I) of all the protein bands was first correct against the signal intensity of the background using the formula: $\Delta I = I_{protein\ band} - I_{background}$. Next, the correct signal intensity (ΔI) of the band of interest was normalised to a loading control on the same immunoblot using the formula: Normalised $I_{protein\ interest} = \Delta I_{protein\ interest} / \Delta I_{loading\ control}$.

2.3.2. Generation of LINE-1 ORF2p-RNase H chimeric constructs

To generate the LINE-1 ORF2p-RNase H fusion constructs, first the coding sequence of the *E. coli RNASEHI*, or the catalytic core of the human *RNASEH1* sequences were amplified from the vectors pMAR673 and pMAR676, as well as their catalytic dead alleles, from the vectors pMAR683 and pMAR684 respectively. PCRs were performed using the Q5® High-Fidelity DNA Polymerase (2.2.1.2.2) in combination with primers against the *E. coli RNASEHI* and human *RNASEH1* that incorporate a unique cloning site (NheI) and a flexible protein linker at the 5' end, as well as a unique cloning site at the 3' end (PacI) (2.2.1.2; ***Table 2.3***; ***Table 6.2***). The flexible linker was made up out of five tandem copies of a glycine (GGC)-serine (TCG) motif, allowing a theoretical free movement of the RNase H domain. The codons used to add the glycine/serine linker were chosen to have a higher GC content, for better polymerase processivity, and avoid creating unwanted splice sites. PCR reactions were resolved on 1.5% agarose gels by electrophoresis, and amplified bands were excised and gel-purified (2.2.1.3; 2.2.1.5; 2.2.1.6.1). Purified PCR fragments and the vector pGEM-T-ORF2C-OR were then subjected to NheI-PacI double digestion, removing the OR from the vector (2.2.1.7; ***Table 2.3***). PCR fragments were then independently ligated into the vector (2.2.1.7) and transformed into chemical-competent DH5α E. coli strain (2.2.2.3.1). Colonies were subjected to colony PCR (2.2.1.2), and miniprep DNA was capillary sequenced to confirm the identity of clones (2.2.1.8).

Confirmed clones and vectors pCEP4-L1.3-*mneoI*, pCEP4-L1.3-RT^{-}-*mneoI* and pCEP4-L1.3-PIPm-*mneoI* were then subjected to BstZ17I-PacI double digestion, releasing the 3' end of ORF2 sequence fused to the RNase H domains from the pGEM-T vectors, and removing the

3'end of the ORF2 sequence from pCEP4 vectors containing the different allelic variants of the L1.3 (2.2.1.7; ***Table 2.3***). Different combinations of PCR fragments with the different vectors were independently ligated (2.2.1.7) and transformed into chemical-competent DH5α *E. coli* strain (2.2.2.3.1). Colonies were subjected to colony PCR (2.2.1.2), and miniprep DNA was capillary sequenced to confirm the identity of clones (2.2.1.8), finally resulting in the vector series pMK001 to pMK0012 (***Table 2.4***).

2.3.3. Retrotransposition assays

2.3.3.1. DNA transfections

For retrotransposition assays using the different RNASEH2B-altered HeLa cell lines and control lines (2.3.1), cells were seeded in a well of a 6-well tissue culture plate at a predetermined cellular density (***Table 2.8***) (Day = 0). 16 - 18h after plating, cells were transfected using FuGENE® 6 Transfection Reagent (Promega, Cat. No. E2691) in combination with Opti-MEM™ Reduced Serum Medium (Gibco, Cat. No. 31985062) following the manufacturer's protocol, using 3 µL of FuGENE 6 (Promega), 97 µL of Opti-MEM (Gibco) and 1 µg of purified vector DNA per well of a 6-well tissue culture plate (Day = 1). All vectors used for transfection were prepared using the ZymoPURE II Vector Midiprep Kit (Zymo Research) (2.2.1.1). FuGENE 6 (Promega) was carefully added, avoiding the walls of the tubes, and gently mixed by pipetting up and down, immediately before adding 100 µL of the transfection mix to each well. Cells were incubated with the transfection mixture for about 24 h, after which the medium was replaced (Day = 2).

Assay details	L1-RNase H fusion	L1-WT	L1-PIP
Vectors used (Table 2.2 and 2.3)	L1.3-eRH; L1.3-eRHm; L1.3-hRH; L1.3-hRHm; L1.3-PIPm-eRH; L1.3-PIPm-eRHm; L1.3-PIPm-hRH; L1.3-PIPm-hRHm; L1.3-RT⁻-eRH; L1.3-RT⁻-eRHm; L1.3-RT⁻-hRH; L1.3-RT⁻-hRHm; pT2neo	L1.3; L1.3-RT⁻; pcDNA6.1	L1.3; L1.3-RT⁻; L1.3-PIPm; pcDNA6.1
Reporter	Neomycin^R	Blasticidin^R	Blasticidin^R
Seeding density (cells/well)	5×10^3	5×10^3	L1.3-RT⁻, L1.3-PIPm: 2×10^4 L1.3, pcDNA6.1: 5×10^3
Selection antibiotic (*Table 2.2*)	Geneticin (600 µg/mL)	Blasticidin (5 µg/mL)	Blasticidin (5 µg/mL)
Length (days)	17	15	15

Table 2.8. Conditions used for the retrotransposition assays.

2.3.3.2. Transfection efficiency

To determine the transfection efficiency of each plasmid within each cell line, cells were transfected with a vector mix containing 50% of an EGFP expression vector (pCEP4-EGFP) and 50% of the L1 vector (2.3.3.1), using 0.5 µg of each DNA vector. Two days after transfection, the cells were collected by trypsinisation to assess transfection efficiency of each vector (Day = 4) (2.2.3.4). The percentage of EGFP positive cells were assessed for each construct.

2.3.3.3. Colony formation assays

For colony formation assays, using *mneoI* or *mblastI* tagged LINE-1s, the density at which the cells were seeded was optimised to allow the formation of individual colonies for the vectors containing each specific L1 allele tested: 2.0 $\times 10^4$ cells/well of a 6-well tissue culture plate for the vectors containing the L1-PIPm, and 5.0 $\times 10^3$ cells/well for the vectors containing a WT L1 (***Table 2.8***). The cell density used for the positive control vectors (pT2Neo and pcDNA6.1) and the retrotransposition negative control vector (L1.3-RT$^-$) were the same used with the lowest (2.0 $\times 10^4$ cells/well) and the highest (5.0 $\times 10^3$ cells/well) assayed cellular density respectively. Each cell line was plated in triplicate for each vector as part of the colony forming assay, and in duplicate for the transfection efficiency assessment of each vector (Day = 0). The next day, DNA transfections were performed (Day = 1) (2.3.3.1), and 3 days later the transfection efficiency was assessed by FACS (Day = 4) (2.3.3.2). For neomycin-based colony assays, antibiotic selection was started on the same day transfection Efficiency was measured, using 600 µg/mL Geneticin (***Table 2.2***), followed by 6 more media changes every other day (***Figure 2.1A***). For Blasticidin-based colony assays, the cell supernatant was replaced by culture media, without selection antibiotic, on day 4 and 6. Antibiotic selection was started on day 7, using 5 µg/mL Blasticidin (***Table 2.2***) followed by one more media change on day 11 (***Figure 2.1B***). After antibiotic selection, attached foci were washed with 1× PBS and fixed using 1 mL 2% formaldehyde solution (Sigma-Aldrich, Cat. No. 252549) in 1× PBS, for 30 min at room temperature. Next, plates were rinsed with tap water, and foci stained using 1 mL/well of a 6-well plate of 0.1% (w/v) Crystal violet solution (Sigma-Aldrich, Cat. No. V5265) in molecular grade H_2O, for 10 min at room temperature. Plates were rinsed again with tap water and left inverted to air-dry overnight.

Colonies were counted manually using eCount Colony Counter (Fisherbrand). The average number of colonies per triplicate was calculated and this was normalised to the percentage of transfection efficiency (# $colonies_{TEnorm}$ = # colonies / % transfection efficiency). The activity (percentage) of L1.3 variant compared to their respective positive control, pT2Neo or pcDNA6.1 transfected cells depending on the assay, was calculated (activity($L1.3_{variant}$)= (# $colonies_{TEnorm}$($L1.3_{variant}$)/ # $colonies_{TEnorm}$(control)) × 100). These values were used to compare the activity of each L1.3 variant in the different cell lines. The significance of results was explored using the statistical analyses software GraphPad Prism 9.

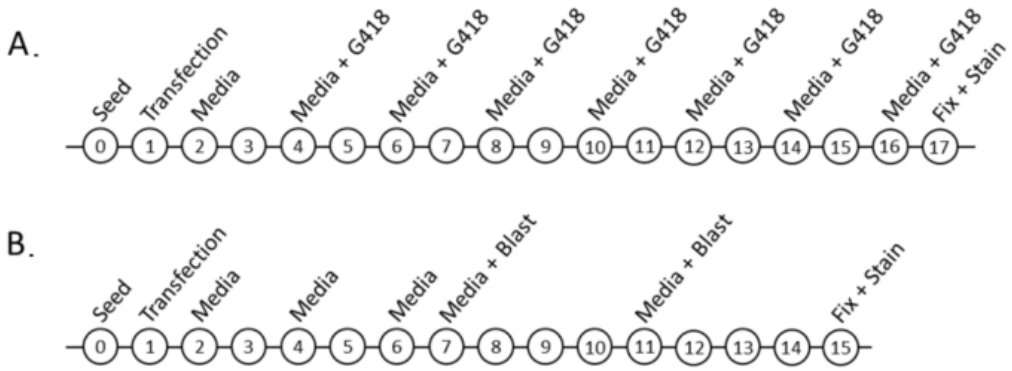


Figure 2.1. Timeline of the different retrotransposition assays - Shown are schemes of colony formation assays using the (A) *mneoI* and (B) *mblastI* tagged retrotransposition vectors.

2.4. Identification of active LINE-1s in the zebrafish genome

2.4.1. Zebrafish husbandry and breeding

Fish were housed at the IGC Zebrafish facility under a 14/10 light/dark cycle. Water temperature and pH were checked daily, and fish were fed two times a day with dry and live food by qualified facility staff. Fish density was always kept under 5 fish/L in system, and under 1 fish/L for breeding tanks.

Fish of the WT background AB and TU, between 1 and 3 years of age, were used for breeding. In the evening, females and males were moved to a breeding tank and separated by a divider. The breeding tank was placed on a hot plate to ensure a water temperature of 28°C. The next morning at 09:00 am, the divider was removed allowing fish to span. Eggs were collected from the breeding tank every 10 min, rinsed with E3 (***Table 2.1***) and kept in E3 medium at 28°C using an incubator.

2.4.2. Sequence analyses of the zebrafish LINE-1 subfamilies

The consensus sequences of seventeen different ZfL1 subfamilies provided by Boissinot & Sookdeo, 2016 were used to retrieve all the full-length copies of each ZfL1 subfamily present in the UCSC genome browser, Zebrafish assembly May 2017 (GRCz10/danRer10), using the BLAT function (https://genome.ucsc.edu/cgi-bin/hgBlat) (Boissinot & Sookdeo, 2016). The sequence of each full-length L1 element, including 150bp of flanking genomic sequence on both sides of each insertion site, were downloaded for further analyses.

The obtained sequences were analysed using the online tool ORFfinder (https://www.ncbi.nlm.nih.gov/orffinder/) (Wheeler et al., 2003). The search parameters were set to 75bp Minimal ORF Length, Standard Genetic Code, and both ATG and alternative ORF initiation codons. The amino acid sequences of the two largest identified ORFs were used for subsequent analyses using the online tool Motif Scan (https://myhits.sib.swiss/cgi-bin/motif_scan) (Artimo et al., 2012). The 'functional' consensus sequence of each of the seventeen different ZfL1 subfamilies was build using SnapGene® 4.3.10 software (Insightful Science).

The amino acid sequences of the proteins encoded in the 'functional' consensus sequences of the fifteen potentially active ZfL1 subfamilies identified with the method described in the previous section [ZfL1-1, -1D, -6, -7B, -8, -10B, -11A, -12A, -12B, -13A, -13B, -13C, -13D, -16B and -17B] were analysed by aligning them to an active human LINE-1 L1-Hs and an active mouse LINE-1 L1-MdA, using the online tool T-Coffee (http://tcoffee.crg.cat/) (Notredame, Higgins, & Heringa, 2000). The parameters to facilitate alignments between distantly related proteins using homology extension, provided by the PSI-Coffee alignment option, were chosen. The presence of conserved amino acids within the functional domains as they were described in Gilbert and Moran (2002) were identified (Christian et al., 2016; Gilbert & Moran, 2002).

2.4.3. ZfL1 expression along zebrafish development

2.4.3.1. Analysis of publicly available RNA-sequencing data

This analysis was performed by a bio-informatician in the team (A. Rubio-Roldan). Briefly, the dataset "Comprehensive identification of long non-coding RNAs expressed during zebrafish embryogenesis" was downloaded from Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) (GEO accession: GSE32898). Using Bowtie2, RNA sequencing data from duplicate samples representing eight zebrafish developmental stages [2-4 cell stage, 1K cell stage, Dome stage, Shield stage, Bud stage, 28 hours post fertilization (28hpf), 2 days post fertilization (2dpf) and 5dpf (Kimmel et al., 1995)], were aligned to the generated functional consensus sequences of the fifteen ZfL1 subfamilies (4.2.1), as well as the consensus sequence of ZfL2-2 element, previously shown in our lab to be expressed along zebrafish development.

To calculate the Fragments Per Kilobase per Million mapped reads (FPKM) values to the full sequence of each zebrafish LINE subfamily, the .bed file containing the zebrafish LINE information (Chromosomal location, sequence, orientation) was fed into the FPKM_count from the RSeqC package (http://rseqc.sourceforge.net/), and the mapped reads were counted and normalized.

The mappability to the first 1 Kb of each zebrafish LINE was obtained by modifying the .bed file to only include the first 1kb of the sequences. Using the same process described above, the modified .bed file was fed into the FPKM_county and the Reads per Million mapped reads (RPM) values were calculated for each zebrafish LINE (sub)family. In this case, the RPM was used instead of the FPKM, since the FPKM has a bias towards short sequences, giving the shorter reads more weight in the fragment.

2.4.3.2. Validation of RNA-sequencing results

2.4.3.2.1. Sample collection at different zebrafish developmental stages

Five females of the WT zebrafish line AB and five males of the WT zebrafish line TU were used for egg production. The eggs were homogenised in ice-cold TRIzol® reagent (Invitrogen, Cat. No. 15596026) (15 eggs/mL) using a disposable Axygen™ Tissue Grinder

(Fisher Scientific, Cat. No. 12649595) and snap frozen on dry ice for later RNA extraction, at different time-points corresponding to the eight different developmental stages analysed by RNA-seq (Table 2.9). The developmental stage of the eggs was confirmed using a stereo microscope (Nikon SMZ645). This process was repeated until the number of eggs specified was obtained for each developmental stage (Table 2.9).

Developmental stage	Egg numbers	Incubation time
2-4 cell stage	60	0.5 h
1K cell stage	45	3 h
Dome stage	30	4.5 h
Shield stage	30	6 h
Bud stage	30	10 h
28 hpf	15	28 h
2 dpf	15	2 days
5 dpf	15	5 days

Table 2.9. Zebrafish developmental stages analysed and their corresponding incubation time in the 28°C incubator.

2.4.3.2.2. Zebrafish embryo RNA

2.4.3.2.2.1. RNA extraction

The homogenised eggs in 1 mL TRIzol reagent (Invitrogen) were thawed in ice. Purification was done by first thoroughly mixing the sample with 200 µL chloroform (Sigma Aldrich, Cat. No. 288306), followed by phase separation through centrifugation (15 min, 17,000 ×*g*, 4°C) and transfer of the aqueous phase (~600 µL) to a fresh tube. Precipitation was done by adding Isopropanol (Fisher Scientific) in ratio 1:1 (v/v), 3M NaAc pH 5.2 (Millipore) in ratio 1:10 (v/v) and 1 µL Glycogen 5 mg/mL (Invitrogen), and incubating the sample for 1 hour at –20°C. RNA was pelleted (30 min, 17,000 ×*g,* 4˚C) and washed using 1 mL ice cold 75% absolute ethanol (Fisher Scientific) followed by centrifugation (15 min, 17,000 ×*g*, 4°C). The pellet was air dried 2-3 min at room temperature and resuspended in 45 µL RNase free H_2O (DPEC-treated water), combining the pellets containing RNA from the same zebrafish developmental stage.

2.4.3.2.2.2. DNA removal

Extracted total RNA was cleaned from carry-over genomic DNA by performing a rigorous DNase treatment using the TURBO DNA-free™ kit (Invitrogen, Cat. No. AM1907). A typical reaction mixture contained: extracted RNA, 1× TURBO DNase Buffer, 4 U TURBO DNase

(2 U/ µL) and RNase free H_2O (DPEC-treated water) up to a total reaction volume of 50 µL. The reaction was incubated at 37°C for 30 min, after which the DNase was inactivated by adding 10 µL of DNase Inactivation Reagent and the reaction incubated at room temperature for 5 min. To remove the DNase Inactivation Reagent beads, the sample was centrifuged (1.5 min, 16,000 ×*g*) and the RNA-containing aqueous phase was transferred to a fresh tube.

2.4.3.2.2.3. Total RNA quality analysis

The removal of the genomic DNA was confirmed by lack of PCR amplification of a ZfL1 copy homozygously present in all zebrafish (4.3.1). A PCR reaction using the genotyping primers and protocol to detect the 5'end of the ZfL1-12B element on chromosome 19 (***Table 6.4***) (2.2.1.2) was performed and analysed by agarose gel electrophoresis (2.2.1.5). As a positive control, a PCR reaction using 1 µL of each sample pre-DNase treatment was included.

If the PCR confirmed an absence of genomic DNA contamination, then samples were further quantified (2.2.1.4), otherwise the DNase step was repeated (2.4.3.2.2.2). 500 ng of each total RNA sample was analysed by electrophoresis on a 1% agarose gel (2.2.1.5) in order to visualise and confirm the integrity of the 28S and 18S ribosomal RNA. The remaining RNA was divided in aliquots of 1 µg, snap frozen on dry ice and stored at –80°C for later use.

2.4.3.2.3. Reverse transcription of RNA

cDNA (DNA complementary to first strand RNA) was generated using SuperScript™ III Reverse Transcriptase (Invitrogen, Cat. No. 18080093) and random primers (Promega, Cat. No. C1181). For each sample, an aliquot containing 1 µg total RNA was thawed on ice (2.4.2.2). A typical reverse transcription reaction contained: 1 µg total RNA, 250 ng of random primers, 500 µM dNTPs (Invitrogen, Cat. No. 10297117), 1× First Strand Buffer, 5mM DTT, 40 U RNaseOUT™ Recombinant RNase Inhibitor (40 U/µl) (Invitrogen, Cat. no. 10777019), 200 U SuperScript™ III RT (200 U/µl) and RNase free H_2O (DPEC-treated water) up to a total reaction volume of 20 µL. For each sample, a "no RT" negative control sample was included. To denature RNA secondary structures, the mix of total RNA, random primers, dNTPs and RNase free H_2O (DPEC-treated water) up to a final volume of 13 µL was first incubated at 65°C for 5 min followed by a 1 min incubation on ice, before addition of the other reagents. The reverse

transcription reaction was performed on a DNA Engine Tetrad 2 thermal cycler (Bio-Rad) using the following program:

Step	Temperature(°C)	Time(min:sec)
Annealing	25	5:00
Reverse transcription	50	60:00
Enzyme deactivation	70	15:00

cDNA was stored at –20°C for later use.

2.4.3.2.4. Quantitative PCR (qPCR)

qPCR was performed using GoTaq® qPCR Master Mix (Promega, Cat. No. A6001) and primer pairs specific to each zebrafish LINE (sub)family. An aliquot containing 20 µL of cDNAs (obtained from 1 µg total RNA), from each zebrafish developmental stage analysed, was thawed on ice (2.4.3.2.3), and molecular grade H_2O was added up to a total volume of 45 µL. A typical qPCR reaction contained: 2 µL diluted cDNA, 1× GoTaq® qPCR Master Mix, 0.5 µM forward primer, 0.5 µM reverse primer (***Table 6.3***) and molecular grade H_2O up to a total reaction volume of 10 µL. Reactions were run in a white LightCycler® 480 Multiwell Plate 384 (Roche, Cat. No. 04729749001) covered with optic sealing foil. The experimental set up involved allocating the primer sets targeting the different ZfL1 subfamilies in groups of 4 on different plates. Each plate also included the primer sets targeting the three control genes: the LINE-2 element ZfL2-2, Elongation Factor 1-alpha (EF1-α), and beta-Actin (Act-β). Triplicate reactions were conducted for each cDNA sample and the corresponding RT negative control, for each of the eight zebrafish developmental stages tested (2.4.3.1). The qPCR reaction was performed on a LightCycler® 480 Instrument II (Roche) using the following program:

Step	Temperature(°C)	Time(min:sec)	Number of cycles
Initial denatu°n	95	2:00	1
Denaturation	95	0:15	40
Annealing	60	1:00	

The most consistently expressed control gene, EF1-α *vs* Act-β, across the different zebrafish developmental stages was determined by calculating their Coefficient of Variation (%CV) using the formula: $\%CV_{gene}$(developmental stages) = (Standard Deviation (Ct_{gene} developmental stages)/ mean (Ct_{gene} developmental stages)) × 100%. The control gene possessing the smallest %CV was used to calculate the Fold change (Fc) of each zebrafish LINE

element (sub)family per developmental stage. First, for each developmental stage, each zebrafish LINE element (sub)family was normalising to the Ct value of the most consistently expressed control gene (determined by the previously described method) (ΔCt), using the formula: $\Delta Ct_{ZfL} = Ct_{ZfL} - Ct_{control\ gene}$. Using this ΔCt, the Fc was calculated using the formula: $Fc = 2^{-\Delta Ct_{ZfL}}$.

2.4.4. Retrotransposition activity of ZfL1 subfamily members

2.4.4.1. Amplification of potential retrotransposition competent ZfL1 copies

Putative active ZfL1 copies, identified in the zebrafish genome (4.2.1), were amplified using the Roche Expand™ Long Range dNTPack in combination with designed primers, covering the 5' and 3' junctions with the zebrafish genome while incorporating two unique restriction sites at both ends of ZfL1s (AscI and SrfI, absent in all ZfL1 sequences analysed and in the cloning vectors pCEP4-*mneoI* and pTOL2-EGFP-NSL vectors, as revealed using SnapGene® 4.3.10 software (Insightful Science)) (2.2.1.2; ***Table 6.5***). Half of this PCR reaction was resolved on a 1% agarose gel by electrophoresis, excised and gel-purified (2.2.1.5; 2.2.1.6.1). The purified PCR fragments were then cloned in TOPO-TA as described later (2.2.1.7; 2.4.4.2). To confirm that the sequence of the amplified ZfL1 elements match that found in the UCSC genome browser, the other half of the PCR reaction was sent for capillary DNA sequencing (2.2.1.8; ***Table 6.6***).

2.4.4.2. Generation of ZfL1 retrotransposition vectors

Two retrotransposition vectors, pCEP4-L1.3-meoI and pTOL2-Zfl2-2-GFP-NSL, previously used by our team to study human LINE-1 activity *in vitro* using cultured cells, and zebrafish LINE-2 activity *in vivo* using zebrafish embryos respectively, were modified to include restriction sites allowing the cloning of the amplified ZfL1 copies (2.4.4.1). Briefly, a DNA fragment containing the following restriction sites from 5' to 3': NotI, EcoRV, AscI, SrfI, ClaI and BastZ17I (5' – AGAGCGGCCGCGATATCGGCGCGCCATAGCCCGGGCATCGATGTATACCCG – 3'), was generated *in vitro* by hybridising 2 ordered oligonucleotides in the presence of salts (***Table 6.5***). The pCEP4-L1.3-*mneoI* and pTOL2-Zfl2-2-EGFP-NSL vectors (***Table 2.3***) were subjected to NotI-BstZ17I and EcoRV-ClaI double digestions respectively, removing the LINE elements from these two vectors (2.2.1.3). The DNA fragment was digested, in separate, by these two

combination of enzymes, ligated into the vectors (2.2.1.7) and transformed into chemical-competent DH5α *E. coli* strain (2.2.2.3.1). Colonies were subjected to colony PCR (2.2.1.2), and miniprep DNA was capillary sequenced to confirm the identity of clones (2.2.1.8).

The vectors pCEP4-*mneoI*-3'UTR and pTOL2-EGFP-NSL-3'UTR, containing the different ZfL1 subfamily (ZfL1-7B, ZfL1-10B and ZfL1-12B) 3'UTRs behind the retrotransposition cassettes, were built from the modified retrotransposition vectors described above. The different ZfL1 3'UTRs were amplified using Q5® High-Fidelity DNA Polymerase (2.2.1.2.2) from one element of each family carrying the standard sequence, and the BamHI and HpaI restriction sites were incorporated at the 5' and 3' end respectively (2.2.1.2) (***Table 6.7***). The amplicons were gel-purified following a 2% agarose gel electrophoresis (2.2.1.5 and 2.2.1.6.1). Both, vectors and amplicons were subjected to BamHI-HpaI double digestion (2.2.1.3), the vectors were gel-purified following a 1% agarose gel electrophoresis (2.2.1.5 and 2.2.1.6.1) and the amplicons were directly purified (2.2.1.6.2). These 3'UTR amplicons were then ligated to the vectors (2.2.1.7) and transformed into chemically-competent DH5α *E. coli* strain (2.2.2.3.1). Colonies were subjected to colony PCR (2.2.1.2), and miniprep DNA was capillary sequenced to confirm the identity of clones (2.2.1.8).

The amplified full-length ZfL1 copies extracted from agarose gel (2.4.4.1) were cloned into TOPO-TA vectors (2.2.1.7) and the vectors were transformed into electro-competent Stbl4 *E. coli* strain (2.2.2.3.2). Colonies were subjected to colony PCR (2.2.1.2) and vectors containing a ZfL1 copy (miniprep) were capillary sequenced (2.2.1.8) using stepping primers ~500bp along the element (2.2.1.2.1; ***Table 6.6***) for full sequencing coverage. When necessary, specific cloning strategies were designed to eliminate PCR-induced mutations. A 'resurrected' ZfL1 matching each ZfL1 L1.7B, L1.10B and L1.12B subfamily consensus was also generated, by combining fragments of different ZfL1 elements copies and/or site directed mutagenesis (***Table 6.8***). Once a given ZfL1 element copy was cloned, matching its genomic sequence or the consensus sequence, this element was subcloned in the *in vitro* and *in vivo* retrotransposition vectors (i.e., pCEP4-*mneoI* and pTOL2-EGFP-NSL respectively) using the AscI and SrfI sites. The elements were amplified from PCR-mutation-free copies in TOPO vectors using Q5® High-Fidelity DNA Polymerase (2.2.1.2.2), removing the poly(A) signal in its 3'UTR and incorporate an SrfI restriction (***Table 6.5***). These PCRs and the modified *in vitro* and *in vivo* retrotransposition vectors were subjected to AscI-SrfI double digestion (2.2.1.3). The correct

fragments were gel-purified following a 1% agarose gel electrophoresis (2.2.1.5 and 2.2.1.6.1), and inserts were ligated into the vectors (2.2.1.7) and transformed into electro-competent Stbl4 *E. coli* strain (2.2.2.3.2). Colonies were subjected to colony PCR (2.2.1.2), and miniprep DNA was capillary sequenced (2.2.1.8). To also clone these elements into the *in vitro* and *in vivo* retrotransposition vectors pCEP4-*mneoI*-3'UTR and pTOL2-EGFP-NSL-3'UTR, the elements were amplified from PCR-mutation-free copies in TOPO vectors using Q5® High-Fidelity DNA Polymerase (2.2.1.2.2), incorporating the SrfI site immediately downstream of the ORF2 sequence (2.2.1.2) (***Table 6.7***). They were next cloned into the AscI-SrfI sites of the pCEP4-*mneoI*-3'UTR and pTOL2-EGFP-NSL-3'UTR as for the constructs above, and further processed in the same way.

To also track translation of *In Vitro* Transcribed (IVT) RNA (2.4.4.3), a new reporter system was designed, with ZfL1 and ZfL2-2 elements containing the mCherry sequence fused in-frame to the ORF1p or ORFp respectively, using a self-processing 2A sequence (***Table 6.9***). The mCherry sequence was amplified from a standard mCherry containing vector using Q5® High-Fidelity DNA Polymerase (2.2.1.2.2), incorporating a 2A sequence at its 3'end. The 5' of the consensus ZfL1 sequences, and ZfL2-2 were amplified using Q5® High-Fidelity DNA Polymerase (2.2.1.2.2), using a forward primer upstream of the pTOL2 cloning site (pTOL2_F) and a reverse primer just before the start of the ORF1 or ORF of the ZfL1 and ZfL2-2 respectively, incorporating 19 bp of the 5'end of the mCherry sequence. Also the 5' end of the ORF1 or ORF of the ZfL1 and ZfL2-2 respectively were amplified using Q5® High-Fidelity DNA Polymerase (2.2.1.2.2), incorporating 19 bp of the 3' end of the previously generated mCherry-2A amplicon. These 3 PCR reactions were then mixed in equal proportions and used as a template for a final round of amplification using Q5® High-Fidelity DNA Polymerase (2.2.1.2.2), using the (pTOL2_F) and the reverse primer in the ORF (used for the last described PCR reaction). The amplicons were gel-purified following a 2% agarose gel electrophoresis (2.2.1.5 and 2.2.1.6.1), and cloned into the pTOL2-EGFP-NSL-3'UTR containing the matching ZfL1 element with consensus sequence, or the pTOL2-ZfL2-2-EGFP-SL (***Table 2.4***) making use of the AscI and a unique cloning site in each ORF upstream of the utilised primer. pTOL2-ZfL2-2-2A-mCherry-EGFP-SL allelic variants were built in a similar way using this vector as a starting point (***Table 6.10***).

2.4.4.3. In vitro transcription of zebrafish LINE RNAs for retrotransposition assays

Zebrafish LINE RNAs for zebrafish embryo injection were IVT from the pTOL2 vectors (***Table 2.4***, ZfL2-2-WT; ***Table 2.4***, pMK013 - pMK035, and pMK058 - pMK062). The vectors were first linearized utilising a restriction site located downstream of the antisense EGFP reporter cassette (***Table 2.4***, pMK014 - pMK024) or located downstream of the zebrafish LINE 3'UTR, after the antisense EGFP reporter cassette (***Table 2.4***, ZfL2-2-WT; ***Table 2.4***, pMK013, pMK025 - pMK035, and pMK058 - pMK062). *In vitro* transcription of these linearized vectors was carried out using the using the mMESSAGE mMACHINE™ T7 Transcription Kit (Invitrogen, Cat. No. AM1344). The resulting RNAs were *in vitro* poly-adenylated using the Poly(A) Tailing Kit (Invitrogen, Cat. No. AM1350). Finally, the RNA quality was assessed by confirming their integrity (no degradation) and purity (no vector contamination).

2.4.4.3.1. Vector linearization

3 µg of the ZfL1 and ZfL2 containing vectors were linearized using HpaI and BamHI digestion respectively (2.2.1.3), for 3 hours incubation at 37°C. 1 µL of the digested vectors were analysed by electrophoresis on 1% agarose gel to confirm linearization (2.2.1.5). The digestion was purified (2.2.1.6.2), the DNAs were resuspended in 12 µL RNase free H_2O (DPEC-treated water) and the DNA concentration of samples was quantified (2.2.1.4).

2.4.4.3.2. RNA *in vitro* transcription

A typical *in vitro* transcription reaction using the mMESSAGE mMACHINE™ T7 Transcription Kit (Invitrogen) contained: 1 µg linearized vector, 1× NTP/CAP2, 1× reaction buffer, 1 mM GTP, 2 µL Enzyme mix and RNase free H_2O (DPEC-treated water) up to a total reaction volume of 20 µL. The reaction was incubated 3 hours at 37°C. The vector DNA was removed by adding 1 µL TURBO DNase and incubating 30 min at 37°C.

2.4.4.3.3. *In vitro* poly-adenylation

A typical *in vitro* poly-adenylation reaction using the Poly(A) Tailing Kit (Invitrogen) contained: 20 µL mMESSAGE mMACHINE reaction, 1× E-PAP buffer, 2.5 mM $MnCl_2$ (25 mM) , 1 mM ATP (10 mM), 8 U E-PAP (2 U/ µL) and RNase free H_2O (DPEC-treated water) up to a total reaction volume of 100 µL. The reaction was incubated 1 hour at 37°C. The RNA was

precipitated by adding 150 µL RNase free H_2O (DPEC-treated water) and 150 µL Lithium Chloride, and incubating overnight at –20°C. RNA was pelleted (1 hour, 17,000 ×*g*, 4°C), washed with 70% Ethanol and resuspended in 44 µL RNase free H_2O (DPEC-treated water).

2.4.4.3.4. IVT RNA quality analysis

Any residual vector DNA was removed by performing a rigorous DNase treatment using the TURBO DNA-free kit (Invitrogen) (2.4.3.2.2.2). The absence of vector DNA was confirmed by lack of PCR amplification of a vector region downstream of the restriction site used to linearize the vector. A PCR reaction using the DreamTaq Green PCR Master Mix in combination with a primer pair targeting this vector region was performed (2.2.1.2) (***Table 6.11***). IVT RNAs were analysed by electrophoresis on 1% agarose gels to confirm good integrity of the RNAs (2.2.1.5). Then RNAs were divided in 2 µg aliquots and stored at –80°C.

2.4.4.4. ZfL1 retrotransposition assays

2.4.4.4.1. *In vitro* retrotransposition assays

2.4.4.4.1.1. Vector-based assays

Vector-based *in vitro* retrotransposition assays were done using the ZfL1 elements cloned in the pCEP4 vectors (***Table 2.4***). Vectors containing a retrotransposition competent human LINE-1 L1.3 (JM101), or a retrotransposition incompetent LINE-1, L1.3-RT^{-} (JM105) were used as positive and negative control respectively (***Table 2.3***). Triplicates were done for each tested vector. HeLa cells were seeded in 6-well tissue culture plates at a cellular density of 2 ×10^5 cells/well (Day = 0), and transfected 16-18h later as described (Day = 1) (2.3.3.1). On day 4, episomal vector selection started using 200 µg/mL Hygromycin (***Table 2.2***) for 5 days, using untransfected cells as a negative control. The Hygromycin resistant (HygR) cells were trypsinized, counted and reseeded into 100-mm dishes at densities of ∼1×10^5 cells/dish (2.2.3.1; 2.2.3.2). On day 10, selection for retrotransposition events started using 600 µg/mL Geneticin (***Table 2.2***) for 12 days. On day 21, cells were fixed using 2% formaldehyde solution (Sigma-Aldrich) in 1× PBS and stained with 0.1% (w/v) crystal violet solution (1%) (Sigma-Aldrich) in H_2O (2.3.3.3). After washing and drying, G-418 resistant (G418R) colonies were manually counted and the retrotransposition frequency was calculated as the number of G-

418R colonies per plate. The significance of results was explored using a nonparametric t-test using the statistical analyses software GraphPad Prism 9.

2.4.4.4.1.2. RNA transcript-based assays

IVT ZfL1 RNA were used for *in vitro* retrotransposition assays using a HeLa cell line, with IVT RNAs from ZfL2-2-WT and ZfL2-2-RT^{-} RNA as positive and negative controls respectively (2.4.4.3). In each independent experiment, each vector was tested in triplicate. HeLa cells were seeded in a 6-well tissue culture plate at a density of 2 ×10^5 cells/well (Day = 0). 16-18h later, the media was changed to penicillin/streptavidin-free media, and cells were transfected using Lipofectamine 2000 Transfection Reagent (Invitrogen, Cat. No. 11668019) according to manufacturer's protocol (Day = 1). Briefly, for each well of a 6-well tissue culture plate, 5 µL Lipofectamine 2000 Transfection Reagent (Invitrogen) was mixed with 250 µL Opti-MEM Medium (Gibco) and incubated 5 min at room temperature. In parallel, for each well, 1 µg of IVT RNA was mixed with 250 µL Opti-MEM Medium (Gibco). Both mixes were combined and incubated for 20 min at room temperature. 500 µL of the final mix was added to each corresponding well. Media was changed after 4 h. The set of transfection replicates were performed twice, using one set for analysing the percentage of cherry and/or EGFP positive cells and fluorescence intensity, 1 day and 2 days after transfection, using the BD LSR Fortessa Cell Analyser (Biosciences) (2.2.3.4). The significance of results was explored using a nonparametric t-test using the statistical analyses software GraphPad Prism 9.

2.4.4.4.2. *In vivo* retrotransposition assays

In vivo retrotransposition assays were done using the ZfL1 IVT and poly-adenylated RNAs, and IVT ZfL2-2 RNA as a positive control (2.4.4.3). Uninjected embryos were used as a negative control. In each independent experiment, each vector was injected in ±50 embryos. RNAs were diluted in RNase free H_2O (DPEC-treated water) to concentrations of 50, 100 or 200 ng/µL and kept on ice until injection. Microneedles were pulled from 1.00 mm × 0.80 mm × 10 cm OD glass capillary (Intrafil) using a micro-electrode puller instrument (CFP) with settings, 4 initial pull, 8 collets, 7 furnace. A needle was loaded with zebrafish LINE RNA, inserted into Picospritzer III microinjector (Intracel) with Narishige needle arm (Intracel) and the pressure was adjusted to eject 1 nL droplets (0.2 mm). Eggs were collected from the breeding tanks and

single-cell zebrafish embryos were injected with RNA. Zebrafish larvae were analysed by fluorescence microscopy at 1 dpf, 2 dpf and 3 dpf using a stereomicroscope (Leica). At 3 dpf, larvae were incubated in proteinase K containing lysis buffer at 60°C for 6 h, followed by proteinase K inactivation by incubation at 95°C for 10 min. The lysate was used for PCR validation of integrated EGFP, using DreamTaq Green PCR Master Mix in combination with primers specific to the EGFP gene (2.2.1.2) (*Table 6.11*). To ensure positive reading was not due to integration of residual vector DNA, we included a control PCR targeting a vector region downstream of the restriction site used to linearize the vector before the IVT reactions (2.4.4.3) (*Table 6.11*).

Chapter 3

RNase H involvement in LINE-1 retrotransposition

3. RNase H involvement in LINE-1 retrotransposition

Background

As described in the introduction, LINE-1 mobilisation occurs through TPRT. During TPRT, the EN domain of L1-ORF2p recognises and cleaves a consensus sequence (5´TTTTT/AA and variants; / denotes the site of cleaving) to expose a 3´OH group (Feng et al., 1996; Flasch et al., 2019; Jurka, 1997). The 3'end of the LINE-1 RNA anneals with this exposed genomic DNA flap, which now functions as a primer, initiating the reverse transcription of the LINE-1 RNA into cDNA by the RT domain present in ORF2p (Monot et al., 2013). This creates an RNA:DNA hybrid with the DNA moiety covalently linked to the genome. To complete the integration of the new LINE-1 copy, the RNA moiety has to be removed/displaced to allow the use of the cDNA as a template for second strand synthesis (Cost et al., 2002). However, how second strand cDNA synthesis occurs and retrotransposition intermediates are resolved remains unclear. Although enzymatic activities from L1-ORF2p might be involved in these processes, host cellular factors have been suggested to also play a vital role in the completion of LINE-1 retrotransposition (Benitez-Guijarro et al., 2018; Gasior et al., 2006; Suzuki et al., 2009; Taylor et al., 2013).

3.1. PCNA and retrotransposition

One host factor that has been shown to be involved in the LINE-1 retrotransposition cycle is PCNA. PCNA is a 87 kDa, homotrimeric ring shaped protein, known as a DNA sliding clamp, that encircles and slides along double stranded DNA (Ellison & Stillman, 2003; Tsurimoto & Stillman, 1990). It plays a key role in DNA replication and repair, where it functions as a scaffolding protein to tether other factors involved in these processes to the genomic template, and coordinates their interactions (De March et al., 2017; Essers et al., 2005; reviewed in Moldovan, Pfander, & Jentsch, 2007). Particularly, it has been shown to interact and regulate the activities of many proteins involved in the maturation of Okazaki fragments (Levin et al., 1997; Warbrick et al., 1997), mismatch and nucleotide excision repair (Gary et al., 1997; Umar et al., 1996), and translesion synthesis (Haracska et al., 2001a, 2001b, 2002). These different proteins can interact with PCNA through the outer surface of its ring structure,

primarily through a hydrophobic pocket present on the front face (Bruning & Shamoo, 2004; Gulbis et al., 1996; Hishiki et al., 2009). This face points in the direction of DNA synthesis, allowing interacting proteins to access the primer terminus of the replicating DNA (Jónsson, Hindges, & Hübscher, 1998). PCNA interacting proteins generally contain one or more PIP motifs which facilitate their interaction with PCNA's hydrophobic pocket (Prestel et al., 2019). Recently, a functional PIP motif has been characterised in L1-ORF2p, allowing its interaction with PCNA (Taylor et al., 2013). The L1-ORF2p PIP motif is situated in the Z domain, located adjacent to the EN domain (Christian et al., 2016; Taylor et al., 2013). The presence of the PIP motif is highly conserved across LINE-1 and LINE-1-like elements in different species (Taylor et al., 2013). Indeed, deletion or mutation of this motif has been shown to dramatically reduce retrotransposition of both LINE-1 and *Alu* elements (mobilised in-trans) (Christian et al., 2016; Clements & Singer, 1998; Taylor et al., 2013). Investigating the role of PCNA in retrotransposition, in 2013 Taylor and co-workers demonstrated that PCNA KD also resulted in a reduction of LINE-1 retrotransposition, which correlated with decreased PCNA levels. Further experiments assessing the association of PCNA with LINE-1 mutants lacking EN and RT catalytic activity respectively, suggested that LINE-1/PCNA interaction takes place after the EN domain cleaved the target DNA, and during/after L1 cDNA synthesis.

Based on these findings, and the known function of PCNA in DNA replication and repair, three hypothesis were proposed to explain how PCNA facilitates LINE-1 retrotransposition: (1) PCNA may tether ORF2p to genomic DNA enabling it to efficiently scan for a preferred target site for integration; (2) PCNA may increase the processivity of the L1 RT; or (3) the presence of PCNA may facilitate the completion of the final integration steps of retrotransposition, by promoting the repair of the formed DNA nicks or gaps, or removing RNA flaps at the junctions between retrotransposon and host DNA. More recently, a fourth hypothesis was proposed based on work from our research team in collaboration with A. Jackson's group: (4) that PCNA may facilitate the interaction between L1-ORF2p and other host factors involved in the LINE-1 retrotransposition cycle, specifically RNase H2 (Benitez-Guijarro et al., 2018).

3.2. RNase H and Retrotransposition

Cellular RNase H activity, amongst others provided by RNase H2, is a proposed likely candidate for the removal of the LINE-1 RNA from the RNA:DNA hybrid generated during retrotransposition (Benitez-Guijarro et al., 2018; Malik et al., 1999). Retroviruses, LTR-retrotransposons and some non-LTR retrotransposons, mainly present in the genomes of plants and lower eukaryotes, encode their own functional RNase H domain (Beilhartz & Götte, 2010; Fawcett et al., 1986; Malik et al., 1999; Mcclure, 1991; Olivares et al., 2002). However, most LINEs lack RNase H domains, as is the case for the L1 clade of retrotransposons, which includes mammalian LINE-1 elements, suggesting that cellular RNase H activity can readily provide this activity during TPRT. Mammalian cells possess two types of RNases H: RNase H1 and RNase H2; both sharing the ability to enzymatically degrade RNA from RNA:DNA hybrids (reviewed in Cerritelli & Crouch, 2009). RNase H2 has the additional ability to cleave the 5'-phosphodiester bond of genomic sites containing misincorporated single ribonucleotides, which is critical for the repair of this common DNA lesion (Hiller et al., 2012; Mcelhinny et al., 2010; Reijns et al., 2012). Both enzymes are localised in the nucleus, where RNASEH1 is heavily involved in the prevention and removal of R-loops and consequential transcription-replication conflicts (Nguyen et al., 2017; Parajuli et al., 2017; Shen et al., 2017), while RNase H2 is predominantly responsible for RNA:DNA hybrid degradation (reviewed in Reijns & Jackson, 2014). RNASEH1 is monomeric, while RNase H2 is a heterotrimeric complex consisting of three subunits: RNASEH2A, B and C (Jeong et al., 2004). Its catalytic core is present in the RNASEH2A subunit and, although the precise role of the other two subunits is unclear, all are required for the stability of the complex and hence RNase H2 activity (Chon et al., 2009). Notably, a PIP motif has been described in the RNASEH2B subunit, which allows the RNase H2 complex to interact with PCNA and direct its activity towards replication and repair foci (Bubeck et al., 2011; Chon et al., 2009).

RNase H2 has recently been described as being required for efficient LINE-1 retrotransposition in human cells (Bartsch et al., 2017; Benitez-Guijarro et al., 2018). Benitez-Guijarro and co-workers took a closer look at the mechanisms underlying this phenomenon in a study performed by our research team in collaboration with A. Jackson's group. We found that RNASEH2A CRISPR/cas-9 KO in HeLa, HCT116 and U2OS cell lines resulted in the reduction of retrotransposition activity of an engineered WT human LINE-1 element. Consistently, we found reduced retrotransposition of a zebrafish LINE-2 retrotransposon in RNase H2 KO cells,

that also lack a proper RNase H domain (Sugano et al., 2006). In contrast, retrotransposition of a mouse LTR-retrotransposon that codes its own RNase H domain was not influenced by the cellular RNase H2 status. Complementation of these HeLa KO cell lines with a "Separation of Function" (SoF) RNASEH2A mutant, lacking its activity against single-embedded ribonucleotides but maintaining the ability to degrade RNA from RNA:DNA hybrids, suggested that the function of RNase H2 in the retrotransposition cycle is dependent on the later activity. Finally, the fact that RNASEH1 overexpression was only able to partially rescue the retrotransposition deficit observed in RNASEH2A KO cell lines, suggested that other factors besides the activity against RNA:DNA hybrid are important in the role of RNase H2 in the LINE retrotransposition cycle. We proposed that the presence of the PIP motif in the RNASEH2B subunit of RNase H2 (Bubeck et al., 2011; Chon et al., 2009), not present in RNASEH1, might be one of these factors.

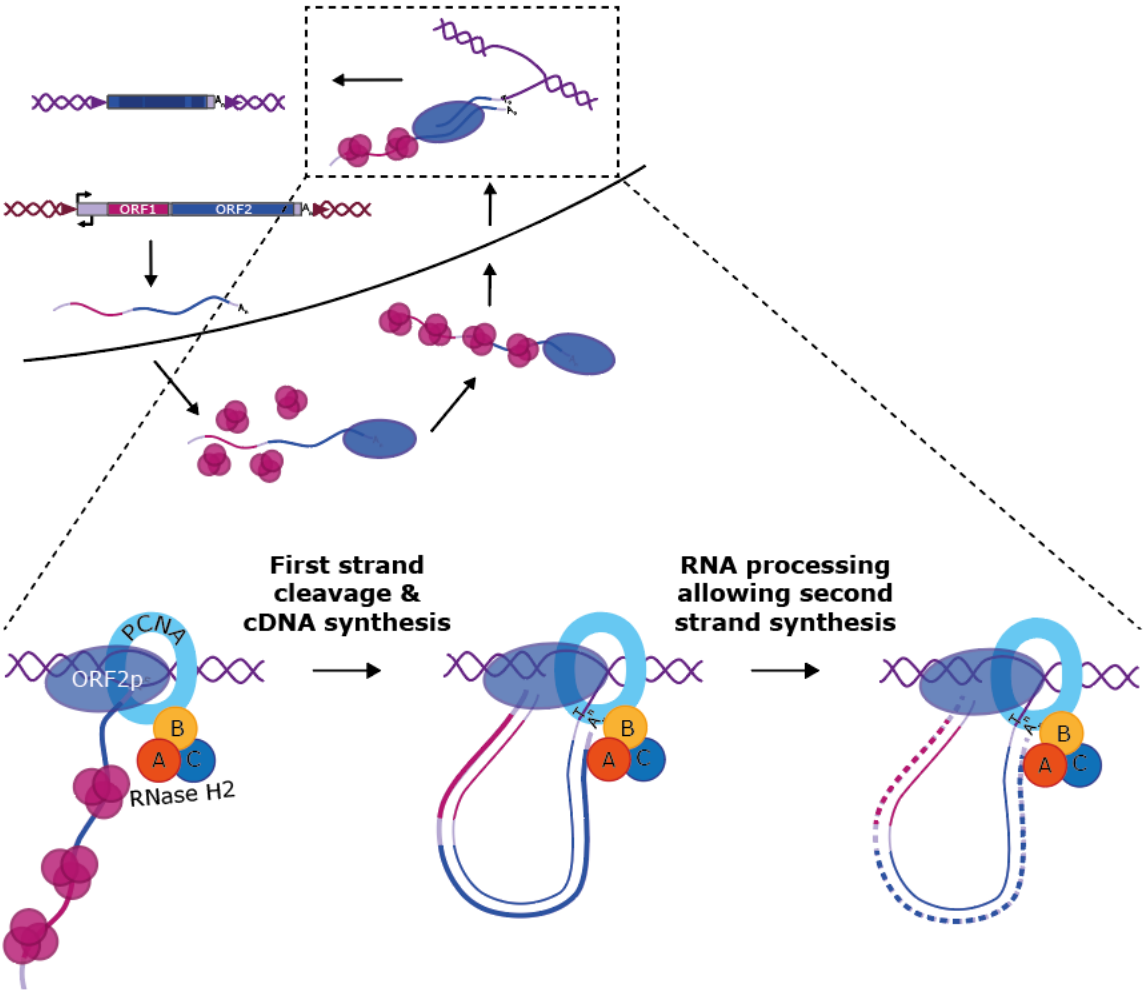


Figure 3.1. Hypothesis of PCNA and RNase H2 involvement in the LINE-1 retrotranspositon cycle – Schematic representation of the hypothesised involvement of PCNA and RNase H2 in the LINE-1 retrotransposition cycle. Figure adapted from Benitez-Guijarro et al. (2018).

Indeed, the timing of LINE-1/PCNA interaction aligns well with the need for RNase H activity within the retrotransposition cycle. Therefore, we hypothesized here that PCNA might be functioning as an anchor protein for RNase H2 and L1-ORF2p through their respective PIP motifs, ensuring their co-localisation during LINE-1 TPRT, when L1-RNA removal from the RNA:cDNA hybrid is required to complete a round of retrotransposition (***Figure 3.1***).

3.3. Aims and objectives

In this chapter, I describe a series of experiments aimed to get a better understanding of the role of RNase H2, and indirectly PCNA, in the retrotransposition cycle. I particularly focus on (1) the involvement of the PIP motif of RNase H2 and L1-ORF2p in TPRT, (2) the level of RNase H2 activity necessary to support efficient LINE-1 retrotransposition, and (3) whether introducing a functional RNase H domain in the ORF2p of a human LINE-1 eliminates its dependence on host-encoded RNase H2.

Results

3.4. The involvement of the PIP motif of RNase H2 and L1-ORF2p in LINE-1 retrotransposition

3.4.1. Generation and characterisation of RNASEH2B HeLa cell models

3.4.1.1. Generation and characterisation of RNASEH2B KO cell lines

To investigate the role of the PIP motif of RNase H2 in TPRT, I first generated HeLa RNASEH2B KO cell lines using CRISPR-Cas9:sgRNA RNPs (2.3.1.1). The HeLa cell stock used to generate this KO cell line was the same used in the study by Benitez-Guijarro and co-workers in 2018, where the role of RNASEH2A on retrotransposition was explored. This was decided in order to avoid potential variables introduced by differences in HeLa cell lines (Liu et al., 2019), making the work fully comparable. The RNASEH2B subunit was targeted since it harbours the PIP motif of the RNase H2 enzyme. Using the Synthego CRISPR Design Tool, three sgRNA targeting the exon 3 of the human *RNASEH2B* gene were designed (***Figure 3.2***).

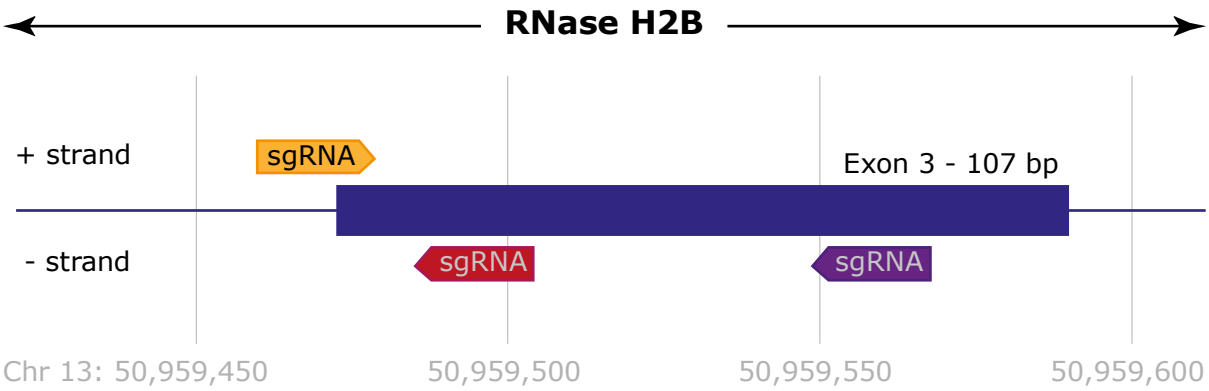


Figure 3.2. **Synthego sgRNAs** – Location of the three sgRNAs designed using the online Synthego CRISPR Design Tool, targeting exon 3 of the human RNAase H2B gene.

The experimental workflow is summarised in ***Figure 3.3***. Briefly, HeLa cells were electroporated with Cas9:sgRNA RNPs containing an equal mix of the three sgRNAs (***Figure 3.2***) (2.3.1.1). Transfected cells were subjected to single cell sorting (2.3.1.3) and clonal expansion. The region of the *RNASEH2B* gene targeted by the sgRNAs of 192 expanded clones were analysed using PCR and capillary DNA sequencing, to confirm the presence/absence of deleterious mutations (2.2.1.8). The capillary DNA sequencing suggested none of the analysed clones possessed a true WT genotype (***Figure 6.9***) (which is consistent with the observed discrepancy of measured RNase H2 activity and LINE-1 retrotransposition acvitivity in the selected control RH2B KO cells (3.4.1.2; 3.7)). Clones with near WT genotypes, as well as clones containing inactivating mutations in the *RNASEH2B* genes, were further expanded and their RNase H2 activity against single-embedded ribonucleotides was assessed (***Figure 3.4A***; 84 analysed, but only 26 shown) (2.3.1.3.2). Based on the measured RNase H2 activity levels, the general cell morphology and growth rates, a set of clonal cell lines was selected for further experimentation, including complete RNase H2 KO clonal cell lines as well as clonal cell lines with near WT RNase H2 activity levels. These selected clonal cell lines were also characterised by western blot probing for the three RNase H2 subunits (2.3.1.3.3) (***Figure 3.4B***).

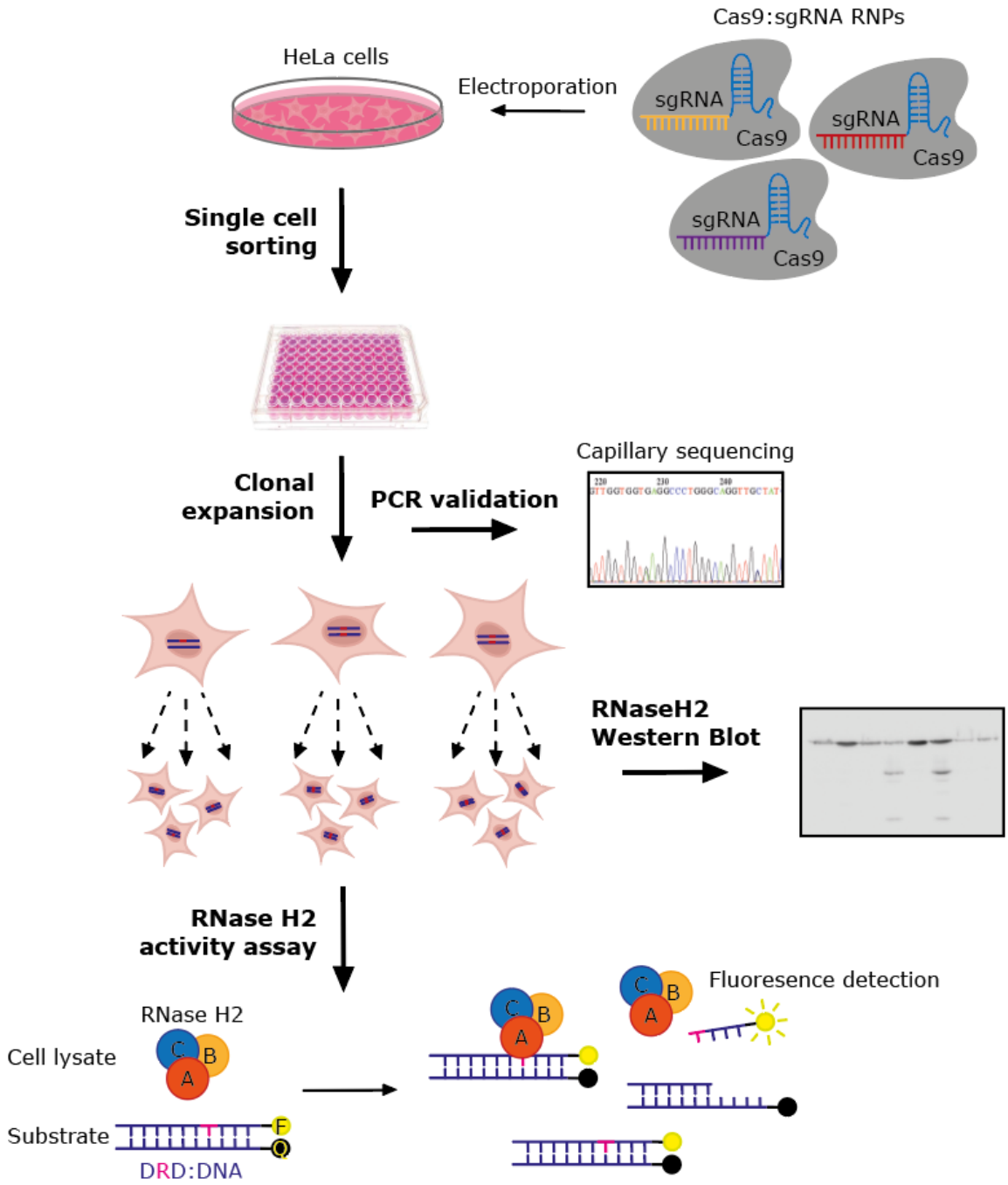


Figure 3.3. Generation and characterisation of HeLa RNASEH2B KO cell lines –Workflow used to generate and characterise the HeLa RNASEH2B KO cell lines is shown. For the RNase H2 activity assay, fluorescence is only measured when the RNase H2 cleaves the substrate at the strand containing the ribonucleotide (DRD), separating the attached fluorophore (F) from the quencher (Q) on the other strand. Further details can be found in the main text.

Capillary DNA sequencing of PCRs amplifying the sgRNA targeted region of the *RNASEH2B* gene indicated that all of the analysed clones that had undergone the CRISPR KO process, had all or some of their *RNASEH2B* alleles altered. Nonetheless, two of the clonal cell lines (16 and 26) subjected to the RNase H2 activity assay possessed WT activity levels

comparable to that found in the parental cell line (HeLa WT) (*Figure 3.4A*). These cell lines were selected as internal negative controls (i.e., control lines). All other analysed clones showed RNase H2 activity levels comparable to that of a previously established RNASEH2A KO cell line model (RH2A KO) (*Figure 3.4A*; 24 of the 82 shown). Of these clonal cell lines, six were selected for further experiments (3, 6, 8, 12, 18 and 23), paying special attention to their morphology and growth rates, selecting those resembling that of the parental HeLa cells (HeLa WT) (data not shown). Western blot of the three RNase H2 subunits indicated that protein levels correlated with the reported levels of RNase H2 activity (*Figure 3.4B*). To note, the antibody used to detect the three RNase H2 subunits, has preferential binding to the RNASEH2B subunit. Nonetheless, it is clear that RNASEH2B depletion also leads to a significant reduction of the other two subunits, RNASEH2A and C.

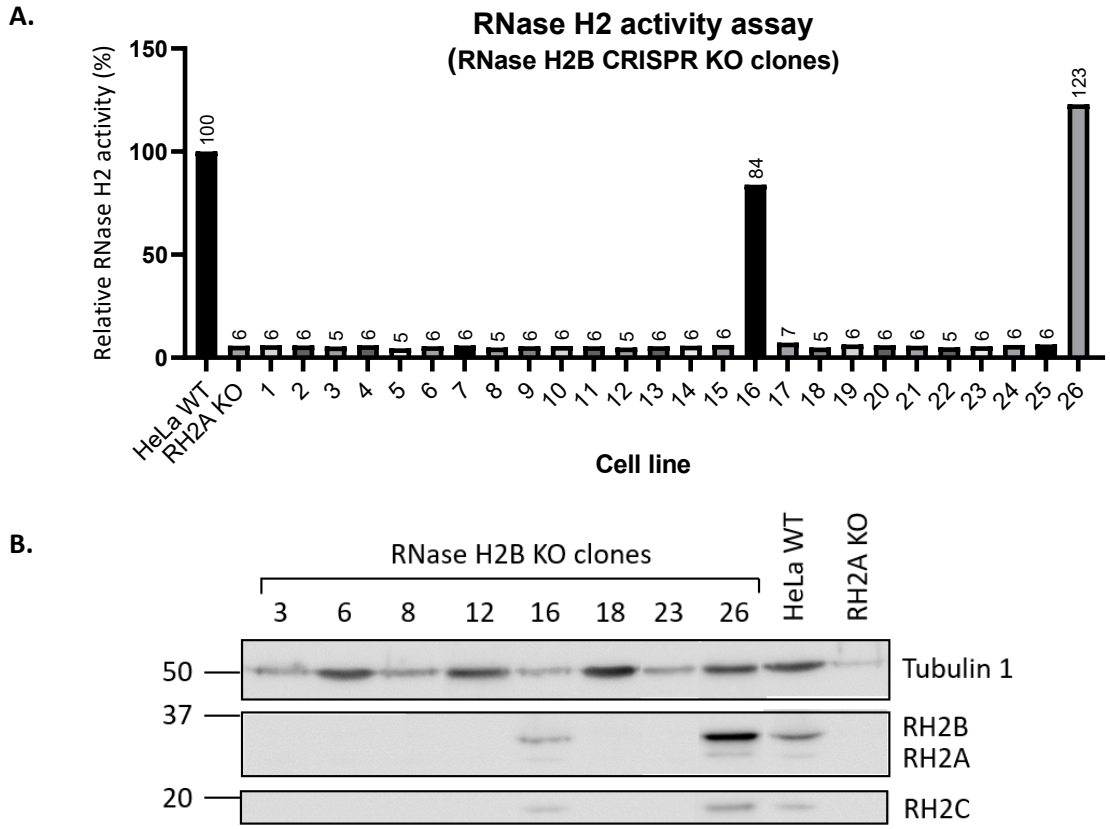


Figure 3.4. RNase H2 activity assay and western blot for RNASEH2B KO clones – (A) Results from the RNase H2 activity assay conducted on RNASEH2B CRISPR KO clones relative to that of the parental (HeLa WT) cell line (84 clonal cell lines analysed but only 26 shown in figure, including the eight selected clonal cell lines). An internal positive control lacking RNase H2 activity (an RNASEH2A KO cell line) is included (RH2A KO). Values above bars indicate relative RNase H2 activity (%). (B) Protein levels of the three RNASEH2 subunits A (33 kDa), B (35 kDa) and C (18 kDa), as detected using an RNASEH2 specific antibody. Tubulin 1 (50 kDa) was used as a loading control. RNASEH2A KO cells and the HeLa parental cell line were used as internal controls.

3.4.1.2. Complementation of RNASEH2B KO cell lines

To control for potential off-target effects of the CRISPR KO process, and to generate a cell line expressing an RNASEH2B subunit containing a mutation in its PIP motif (therefore abolishing its interaction with PCNA), I used a RNASEH2B KO clonal cell line for complementation. Clone 6 was selected, and was complemented with cDNA coding for RNASEH2B-WT or RNASEH2B-PIPm (F300A/F301A), abolishing interaction with PCNA (Bubeck et al., 2011), using retroviral transduction (2.3.1.2). The experimental workflow is summarised in ***Figure 3.5***. In brief, retroviral particles conferring resistance to zeocin, and containing either *RNASEH2B*-WT, *RNASEH2B*-PIPm or an EV (***Table 2.3***: pMAR761, pMAR762, pJKN255), were generated and used for transduction of the clonal HeLa RNASEH2B KO cell line 6. Transduced cells underwent zeocin selection to ensure all surviving cells contained at least one vector insertion. The resulting zeocin-resistant foci (i.e., complemented cell lines) were expanded and characterised using western blot probing for RNASEH2 subunits (2.3.1.4.2; 2.3.1.4.4). The blots were analysed using Fiji-ImageJ to determine the achieved level of complementation compared to WT HeLa cells.

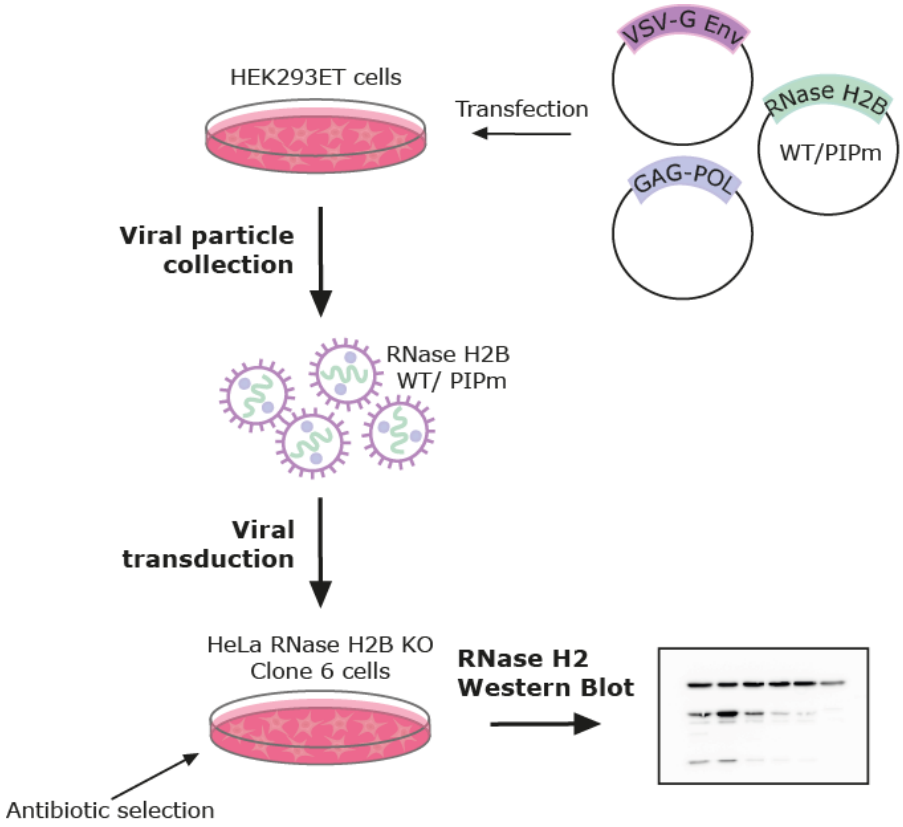


Figure 3.5. **Generation and characterisation of complemented clone 6 HeLa RNASEH2B KO cell lines** – Workflow used to generate clone 6 RNASEH2B-WT and -PIPm complemented cell lines.

This process was initially repeated twice for *RNASEH2B*-WT (+WT1 and +WT2) and -PIPm (+PIP1 and +PIP2). Western blot analyses revealed that the first viral transduction was more efficient than the second. Quantification using Fiji-ImageJ showed complementation levels of 232% and 22% for +WT1 and +PIP1 respectively, compared to 82% and 4% for +WT2 and +PIP2 respectively (***Figure 3.6***). Due to the low efficiency of RNASEH2B-PIPm complementation, two more independent rounds of viral particle production and subsequent complementation using this vector were performed, each resulting in a similar or lower level of complementation than that observed in the first two rounds (data not shown). While further research is needed, it is possible that mutation of the PIP motif destabilises RNASEH2B and cause the low complementation levels detected. I therefore proceeded with the generated cell lines and selected the HeLa RNASEH2B KO clone 6 +WT2, +PIP1 and +EV cell lines for further experimentation, since their RNASEH2B complementation levels were closer to that of the WT HeLa cell line (HeLa WT).

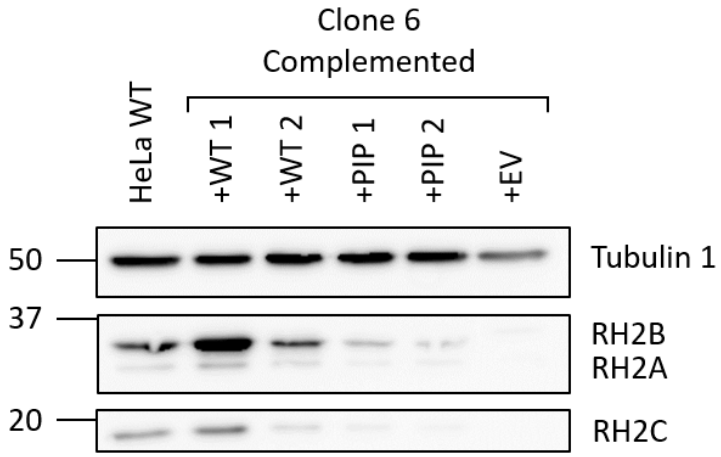


Figure 3.6. RNASEH2B western blot on complemented RNASEH2B KO Clone 6 cells – Protein levels of the three RNASEH2 subunits A (33 kDa), B (35 kDa) and C (18 kDa), were detected using an RNASEH2B specific antibody. Tubulin 1 (50 kDa) was used as a loading control. 1 and 2 refer to two independent replicates of the complementation of clone 6 with RNASEH2B WT and PIPm.

3.4.2. The retrotransposition activity of WT LINE-1s in HeLa RNASEH2B KO cells

The characterised HeLa RNASEH2B CRISPR KO cell lines were used to explore changes in LINE-1 retrotransposition, using engineered LINE-1 mobilisation assays. The goal of these experiments was to confirm whether RNASEH2B KO in HeLa cells reproduces the same LINE-1

deficit observed for HeLa RNASEH2A KO cell lines (Benitez-Guijarro et al., 2018). To allow direct comparison with our previous study, I included in these experiments three HeLa RNASEH2A KO (RH2A KO) and two RNASEH2A control (Control RH2A) clonal cell lines, all generated by Benitez-Guijarro and co-workers in 2018. To note, to rule out potential off-target effects, the control clonal cell lines used for these experiments were established from cells that had gone through the RNASEH2A and B CRISPR KO process respectively, but possessed near WT RNase H2 activity levels as measured by the RNase H2 activity assays (this thesis and Benitez-Guijarro et al., 2018) (3.1.1.1). These cell lines were used to perform an antibiotic-selection colony-formation retrotransposition assay (2.3.3.3), relying on the activation of a blasticidin resistance gene only after a round of retrotransposition has taken place. The rationale of the L1 retrotransposition assay and the experimental workflow is summarised in ***Figure 3.7***. Briefly, the vectors containing a human LINE-1 tagged with a blasticidin resistance based retrotransposition indicator cassette were transfected into the different HeLa cell lines (2.3.3.1), and cells were subjected to antibiotic selection using blasticidin. In parallel, the transfection efficiency of the different vectors in the different cell lines was measured (2.3.3.2). As an additional control, the different cell lines were transfected with a vector conferring constitutive resistance to blasticidin, allowing to control for changes in clonability/toxicity. The efficiency of LINE-1 retrotransposition is calculated based on the number of surviving colonies for each vector, normalised to the transfection efficiency and the number of colonies in the clonability/toxicity control, for each cell line respectively. All experiments were done using three technical replicates for each vector, and were repeated in three independent experiments (i.e., biological replicates, at different cell passages). The data shown in figures is the average of the three biological replicates.

The retrotransposition colony assay was performed using vectors containing different allelic LINE-1 variants, WT and RTm (D702A) (Moran et al., 1996), all tagged with an *mblastI* cassette (***Table 2.3***: JJ101, JJ116). pcDNA6.1 was used in parallel in toxicity/clonability control assays (***Table 2.3***: pcDNA6.1). The L1-RTm was used to control for any acquired blasticidin resistance, not mediated by L1 retrotransposition, and was expected to not generate any viable colonies (i.e., background levels). The pCDNA6.1, which constitutively express the blasticidin deaminase gene, was used to control for cell line differences in growth and sensitivity to blasticidin, and normalise the readout of LINE-1 retrotransposition (2.3.3.3).

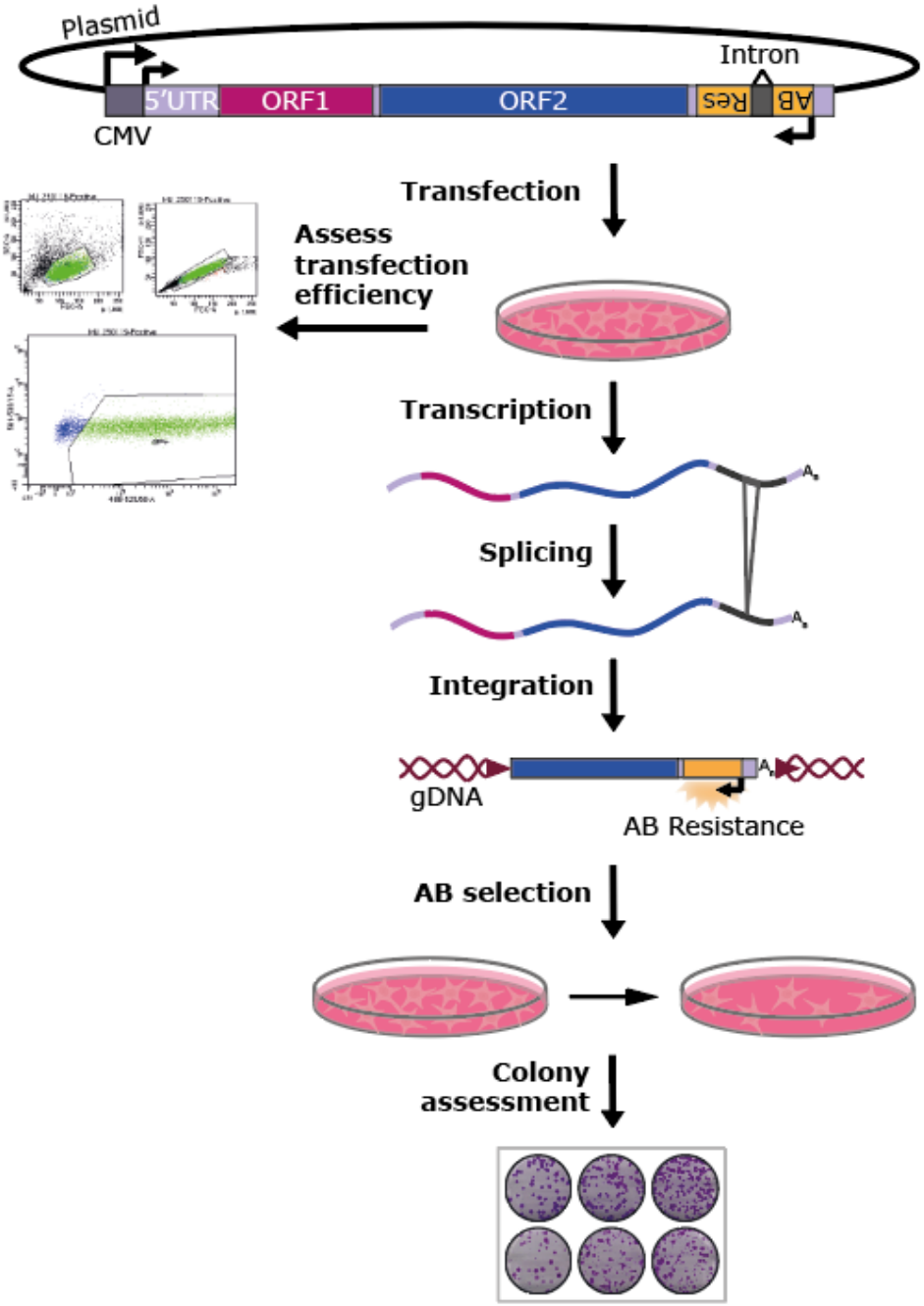


Figure 3.7. In vitro retrotransposition assay using antibiotic selection –Workflow used to perform the colony retrotransposition assay, including a schematic of a generic LINE-1 reporter construct under control of an external CMV promoter (light grey), tagged with an antisense retrotransposition indicator cassette containing its own promoter and polyadenylation signal (yellow). Promoters depicted as arrows.

Comparison by a Tukey's multiple comparison test showed there was no significant difference between the levels of L1-WT retrotransposition in the parental HeLa cell line (HeLa WT) and the two control cell lines of the RNASEH2A and B KO (Control RH2A and Control RH2B), although retrotransposition levels in the RH2B control cell lines were close to 50% of that found

in the parental HeLa cell line (***Figure 3.8***). The lack of statistical significance of the 50% reduction in retrotransposition levels in the control RH2B cells compared to the WT HeLa cell line could be due to limited number of samples. More repeats or another statistical test such as one-sample t-test might provide more power to the analysis. The reduction in retrotransposition levels of the control RH2B cells might be due to the lack of true WT genotype in these cells, as indicated by capillary DNA sequencing (3.4.1.1; ***Figure 6.9***). As observed by Benitez-Guijarro and co-workers in 2018, there was a significant reduction in retrotransposition levels in the RNASEH2A KO clonal cell lines (RH2A) compared to their control cell lines (Control RH2A) ($p = 0.008$). This was not observed for the RNASEH2B KO clonal cell lines (RH2B) compared to their control cell lines (Control RH2B) ($p = 0.109$). There was however, a significant reduction in LINE-1 retrotransposition in the RNase 2B KO cell lines (RH2B KO) when compared to the parental cell line (HeLa WT) ($p = 0.011$). In fact, L1-WT retrotransposition levels were very similar in the RNASEH2A and B KO clonal lines (***Figure 3.8***).

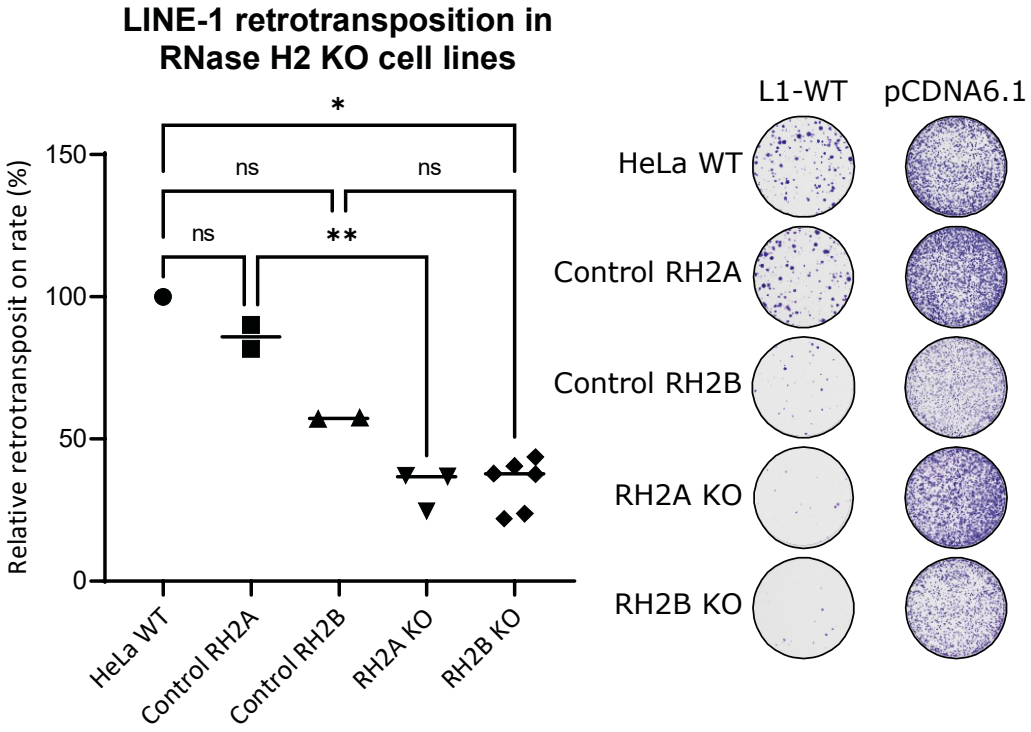


Figure 3.8. RNase H2 KO decreases LINE-1 retrotransposition – The retrotransposition rate of the standard L1-WT element, as measured using the colony assay in, the control clonal cell lines for the RNASEH2A (Control RH2A) (n = 2) and RNASEH2B (Control RH2B) (n = 2) KO cell lines, as well as the RNASEH2A (RH2A KO) (n = 3) and RNASEH2B (RH2B KO) (n = 6) KO cell lines relative to that of the parental HeLa cell line (HeLa WT) (n = 1) is shown. Each point represents the average of three biological replicates. Data was compared using a Tukey's multiple comparison test (ns = not significant; * = p-value < 0.05; ** = p-value < 0.01). Pictures on the right show representative retrotransposition assays of the L1-WT and the control pCDN6.1 plasmid in the different cell lines.

3.4.3. LINE-1 retrotransposition in HeLa RNASEH2B-PIPm cells

To control for potential undesired changes induced by the CRISPR KO process, which can not be measured using the RNase H2 activity assay against misincorporated single ribonucleotides or western blot, the following experiments were performed using the complemented cell lines (3.1.1.2). To determine the role that the PIP motif of RNASEH2B might have in the LINE-1 retrotransposition cycle, I performed a retrotransposition assay using the selected complemented cell lines, which were derived from HeLa RNASEH2B KO clone 6 complemented with i) RNASEH2B-WT (Clone 6 +WT), ii) RNASEH2B-PIPm (Clone 6 +PIPm) and ii) EV (Clone 6 +EV) (3.1.1.2; ***Figure 3.7***; ***Figure 3.9***). I used the same experimental set up and data processing steps described above (3.1.2.1).

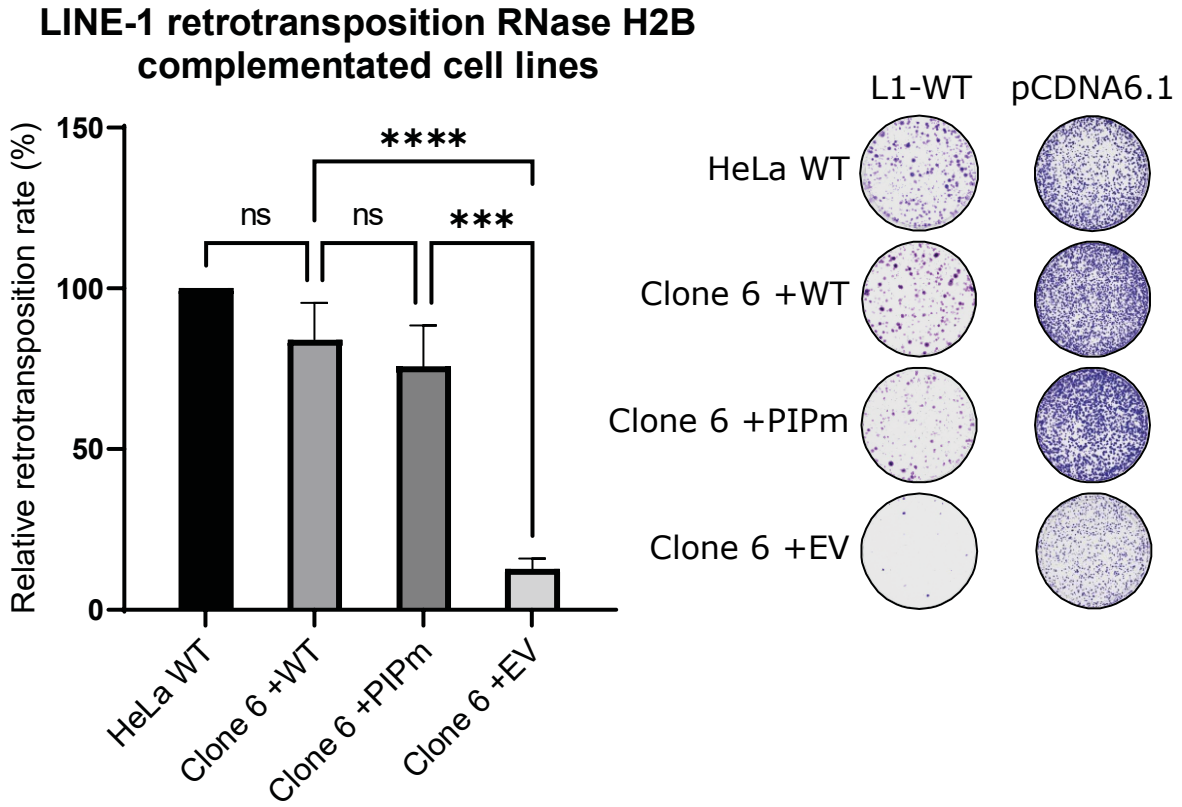


Figure 3.9. **RNASEH2B complementation rescues LINE-1 retrotransposition** – The retrotransposition rate of the standardised L1-WT relative to its retrotransposition rate in WT HeLa cells, as measured using the colony assay, in a WT HeLa cell line (HeLa WT), the RNASEH2B KO clone 6 complemented with RNASEH2B-WT (Clone6 +WT) and EV (Clone6 +EV). The data represents the average of three biological replicates. Data was compared using a Tukey's multiple comparison test (ns = not significant; *** = p-value < 0.001; **** = p-value < 0.0001). Error bars represent the standard deviation of independent triplicates. Pictures on the right show representative retrotransposition assays of the L1-WT and the control pCDN6.1 plasmid in the different cell lines.

Using a Tukey's multiple comparison test, I first observed a significant increase of retrotransposition activity of the L1-WT in Clone 6 complemented with WT RNASEH2B compared to EV (***Figure 3.9***; $p < 0.0001$). Notably, the L1-WT retrotransposition in Clone 6 complemented with RNASEH2B-WT was not significantly different from the levels detected in the WT HeLa cell line (***Figure 3.9***; $p = 0.19$), suggesting a rescue of the LINE-1 retrotransposition deficit back to WT levels after RNASEH2B complementation. Furthermore, the retrotransposition levels measured in the Clone 6 complemented with RNASEH2B-PIPm were not significantly different from those measured in the RNASEH2B-WT complemented cell line (***Figure 3.9***; $p = 0.66$). This result suggests that the presence/absence of a functional PIP motif has little to no effect on the ability of RNase H2 to facilitate retrotransposition (note that complementation levels of RNASEH2B-WT and -PIPm were 82% *vs* 22% of WT levels respectively; 3.2.1.2; ***Figure 3.6***).

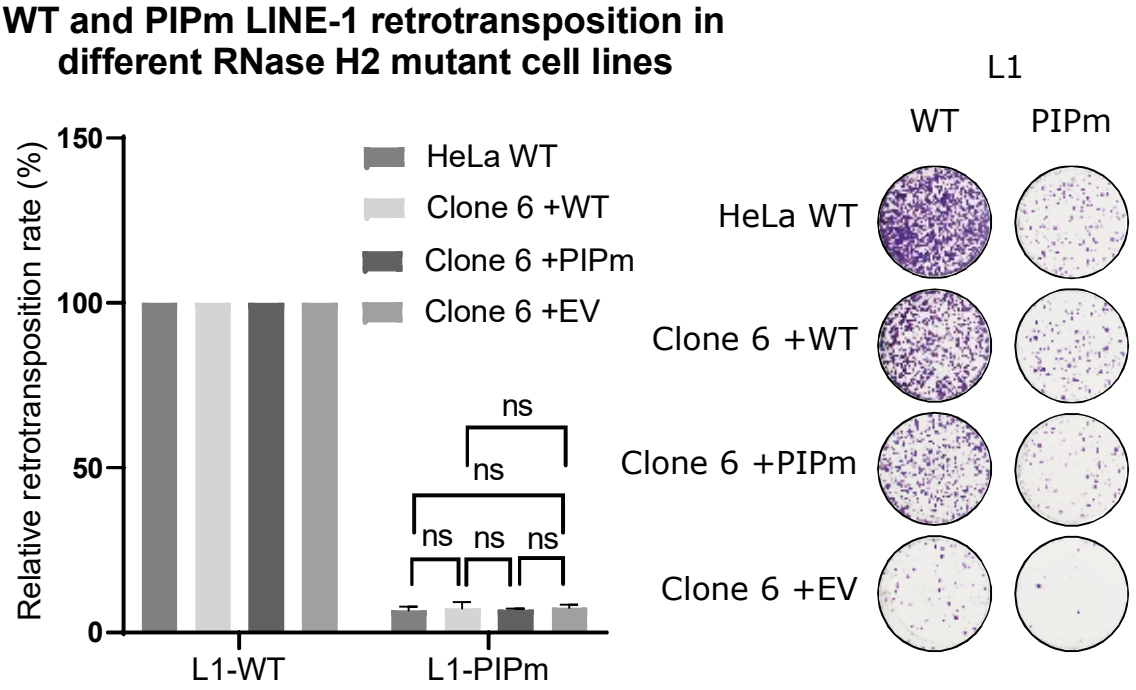


Figure 3.10. RNase H2 status impacts WT and PIPm LINE-1 retrotransposition equally – The retrotransposition rate of the standardised L1-WT and L1-PIPm relative to the retrotransposition rate of the L1-WT in each respective cell line, as measured using the colony assay in a WT HeLa cell line (HeLa WT), the RNASEH2B KO clone 6 complemented with RNASEH2B-WT (Clone6 +WT), RNASEH2B-PIPm (Clone 6 +PIPm) and EV (Clone6 +EV) is shown. Data represents the average of three biological replicates. The relative retrotransposition of the L1-PIPm in each cell line was compared using a Tukey's multiple comparison test (ns = not significant). Error bars represent the standard deviation of biological replicates. Pictures on the right show representative retrotransposition data of the L1-WT and -PIPm in the different cell lines tested (Note that, due to its reduced retrotransposition potential, four times more cells were seeded for the L1-PIPm retrotransposition assays 2.3.3.1; ***Table 2.8***).

To corroborate this data and gain mechanistic insights in the role of the PIP motif of RNase H2 and LINE-1-ORF2p in retrotransposition, I conducted retrotransposition colony assays in these cell lines, including the LINE-1 mutant variant containing a mutation in its PIP motif (Y414A/Y415A) abolishing ORF2p interaction with PCNA (Taylor et al., 2013) (***Figure 3.7***; ***Table 2.3***: JJ101, JJ116, JJ107). To compare the effect of cellular RNase H2 status on the retrotransposition activity of the L1-WT and L1-PIPm, the measured retrotransposition activity of the L1-PIPm was normalised to that of the L1-WT in each respective cell line (giving rise to the relative retrotransposition rate (%) depicted in ***Figure 3.10***). Consistently with previous findings, comparing the relative retrotransposition rate of the L1-PIPm in the different cell lines showed there was no significant difference, using a Tukey's multiple comparison test (p-value > 0.62).

3.5. Low RNase H2 activity levels are sufficient to support efficient LINE-1 retrotransposition in cultured cells

3.5.1. Generation and characterisation of HeLa cell lines with different levels of RNase H2 activity

The finding that complementation with <25% WT RNase H2 levels can restore LINE-1 retrotransposition rates back to WT levels, as observed for the HeLa RNASEH2B KO clone 6 +PIPm cell line, drew our attention (3.2.2.2). Thus, I aimed to further investigate the level of RNase H2 activity necessary to support efficient LINE-1 retrotransposition in HeLa cells by generating several subclonal HeLa cell lines with different levels of RNase H2 activity, using a workflow similar to that depicted in ***Figure 3.3***. First, the characterised HeLa RNASEH2B KO clone 6 complemented with RNASEH2B-WT and EV cell lines were single cell sorted and these subclonal cell lines were generated and expanded (3.2.1; 2.3.1.3). Then, RNase H2 activity level of the resulting subclonal cell lines was assessed by RNase H2 activity assays (2.3.1.4.3). As expected, no significant RNase H2 activity was found in any of the subclonal lines complemented with EV. In contrast, RNase H2 activity of RNASEH2B-WT complemented subclonal cell lines ranged from 11% to 192% that of WT levels (data not shown). Three subclonal EV cell lines showing background RNase H2 activity levels (Subclone 6 +EV1, Subclone

6 +EV2 and Subclone 6 +EV3), and six cell lines with RNase H2 activity levels ranging from 26% to 192% of WT levels (Subclone 6 +26%, Subclone 6 +52%, Subclone 6 +94%, Subclone 6 +109%, Subclone 6 +148% and Subclone 6 +192%) were selected for further experimentation. These cell lines were then analysed over time to establish potential variations in RNase H2 activity (***Figure 3.11***) and protein levels (***Figure 3.12***) between three different time points. For this purpose, cells were collected from three consecutive cell passages, corresponding with the start of the biological replicates of the related colony retrotransposition assays discussed below (3.3.2). RNase H2 activity assays (***Figure 3.11***) and western blots (***Figure 3.12***) indicated a rough correlation between the measured average activity and protein levels with the expected values, based on my initial measurements.

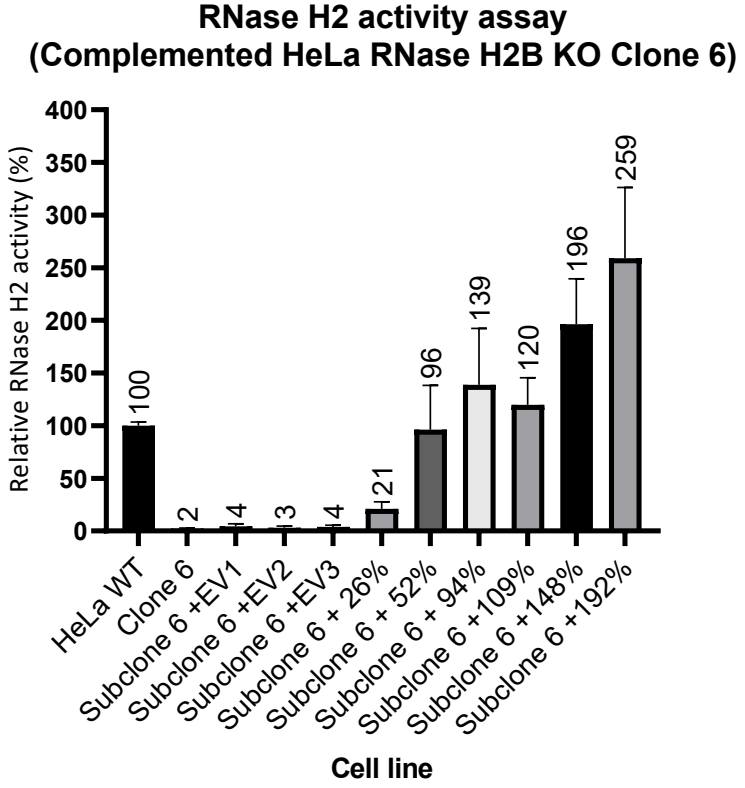


Figure 3.11. RNase H2 activity assay of Clone 6 HeLa RNASEH2B KO complemented subclonal cell lines – The average RNase H2 activity of three different time points, of RNASEH2B CRISPR KO clone 6 (Clone 6), three subclonal cell lines generated from clone 6 complemented with EV (Subclone 6 +EV1, Subclone 6 +EV2 and Subclone 6 +EV3) and six subclonal cell lines generated from clone 6 complemented with WT RNASEH2B, previously identified as having different levels of RNase H2 activity (Subclone 6 +26%, Subclone 6 +52%, Subclone 6 +94%, Subclone 6 +109%, Subclone 6 +148% and Subclone 6 +192%), relative to that of a WT HeLa cell line (HeLa WT) is shown. Error bars represent the standard deviation of the three different time points (i.e., the three passages). Values above error bars indicates the average relative RNase H2 activity (%); values in the X-axis (%) refer to the original level detected when these subclonal cell lines were established.

As a notable exception, Subclone 6 +52% has almost double (~100%) the average activity level of its initial measurement over the three time points (***Figure 3.11***), which is also reflected in measured protein levels (~150%) (***Figure 3.12B***). Furthermore, Clone 6 +94% has much higher protein levels (~300%) than expected based on the measured RNase H2 activity level (~130%). Since both measurements were performed on the same cell lysates, this might indicate that protein and activity levels don't necessarily always correlate, although in general this does seem to be the case.

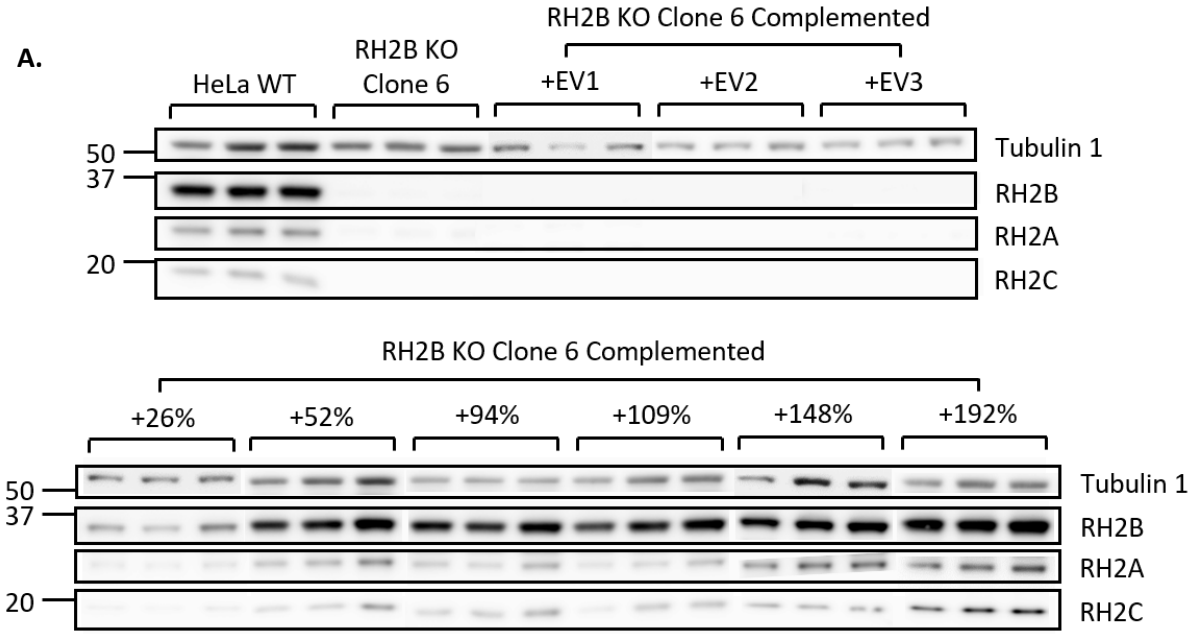


Figure 3.12. RNase H2 protein levels of Clone 6 HeLa RNASEH2B KO complemented subclonal cell lines – (A) Western blot of the three RNASEH2 subunits A (33 kDa), B (35 kDa) and C (18 kDa), as detected by an RNase H2 specific antibody, as well as the loading control Tubulin 1 (50 kDa) at three different time points, of a WT HeLa cell line (HeLa WT), RNASEH2B KO clone 6 (Clone 6), three cell lines generated from a subclonal expansion of clone 6 complemented with EV (Subclone 6 +EV1, Subclone 6 +EV2 and Subclone 6 +EV3) and six cell lines generated from a subclonal expansion of clone 6 complemented with WT RNASEH2B, previously identified as having different levels of RNase H2 activity (Subclone 6 +26%, Subclone 6 +52%, Subclone 6 +94%, Subclone 6 +109%, Subclone 6 +148% and Subclone 6 +192%). Tubulin 1 (50 kDa) was used as a loading control. (B) [in the next page] Western blot quantification. Bar graph represents the average RNASEH2B protein levels of triplicates obtained at different time points, relative to that of the WT HeLa cell line. Error bars represent the standard deviation of three different time points. Values above error bars indicates the average relative RNASEH2B protein level (%), while values in the X-axis (%) refer to the original activity level detected when these subclonal lines were established.

B.

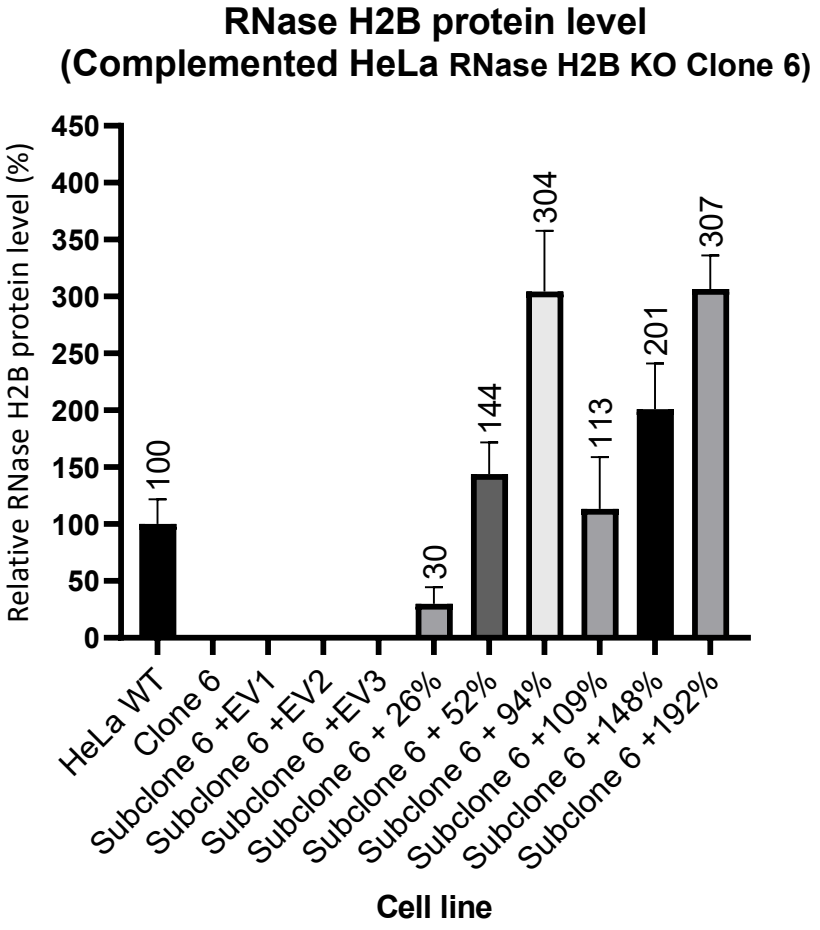


Figure 3.12. RNase H2 protein levels of Clone 6 HeLa RNASEH2B KO complemented subclonal cell lines – Legend in the previous page.

3.5.2. LINE-1 retrotransposition efficiency does not correlate with RNase H2 activity levels in cultured HeLa cells

To determine whether RNASEH2B expression levels influence L1 integration, I used a colony retrotransposition assay to assess LINE-1 retrotransposition rate in the previously generated subclonal cell lines with different levels of RNase H2 activity (3.3.1), following the experimental set up and data processing described before (3.1.2.1; ***Figure 3.7***; ***Figure 3.13***). The relative retrotransposition rate of the standardised L1-WT in the different cell lines, of three independent experiments, was plotted against the relative RNase H2 activity level measured in the sample collected at the start of that retrotransposition assay (3.3.1; ***Figure 3.12***, ***Figure 3.13***). The values were analysed using simple linear regression, finding that although there is a relationship between RNase H2 activity levels and retrotransposition rate, it does not explain

the different retrotransposition rates between the clones (r=0.313, ***Figure 3.13***). In fact, the clonal cell line possessing 21% relative RNase H2 activity levels showed >100% LINE-1 retrotransposition rates. The relative LINE-1 retrotransposition rate detected in the complemented cell lines ranged between 50% and 160% that observed in the WT HeLa cell line. On the other hand, the clonal cell lines derived from the RNASEH2B KO HeLa cells complemented with EV, possessing between 2% and 4% relative RNase H2 activity levels, supported a maximum of 50% the retrotransposition rate measured in WT HeLa cell lines, with an average value of 30%.

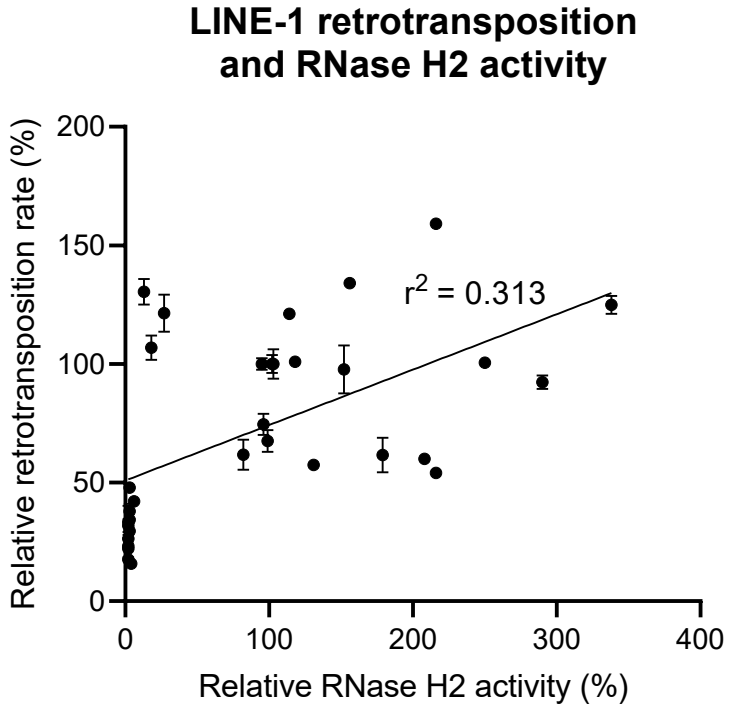


Figure 3.13. LINE-1 retrotransposition and RNase H2 activity – The retrotransposition rate of the standard L1-WT in the RNASEH2B KO clone 6 (Clone 6), three cell lines generated from a subclonal expansion of clone 6 complemented with EV (Subclone 6 +EV1, Subclone 6 +EV2 and Subclone 6 +EV3) and six cell lines generated from a subclonal expansion of clone 6 complemented with WT RNASEH2B, previously identified as having different levels of RNase H2 activity (Subclone 6 +26%, Subclone 6 +52%, Subclone 6 +94%, Subclone 6 +109%, Subclone 6 +148% and Subclone 6 +192%), relative to its retrotransposition rate in a WT HeLa cell line (HeLa WT), plotted against their corresponding relative RNase H2 activity level measured is shown. The data represents individual values of three biological replicates (11 cell lines x 3 replicates = 33 data points). Data was analysed using simple linear regression (r^2=0.313). Error bars represent the standard deviation between technical replicates.

3.6. Dependence of the human LINE-1 RNase H chimeric construct on cellular RNase H2

3.6.1. Introduction of a functional RNase H domain in human LINE-1s

To analyse the involvement of RNase H activity in TPRT using a different approach, we explored whether "*in vitro* evolution" could be used to further clarify the need for such enzymatic activity during LINE-1 integration. Indeed, evolutionary analyses suggest that RNase H domains were acquired recently during the evolution of the LINE clade. Thus, mimicking recent evolution, we explored whether the addition of a functional RNase H domain in human L1-ORF2p, emulating the configuration described for other LINE-1-like elements (3.1), would alleviate the need for cellular RNase H activity during retrotransposition. To do that, we considered two independent functional RNase H domains: the *RNASEHI* gene from *E. coli* (constructs L1-eRH); or the catalytic domain from the human *RNASEH1* (constructs L1-hRH). Each domain was cloned in-frame with, and fused to, the C-terminus of L1-ORF2 (2.2.1.7; ***Table 2.4***: pMK001 – pMK012; ***Figure 3.14***), adding a flexible protein linker between L1-ORF2p and each RNase H domain, allowing a theoretical free movement of the RNase H domain (linker made up out of 5 tandem copies of a glycine (GGC)-serine (TCG) motif). The codons used to add the glycine/serine linker were higher in GC content, for better polymerase processivity, and avoided creating unwanted splice sites. As internal negative controls, we also generated allelic L1 constructs containing catalytic dead versions of each RNase H domain *E. coli RNASEHI* (D70N; L1-eRHm) and human *RNASEH1* (D119N; L1-hRHm)). These chimeric L1 constructs were constructed using three allelic variants of the human LINE-1, WT, RTm and PIPm, and all were tagged with the *mneoI* retrotransposition cassette.

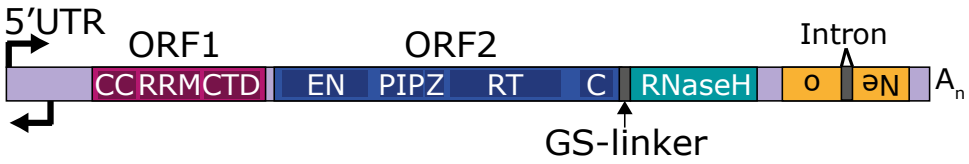


Figure 3.14. **LINE-1 ORF2-RNase H fusion constructs** – A schematic representation of the generated LINE-1 ORF2-RNase H chimeric constructs, tagged with an antibiotic resistance, retrotransposition indicator cassette. All variants of this chimeric construct follow the same structure depicted here.

3.6.2. Retrotransposition of LINE-1 ORF2-RNase H chimeric constructs in RNase H2 KO cells

These generated LINE-1 ORF2-RNase H fusion constructs were then tested in a colony retrotransposition assay to first assess their retrotransposition activity compared to the WT L1 and the L1-PIPm, in a WT HeLa cell line, following the same experimental set up and data processing described before (3.1.2.1; ***Figure 3.7***; ***Figure 3.15***). Although the chimeric LINE-1 RNase H constructs were able to retrotranspose in cultured HeLa cells, their activity levels were significantly lower than that of the parental L1 element as assessed using a Tukey's multiple comparison test (p-value < 0.01), with the exception of the L1-WT - eRHm (***Figure 3.15A***). The chimeric constructs containing the *E. coli RNASEHI* domain retrotransposed at a higher level than the constructs containing the human *RNASEH1* domain. Notably, when compared to their catalytic dead counterpart, the L1-WT - hRH construct shows a significant reduction in its activity as assesed using a Tukey's multiple comparison test (p-value < 0.001). This was not observed for the L1-WT - eRH constructs. Comparing the retrotransposition activity of the L1-PIPm ORF2-RNase H chimeric constructs to the L1-PIPm, I observed a similar pattern, although the low colony numbers associated to the low retrotransposition activity of the L1-PIPm resulted in larger error bars and less significance (***Figure 3.15B***). As expected, none of the RT-mutant allelic constructs tested retrotransposed efficiently in HeLa cells (data not shown).

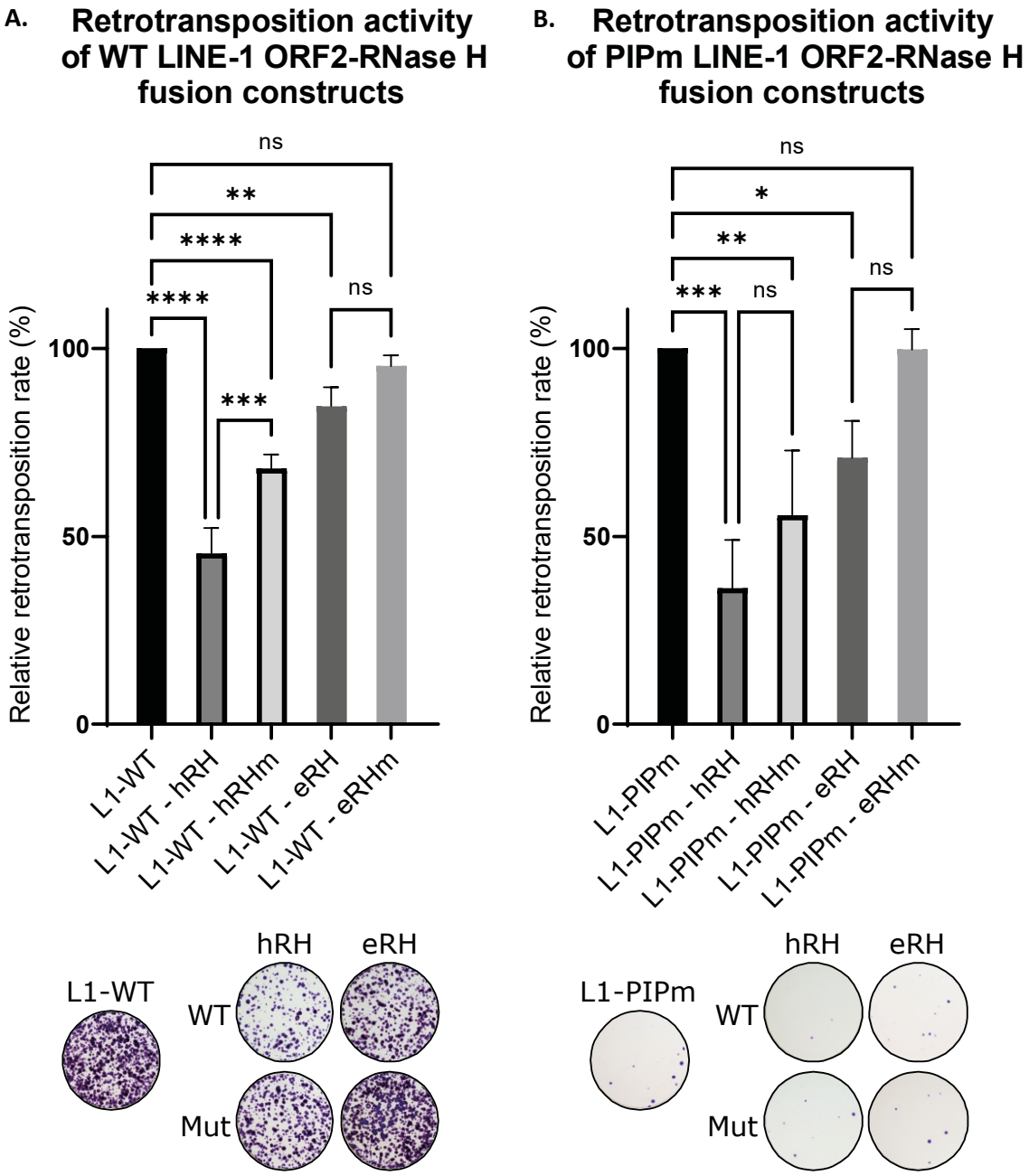


Figure 3.15. Retrotransposition activity of chimeric LINE-1 ORF2-RNase H constructs – (A) The retrotranposition activity of the WT L1 element containing the catalytic domain of the human *RNASEH1* (L1-WT - hRH) and a catalytic dead mutant (L1-WT - hRHm) respectively, and the *E. coli RNASEHI* (L1-WT - eRH) and a catalytic dead mutant (L1-WT - eRHm) respectively, relative to the WT L1, in a WT HeLa cell line. (B) As in A, but for the L1-PIPm ORF2-RNase H chimeric constructs. Data was compared using a Tukey's multiple comparison test, and its significance indicated in the graph (ns = not significant; * = p-value < 0.05; ** = p-value < 0.01; *** = p-value < 0.001; **** = p-value < 0.0001). Error bars represent the standard deviation from two biological replicates. Pictures at the bottom show representative retrotransposition data of the L1-WT and L1-PIPm.

To complement this result, I tested the different L1-WT ORF2-RNase H chimeric constructs in a subset of the previously generated RNase H2 KO cell lines (3.2.1.1). There was no significant difference in relative retrotransposition activity of the WT and catalytic dead mutants when comparing any of the cell lines (***Figure 3.16***). In line with previous results, comparing the activity of the catalytic dead RH mutants to their WT counterparts in each cell line indicated the catalytic dead mutant had a higher retrotransposition activity in all cell lines.

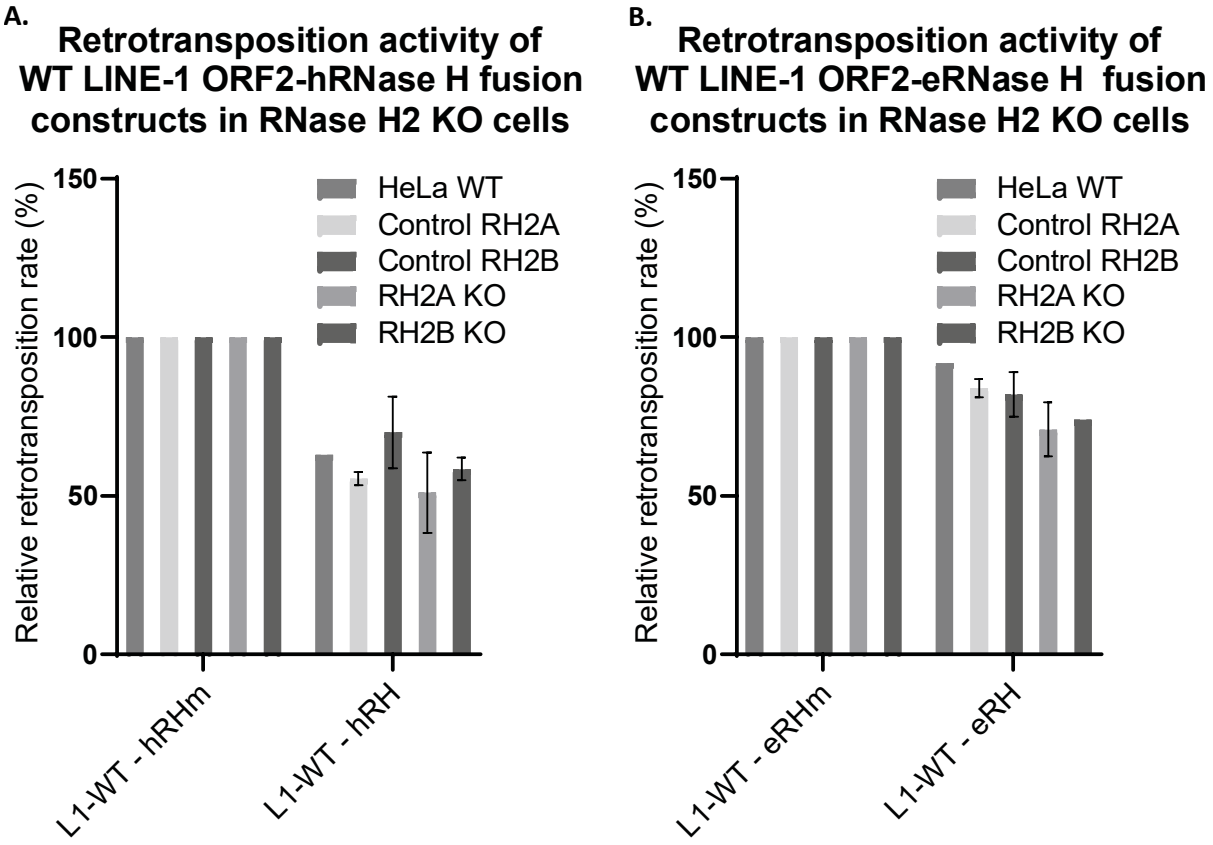


Figure 3.16. Retrotransposition activity of chimeric LINE-1 ORF2-RNase H constructs in RNase H2 KO cells – (A) The retrotranposition activity of the WT L1 element containing the catalytic dead mutant of the human *RNASEH1* (L1-WT - hRHm), relative to its WT variant (L1-WT - hRH), in a WT HeLa cell line (n = 1), control RH2A (n = 2) and RH2B (n = 2) cell lines, and RH2A KO (n = 2) and RH2B KO (n = 2) cell lines. (B) As in A but using the WT L1 element containing the catalytic dead mutant of the *E.coli RNASEHI* (L1-WT - eRHm) and WT variant (L1-WT - eRH). N indicates number of cell lines; graphs indicate average of two biological replicates; error bars indicate standard deviation between the different cell lines. Data from each construct in the different cell lines was compared using a Tukey's multiple comparison test.

Discussion

LINE-1 activity, expression and/or retrotransposition has been observed in association with various diseases and disorders (1.3). Although in most cases LINE-1's contribution to the

cause and symptomology of these pathologies is unclear, their mutagenic potential and ability to induce the activation of the innate immune system can lead to a worsening of the prognosis. Understanding which cellular factors are involved in the LINE-1 life cycle, what role they play and how they fulfil these roles, lies at the base of learning how to control aberrant LINE-1 activity, and predict when they may become a threat to human health. Building from previous work performed by our research team (Benitez-Guijarro et al., 2018), I set out to explore how cellular RNase H2, a protein frequently mutated in AGS and recently described to facilitate retrotransposition (1.3.3), fulfils its function in the LINE-1 retrotransposition cycle. In this chapter, I will discuss the outcomes of a series of experiments I performed to address the aims and objectives set out.

3.7. RNase H2 promotes LINE-1 retrotransposition in a PCNA-independent manner

To shed more light on the involvement of the PIP motifs of RNase H2 and L1-ORF2p in the retrotransposition cycle, I set out to generate a cellular model with a mutation in the PIP motif of its RNase H2 (located at the RNASEH2B subunit), abolishing its interaction with PCNA. The effect of the utilised mutation on the ability of RNase H2 to interact with PCNA was previously characterised (Bubeck et al., 2011). At the lab, we previously observed that RNASEH2A KO significantly affects the ability of cells to support LINE-1 retrotransposition using an exogenous WT LINE-1 vector. Since deleting any subunit of RNase H2 leads to the absence of the entire heterotrimer (Chon et al., 2009), I generated RNASEH2B CRISPR KO HeLa cells using the exact HeLa cell line previously used to create RNASEH2A KO cells (RH2A KO) (3.2.1.1). The only two clones with WT RNASEH2B protein and RNase H2 activity levels, obtained from the genome editing process (3.2.2), were used as controls to eliminate potential off-target effects of CRISPR-Cas9 (Control RH2B). However, unlike for RH2A KO, we found the RH2B KO cell lines support LINE-1 retrotransposition only slightly less than the Control RH2B cell lines, yet not significantly (***Figure 3.8***). However, in the same assay, the RH2A KO and Control cell lines did show the expected phenotypes, as the Control RH2A cell lines clustered with the WT HeLa parental cell line, and the RH2A KO cell line supported significantly less retrotransposition compared to its control. The lack of significance between RH2B KOs and RH2B Controls, was likely due to their ambiguous retrotransposition levels, measuring between the cell lines

supporting and those not supporting retrotransposition. This suggests that, although they showed near WT protein levels and RNase H2 activity levels in an *in vitro* assay, this may not be the case in an *in vivo* setting. Notably, the RNase H2 activity assay I used here only assess the proteins ability to excise ribonucleotides, while its enzymatic function proposed to facilitate retrotransposition is the removal of the RNA from RNA:DNA hybrids. Additionally, it is worth mentioning that neither of the cell lines were found to possess a true WT genotype, as indicated by close inspection of their capillary DNA sequencing performed during the initial characterisation (3.2.1.1; ***Figure 6.9***). The control RH2B cell line derived from clone 16 shows clear alterations of the targeted exon of the *RNASEH2B* gene (***Figure 6.9***). However, the control cell line derived from clone 23 contains an alteration in the intron following the targeted exon (***Figure 6.9***). We are unsure of the origin of this mutation and whether it is related to the editing process. However, it may influence RNA splicing causing alterations in protein function, but futher experimentation is necessary to address this question. These control cell lines can act as potential useful tools in deciphering what aspect of RNASEH2B is essential for retrotransposition. More detailed analyses of the idependent *RNASEH2B* alleles of these cell lines will be necessary to look into this possibility. I speculate that unassessed aspects of the RNase H2 protein function may have been affected by the genome editing process, in a way that still allows heterotrimer assembly and stability, but disqualify them as controls. Due to the efficiency of our genome editing approach (different to the one used previously to create the RH2A KO clonal cell lines), there were no alternative clones to consider as internal negative controls. Nonetheless, the pattern shown by all the other cell lines regarding their ability to support retrotransposition, together with the specific increase in retrotransposition in RH2B KO complemented with WT RNASEH2B (***Figure 3.8***; ***Figure 3.9***), confirms that RNASEH2B depletion results in the same LINE-1 related phenotype observed for RNASEH2A KO cells. This result is consistent with the hypothesis that the ability of RNase H2 to degrade the RNA from RNA:DNA hybrids facilitates LINE-1 retrotransposition (Benitez-Guijarro et al., 2018), and that this protein can not be formed in the absence of any of its 3 subunits (Chon et al., 2009).

The generated RNASEH2B KO clonal cell lines allow challenging the specific role of the RNASEH2B interaction with PCNA in LINE-1 retrotransposition by complementing these cell lines with RNASEH2B containing a mutation in its PIP motif (RNASEH2B-PIPm; abolishing its interaction with PCNA) (3.2.1.2). In this set of experiments, in addition to a WT LINE-1 (L1-WT),

I also assessed the retrotransposition rate of a LINE-1 allelic variant containing a mutation in the ORF2p PIP motif (L1-PIPm; abolishing PCNA interaction with ORF2p) (3.2.3). The fact that WT RNASEH2B complementation significantly increased LINE-1 retrotransposition to the levels of the parental cell line (Clone 6 +WT, HeLa WT), unlike in the uncomplemented cell line (Clone 6 +EV), confirmed that the observed reduction in retrotransposition in the RNASEH2B KO was due to the absence of RNase H2 (***Figure 3.8***; ***Figure 3.9***). Importantly, there were no differences in WT-L1 retrotransposition levels between the cell lines complemented with the different version of RNASEH2B (Clone 6 +WT *vs* Clone 6 +PIPm), indicating that a functional PIP motif is not essential for RNase H2 to fulfil its function in the LINE-1 retrotransposition cycle. This finding is supported by the fact that the retrotransposition activity of the L1-PIPm and L1-WT were equally affected by the cellular RNase H2 status (Clone 6 +WT *vs* Clone 6 +PIPm *vs* Clone 6 +EV) (***Figure 3.10***). In sum, our data suggest that RNase H2 can promote LINE-1 retrotransposition in a PCNA-independent manner, further suggesting that the conserved and required PIP motif of L1-ORF2p might serve additional roles during LINE-1 retrotransposition (Taylor et al., 2013).

3.8. RNase H activity and LINE-1 retrotransposition

To explore what level of cellular RNase H2 is necessary to support efficient retrotransposition, in the context of cultured HeLa cells, WT RNASEH2B complemented clonal HeLa cell lines showing different levels of RNase H2 activity were generated (3.3.1). Assessing the retrotransposition potential of a WT LINE-1 in these cell lines showed no correlation with the RNase H2 activity levels (3.3.2; ***Figure 3.13***). In fact, a rescue of rate of LINE-1 retrotransposition to near WT levels was already observed at 13% relative RNase H2 activity levels. However, the different complemented clonal cell lines supported various levels of retrotransposition activity, ranging between 50-160% of that observed in WT HeLa cells, while the uncomplemented clonal cell lines (with <6% of WT RNase H2 activity) only supported levels ranging between 15-50%. Since all the clonal cell lines analysed in this experiment are derived from the same CRISPR KO clone, the observed variation is not a result of differences in genetic alterations induced by the CRISPR KO process. The observed variability could be the result of regulatory alterations triggered by the location of the retroviral integration during the complementation and/or the influence of this integration or other factors in the different

ability of each clone to partially compensate some of the absent RNASEH2A functions by overexpressing RNASEH1 (3.1.2). Similarly, the pronounced genomic instability of HeLa cells could influence the overall rate of retrotransposition, especially when "artificial bottlenecks" (i.e., cloning and subcloning) have occurred (Liu et al., 2019). Nonetheless, this data suggests that only low levels of RNase H2 activity are necessary for efficient LINE-1 retrotransposition, although other cellular factors might exert additional influences.

Finally, I tested chimeric human LINE-1s carrying their own RNase H activity in scenarios where (1) cellular RNase H2 was absent; and (2) LINE-1 interaction with PCNA was ablated by a mutation in its PIP motif (3.4). Both, *E. coli* RNASEHI (eRH) and the catalytic domain of the human RNASEH1 (hRH) were assayed as potential RNase H domain in the C-terminus of L1-ORF2p (3.4.1). In order to eliminate the potential deleterious effect of a foreign domain in ORF2p performance in these analyses, I included catalytically dead versions for each RNase H domain assayed as controls. Indeed, although all the chimeric elements were retrotransposition competent (L1-WT_hRH, L1-WT_hRHm, L1-WT_eRH and L1-WT_eRHm), they all showed a lower retrotransposition rate than the L1-WT lacking RNase H, only not statistically significant for the L1-WT_eRHm (***Figure 3.15A***). This confirms that the presence of this foreign domain likely interferes with the functionality of the chimeric LINE-1s, by affecting L1 RNA processing and/or interfering with L1-ORF2p function/stability. Nonetheless, the reduced LINE-1 retrotransposition activity observed for the chimeric constructs containing the active RNase H domains compared to their catalytic dead counterparts indicates that the added RNase H activity, in the designed configuration tested here (***Figure 3.14***), interferes rather than aids LINE-1 retrotransposition (***Figure 3.16***), which was also confirmed for the PIPm LINE-1 constructs (***Figure 3.15B***). Overall, this aligns with our previous finding that the observed reduction in retrotransposition activity of the allelic variant (L1-PIPm) is not due to its inability to use cellular RNase H2 to complete its cycle. Although I attempted to mimic a similar configuration observed in the ORFps of LINE-like elements containing an RNase H domain (1.2.3), in nature these domains have co-evolved over long periods of time leading to complex and well-coordinated interplay that may not be occurring in the configurations tested here. Therefore, I am unable to draw a conclusion regarding whether the human LINE-1 could benefit from encoding its own RNase H activity if adequate co-evolution would occur. More research will be necessary to properly address the research question set out to answer here.

3.9. Proposed model

My results clearly show that RNase H2, and the ability of L1-ORF2p to interact with PCNA are necessary for efficient LINE-1 retrotransposition. In fact, only low levels of cellular RNase H2 activity are sufficient to support efficient retrotransposition rates. However, unlike our original hypothesis, the way RNase H2 fulfils this function appears to be independent from its interaction with PCNA. This means that the previous finding that RNASEH1 overexpression can only partially rescue the LINE-1 phenotype observed in RNase H2 KO cells is not due to its inability to interact with PCNA (Benitez-Guijarro et al., 2018). Potentially RNase H2's ability to remove ribonucleotides from the DNA plays an additional role in the LINE-1 retrotransposition cycle. Although this was partially addressed by Benitez-Guijarro and colleagues, in an experiment using HeLa RNASEH2A KO cell lines complemented with RNase H2 SoF mutant (3.2), this experiment had some technical limitations. The RNase H2 SoF mutant showed reduced substrate affinity and an altered cleavage pattern compared to the WT-RNase H2, making interpretation of these results slightly more complicated. Alternatively, RNase H2's interaction with other cellular factors, not mediated through the PIP motif, could be important for its function in the LINE-1 retrotransposition cycle. Nonetheless, these findings suggest that the need for L1-ORF2p to interact with PCNA to allow efficient retrotransposition is unrelated to the role of RNase H2 in this process, leaving the question what is the function of the L1-ORF2p PIP motif. The PIP motif is highly conserved in the LINE and LINE-like elements across species (Taylor et al., 2013), and it is likely that it facilitates the enzymatic activities encoded in the main ORF of LINE-1s. Since the AP EN domain was added to a pre-existing RT domain along LINE evolution, the origin and acquisition of the PIP domain before or after this event could indicate if its role is more related to either the AP EN or the RT activity.

Previously, it was found that the L1 EN domain is closely related to APE1, a human DNA repair endonuclease that recognizes AP sites (Weichenrieder, Repanas, & Perrakis, 2004). However, there are two AP ENs present in humans and highly conserved in eukaryotes (APE1 and APE2) (Tsuchimoto et al., 2001). By aligning human L1-ORF2p to these proteins, I found that its AP EN domain more closely resembles APE2, with the PIP and Z domains aligning to the APE2 PIP and Zf-GRF motifs respectively (Wallace et al., 2017), absent in APE1 (Tsuchimoto et al., 2001) (1.2.3; ***Figure 3.17***; ***Figure 6.10***). This suggest that the PIP motif might have been acquired together with the AP EN domain from an APE2 ancestral gene in a founder

evolutionary event. Therefore, the function of the PIP-mediated interaction between ORF2p and PCNA might mirror the naturally interaction of APE2 and PCNA. APE2 is a cellular endonuclease and 3'-5' exonuclease involved in the long-patch, base-excision DNA repair pathway (Burkovics et al., 2006; Hadi et al., 2002; Unk et al., 2002). Here, it catalyses the endonucleolytic cleavage of the AP site and the 3'-5' resection of a stretch of ~10nt before other polymerases (δ and ε) are recruited to refill the gap and displace the annealed DNA strand that contains the abasic site. The DNA flap is removed by the FEN-1 endonuclease and the resulting nick is sealed by DNA ligase I. APE2, DNA pol δ and ε, FEN-1 and DNA ligase I all possess PIP motifs and interact with PCNA. PCNA has been proposed to handle these different factors and coordinate their function to facilitate DNA repair (toolbelt hypothesis) (Dovrat et al., 2014; Mayanagi et al., 2018), which aligns with Taylor and colleagues previous suggestion that L1-ORF2p could be interacting with PCNA to facilitate the recruitment of DNA repair machinery in order to assist the completion of the insertion process or the synthesis of the second strand (Taylor et al., 2013).

Figure 3.17. **Schematic alignment of human APE1 and 2 with L1-ORF2p from human, mouse and zebrafish –** A schematic representation of the alignment of the amino acid sequences of the human APE1, APE2, and the L1-ORF2p from human, mouse and zebrafish. Indicated are the conserved regions, AP EN domain, the PIP motif and the Zf-GRF of the APE2 protein and the Z domain of the L1-ORF2ps. The L1-ORF2p Z-domain shows homology with the first part of the Zf-GRF domain of the APE2.

Using *in vitro* models, it has recently been described that during TPRT the single protein encoded by LINE-like R2 elements performs template switching to the opposite strand of the genomic DNA after completing the reverse transcription of the R2 RNA, and that this is necessary for the second strand DNA cleavage to occur (Khadgi, Govindaraju, & Christensen, 2019). This template switching will generate a DNA flap that resembles the DNA structure typically repaired by FEN-1 and DNA ligase during base excision repair. Additionally, it will help

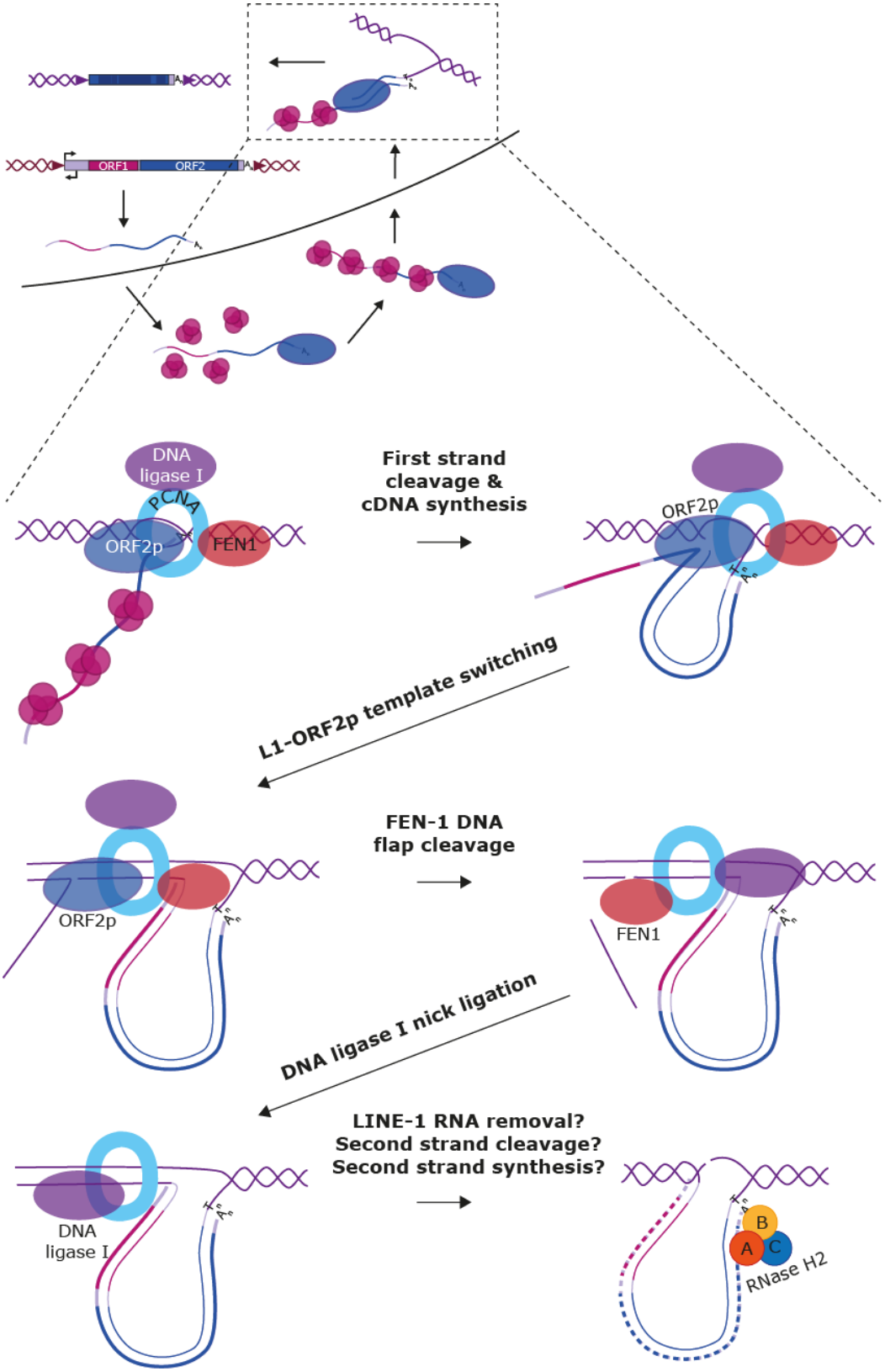


Figure 3.18. Proposed model for the role of L1-ORF2p PIP motif in the retrotransposition cycle – The retrotransposition cycle as shown in Figure 1.2A (top panel). The articulated molecular mechanism proposed here, that integrates the ORF2p interaction with PCNA within the role of APE2-PCNA interaction in the base excision repair pathway (expanded inlet).

preserve the integrity of the DNA strand containing the insertion of the first cDNA before catalysing the second strand cleavage. Notably, the L1-ORF2p has also been shown to possess strong template-switching abilities (Cost et al., 2002). With this new perspective, I hypothesise that L1-ORF2p interaction with PCNA could facilitate: (i) the template-switching by retaining ORF2p at the DNA nick site while reverse transcribing the L1 RNA; and/or (ii) the coordination of FEN-1 and DNA ligase together with L1-ORF2p to mirror the long-patch, base-excision DNA repair pathway, with the L1-ORF2p acting as a combination of APE2/DNA polymerase δ/ε that first incorporates the LINE-1 cDNA before the other factors complete the process in the usual way (*Figure 3.18*). In this proposed model, RNase H2 would fulfil its function at a later stage, when the second strand cDNA is starting to be synthesised, therefore not requiring to be coordinated with L1-ORF2p by PCNA.

Chapter 4

Active LINE-1 elements in the zebrafish genome

4. Active LINE-1 elements in the zebrafish genome

Background

As mentioned in the Introduction (1.2.2), LINE-1 activity in humans has been reported under physiological conditions, during early embryonic development (Garcia-Perez et al., 2007) and in selected brain cells (i.e., neuronal lineage) (Evrony et al., 2016; Sanchez-Luque et al., 2019; Upton et al., 2015). On the other hand, under pathological conditions, LINE-1 activity has been reported in many cancer types (1.3.2) but also in several autoimmune disorders (Ardeljan et al., 2017; Belancio et al., 2010; Harris et al., 2010; Mavragani et al., 2016; Upton et al., 2015) (1.3.3). LINE-1s can impact the genome and cellular function in a myriad of ways, through the expression of their RNA and proteins, as well as acting as insertional mutagens (Belancio et al., 2006; Garcia-Perez, Widmann, & Adams, 2016; Mavragani et al., 2016; Richardson et al., 2015; Roy-Engel et al., 2005; Wallace et al., 2008) (1.3.1). Despite significant research, we know very little regarding the role of LINE-1 in normal cell function, and the extent to which it can cause or contribute to the symptomology of LINE-1-associated disorders. This is partially due to the technical difficulties inherent to retrotransposon research, caused by the repetitive nature of its sequence, the abundance of copies spread throughout the genome, including coding regions, and the polymorphic nature of young and active LINE and SINE copies (Treangen & Salzberg, 2013) (1.4). In fact, even with the development of robust Next Generation DNA sequencing (NGS) methods, it is extremely difficult to identify *de novo* LINE-1 insertions in the germline and soma. Similarly, distinguishing active and inactive copies of LINE/SINE elements is complicated by the presence of different allelic versions of these elements, with some polymorphisms known to impact the retrotransposition potential of elements dramatically (Beck et al., 2010; Sanchez-Luque et al., 2019). Additionally, because of the random nature of LINE-1 retrotransposition and the lack of opportunity to compare patient samples before and after disease occurrence, it is nearly impossible to assign LINE-1 activity as a cause or consequence of a given condition. However, the use of animal models provides researchers a window into the pre-symptomatic phase of L1 related pathologies and represents an opportunity to overcome this significant limitation.

4.1. Animal models in LINE-1 research

Mouse models are frequently used in LINE-1 research to follow retrotransposition *in vivo* (1.4). Transgenic mice containing engineered human or mouse LINE-1 constructs tagged with an EGFP retrotransposition reporter cassette have been a great resource for *in vivo* analyses of LINE-1 biology (An et al., 2008, 2006; Babushok et al., 2006; Kano et al., 2009; Muotri et al., 2005; O'Donnell et al., 2013; Ostertag et al., 2002). However, this research has resulted in discrepant outcomes regarding LINE-1s cell specific activity, most likely influenced by the promoters used to drive the reporter indicating retrotransposition (Muotri et al., 2005; Ostertag et al., 2002; Prak et al., 2003). This occurrence highlights the benefit of investigating LINE-1 activity and regulation using endogenous LINE-1 elements. The mouse genome does contain active endogenous LINE-1 elements, and research using them as a model has greatly expanded our knowledge on the regulation of mammalian LINE-1s (Crichton et al., 2014; Goodier, Ostertag, Du, & Kazazian, 2001; Hardies et al., 2000; Richardson et al., 2017; Trelogan & Martin, 1995). However, working with mouse models comes with limitations; tissue processing can be time consuming and laborious; generating transgenic mice takes a long time because of their relatively slow breeding and small litter size and their maintenance comes with high economical costs, especially for populations-based studies. Additionally, mouse and human LINE-1s differ in certain aspects, such as the structure of their promoters (unitary *vs* multimeric respectively) and the mechanism used to translate ORF2 from the bicistronic L1 RNA (ribosome re-initiation event *vs* IRES respectively) (1.2.3), implying that not all knowledge obtained with mouse LINE-1s can be directly extrapolated to the human LINE-1s. Thus, having access to multiple models containing endogenous LINEs would be of great benefit, as different aspects from different LINEs may model the human LINE-1 more accurately.

Unfortunately, alternative easy to use animal models, such as yeast (*Saccharomyces cerevisiae*), fly (*Drosophila spp*) and nematode (*Caenorhabditis elegans*) have limited value in LINE research. Yeast and *C. elegans* genomes do not contain active LINEs, while LINEs in fly have become subject of molecular domestication, leading to them acting as telomeres/Telomerase. The later implies their regulation/impact is incomparable to that of human LINE-1s. Recently, 17 different LINE-1 subfamilies have been identified in the zebrafish reference genome (Boissinot & Sookdeo, 2016). Although there expression and activity remains unconfirmed, a study assessing LINE-1 retrotransposon lineages in teleost fish,

including zebrafish, has suggested the presence of multiple divergent, active LINE-1 lineages in the genomes of these species (Duvernell et al., 2004). This prompted us to further investigate this possibility.

4.2. Zebrafish as a potential model in LINE-1 research

Zebrafish represent an emerging, amenable and valid model organism to study human disease, particularly brain disorders and behavioural genetics, due to several of their characteristics (Kalueff, Stewart, & Gerlai, 2014; Norton & Bally-Cuif, 2010). Their small size and lack of pigment during the first stages of life (and longer when using certain mutant strains) allows various imaging techniques unachievable in mammalian models, such as real-time *in vivo* imaging (Antinucci & Hindges, 2016; White et al., 2008). Furthermore, transgenic zebrafish are easier, faster and more economical to generate and maintain than transgenic mammals, allowing large scale analysis using a variety of methods (transient transgenics, KD by morpholinos and/or CRISPR-Cas9) (Kearns et al., 2014; Lieschke & Currie, 2007). Nearly half of the zebrafish genome is made up of TEs, perhaps making it the vertebrate with the highest percentage of TE-derived sequences (Howe et al., 2013a). Beside their abundance, it is also notable that many TE-types are still active in zebrafish, including zebrafish LINE-1s, as inferred from the establishment of the first zebrafish genome assembly (Howe et al., 2013a). Considering this and the random integration of human LINE-1s in somatic tissues, in combination with the significant technical advantages zebrafish bring, might make zebrafish an ideal model to infer the impact of retrotransposition under physiological and pathological conditions in the soma. For this and other reasons, I hypothesize that zebrafish could be a valuable model for LINE-1 research. Particularly, it could help to shed light on the long standing question: what is the role, if any, of LINE-1 activity in brain and their effect on behaviour and the development of neurological disorders? However, although genomic analyses have identified LINE-1 elements containing intact ORFs (Boissinot & Sookdeo, 2016), to date no active LINE-1 has been characterised in the zebrafish genome.

A few years ago, our lab started to explore whether zebrafish could be an amenable and robust model to study LINEs (Widmann et al., in preparation). To date we have been limited to the study of two active copies from two distinct zebrafish LINE-2 retrotransposon

subfamilies, Zfl2-1 and Zfl2-2, originally characterised in human cells by the Okada Lab (Sugano et al., 2006). However, LINE-2 retrotransposons are ancient TEs from the Mesozoic era (Jerzy Jurka, Zietkiewicz, & Labuda, 1995), and are no longer active in the human genome. Although we observed similarities in the general biology of zebrafish LINE-2s compared to human LINE-1s (i.e., both are expressed at the same developmental stages), there are fundamental differences between these two clades of LINEs (1.2.3), limiting the impact of our findings. Thus, I propose that zebrafish LINE-1s should be used to realistically understand the impact of LINE-1 retrotransposition *in vivo* using zebrafish, once their activity has been confirmed/validated.

4.3. LINE-1s in zebrafish

As previously mentioned, to date, 17 different LINE-1 subfamilies have been identified in the zebrafish reference genome, but whether this lineage of retrotransposons is expressed and/or currently active in this species is unknown (Boissinot & Sookdeo, 2016). Like the human LINE-1, the zebrafish LINE-1 (ZfL1) possess a unitary promoter, two non-overlapping ORFs separated by an IGR and ends in a 3'UTR (***Figure 4.1***). The absence of an obvious stem-loop in their 3'UTR and the fact that the copies present in the genome end in a poly-A tract suggests both are relaxed non-LTR elements (1.2.3). Comparing the structure of the ZfL1s and human LINE-1s, there are two notable differences: the length of their 5'UTRs and the IGRs (***Figure 4.1***). While the 5'UTR of ZfL1s are much shorter than that of human LINE-1s (on average ~200 bp *vs* ~1000 bp), the IGRs are much longer (on average ~500 bp *vs* ~50 bp). Although it is not surprising to find differences in the 5'UTR, as this region shows little to no homology even among LINE-1 families within the same species (1.2.3), the difference in IGR length could indicate both elements utilise a different mechanisms to allow translation of the second ORF. Finally, different to humans, multiple LINE-1 subfamilies could be active in zebrafish at the same time, as reported in many other vertebrates, such as mice. In fact, humans represent the exception to LINE-1 activity in vertebrates.

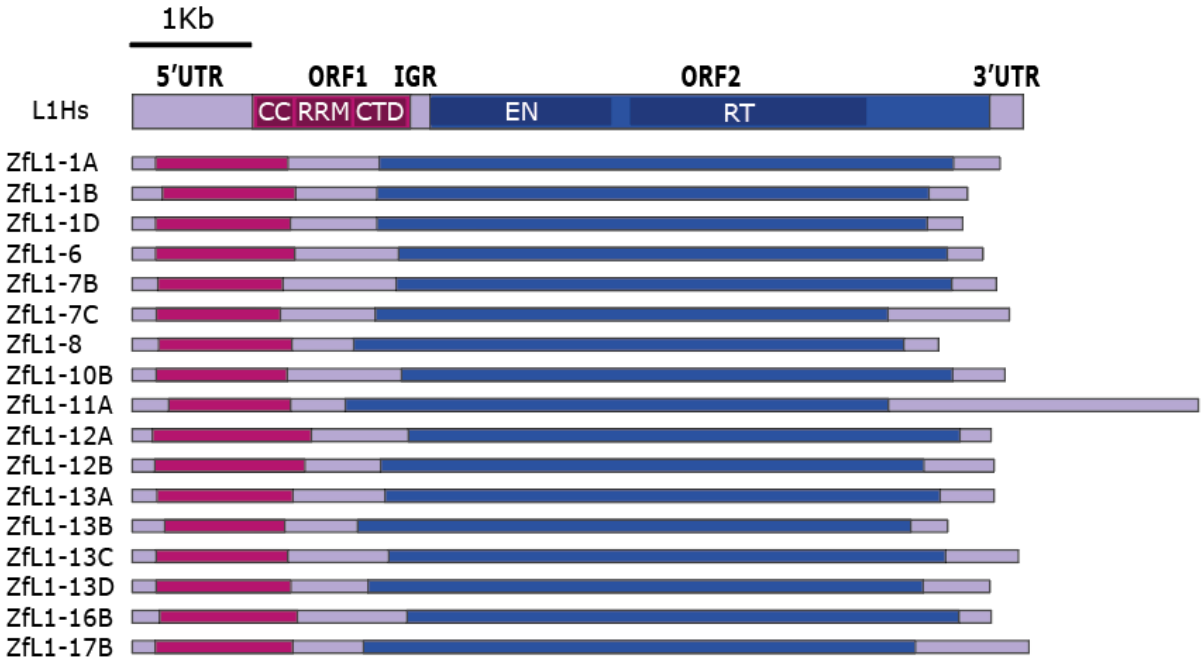


Figure 4.1. LINE-1 subfamilies described in zebrafish – A scaled schematic overview of the 17 Zebrafish LINE-1 (ZfL1) subfamilies described by Boissinot & Sookdeo in 2016, compared to an active human LINE-1 L1Hs consensus. CC, Coiled Coil domain; RRM, RNA Recognition Motif; CTD, Carboxy Terminal Domain; EN, ENdonuclease; RT, Reverse Transcriptase. Colour code: 5'UTR, InterGenic Region (IGR) and 3'UTR, light purple; ORF1, red; ORF2, blue. The scale on the top indicates length (1Kb). Figure adapted from Boissinot & Sookdeo in 2016.

4.3.1. Aims and objectives

In this chapter, I describe a series of analyses and experiments aimed to get a better understanding of whether active LINE-1 copies can be found in the zebrafish genome reference draft. Briefly, I performed *in silico* analyses of the 17 ZfL1 subfamilies to determine: (1) the presence of ZfL1 copies in the zebrafish reference genome with the potential to encode the machinery required for retrotransposition, and (2) their expression levels along zebrafish embryonic development and gastrulation. To complement these analyses, (3) I tested the retrotransposition activity of some of the identified copies using several experimental approaches.

4.4. Identification of potentially active LINE-1 subfamilies in the zebrafish genome

4.4.1. Sequence analyses of full-length LINE-1 copies annotated in the zebrafish reference genome; generation and analysis of functional consensus sequences

A necessary step to determine the presence of retrotransposition competent LINE-1s in the zebrafish genome is to find candidate LINE-1 copies within the different LINE-1 subfamilies. To do this, I first analysed the loci and sequence of all the full-length copies of the 17 different ZfL1 subfamilies present in the zebrafish reference genome (2.4.2) (workflow depicted in ***Figure 4.2***). Briefly, I used the reference genome assembly available at the UCSC genome browser (Assembly version: GRCz10/danRer10), in combination with the consensus sequences provided by Boissinot & Sookdeo (2016) for the 17 different zebrafish LINE-1 subfamilies, to identify the loci containing full-length ZfL1 copies by BLAT (Kent, 2002). Next, I retrieved the sequence of each full-length copy, including 150pb of the flanking genomic sequences upstream and downstream of the insertion. As sporadic mutations are expected to build up over time, and given the reported variability of ZfL1 sequences (reviewed in Furano, Duvernell, & Boissinot, 2004), a number of variables were analysed to determine whether a given LINE-1 copy was recently integrated in the genome: i) the presence of intact TSDs at both sides of the LINE-1; ii) a high % sequence identity with the consensus sequence; and, iii) the presence of a clean poly-A tract at the 3'end of the insertion. The copies of the ZfL1-1B element did not end in a poly-A tract, but instead contained short-repeats of variable length at their 3' end, which are typical for LINE-2 retrotransposons (Sugano et al., 2006). These elements were therefore excluded from further analyses.

Subsequently, each sequence was analysed using the online tool ORFfinder, to identify presence of intact ORFs (i.e., the two expected ORFs described in ZfL1s: ORF1 and ORF2). Here I found that none of the copies of the ZfL1-7B subfamily identified in the zebrafish reference genome possessed intact ORF2s, and this subfamily was therefore excluded from further

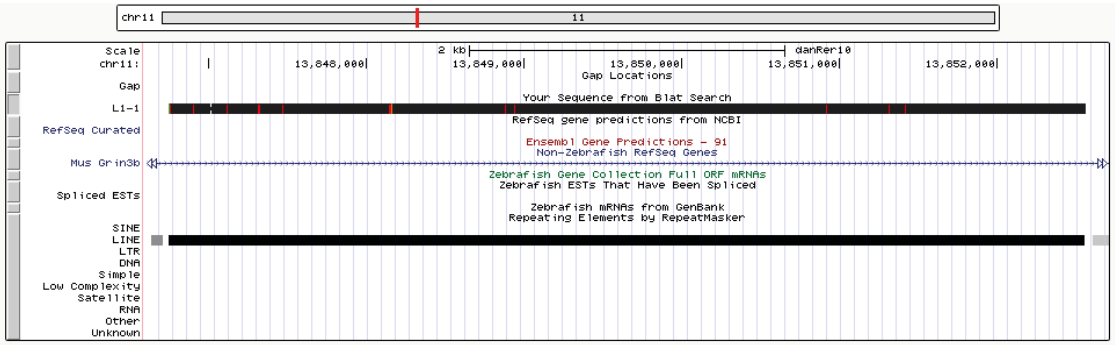
analyses. Additionally, when considering the standard AUG initiation codon, all the identified copies of the ZfL1-12B and ZfL1-16B subfamilies presented with a 5' truncated ORF2p. However, re-analyses considering alternative initiation codons identified the alternative initiation codon TTG as the likely start of translation for these ORFs.

Finally, a preliminary analysis of the amino acid sequences encoded by each ORF was performed using the online tool MotifScan, to determine whether their functional domains remain intact enough to identify the protein families they belong to. The Pfam database (version 32.0) was chosen as a source (El-Gebali et al., 2019), as it contains a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs) (Eddy, 2004). In this analysis, intact ORF1p's were recognised as being related to the transposases proteins, from the L1 transposable element family, while intact ORF2p's were confirmed to contain a domain related to endonuclease/exonuclease/phosphatase proteins followed by a domain related to reverse transcriptase (RNA-dependent DNA polymerase) proteins.

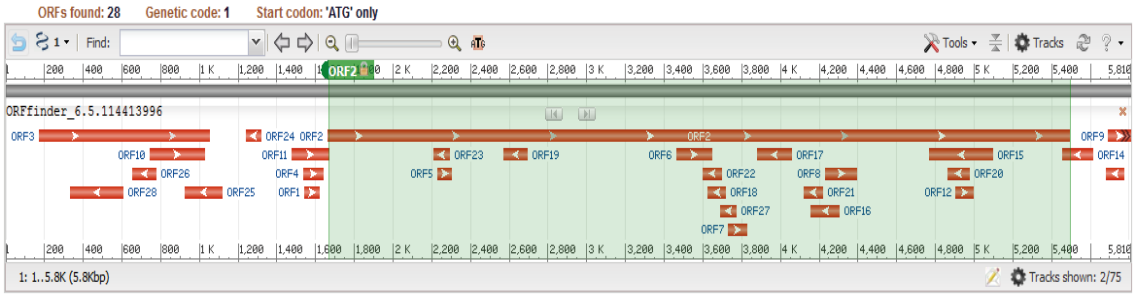
The data from these sequence analyses are summarised in *Table 4.1*. For each of the 15 remaining ZfL1 subfamilies, a functional consensus sequence was build using the sequence of all the identified full-length copies containing the two intact ORFs (ORF1 and ORF2). The functional consensus sequences of the three subfamilies later included in the functional analyses can be found in the Appendix (*Sequence 6.1 – 6.3*). Next, the amino acid sequences of the ZfL1 ORFs encoded by these consensus sequences were compared to human (L1Hs) and mouse (L1-MdA) consensus sequences to further analyse the presence of key conserved amino acids required for the enzymatic activities of these domains (Christian et al., 2016; Gilbert & Moran, 2002).

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	L1-1	5767	1	5800	5801	99.8%	11	-	13846748	13852556	5809
browser details	L1-1	5766	1	5801	5801	99.8%	12	+	9946386	9952191	5806
browser details	L1-1	5753	1	5801	5801	99.7%	14	-	7914582	7920382	5801
browser details	L1-1	5747	4	5801	5801	99.7%	4	+	34649955	34655753	5799
browser details	L1-1	5734	13	5801	5801	99.6%	3	+	27340460	27346230	5771
browser details	L1-1	5724	1	5801	5801	99.4%	4	+	60329909	60335695	5787
browser details	L1-1	5722	1	5801	5801	99.4%	13	+	110768	116543	5776
browser details	L1-1	5712	1	5796	5801	99.5%	2	-	58303840	58309644	5805

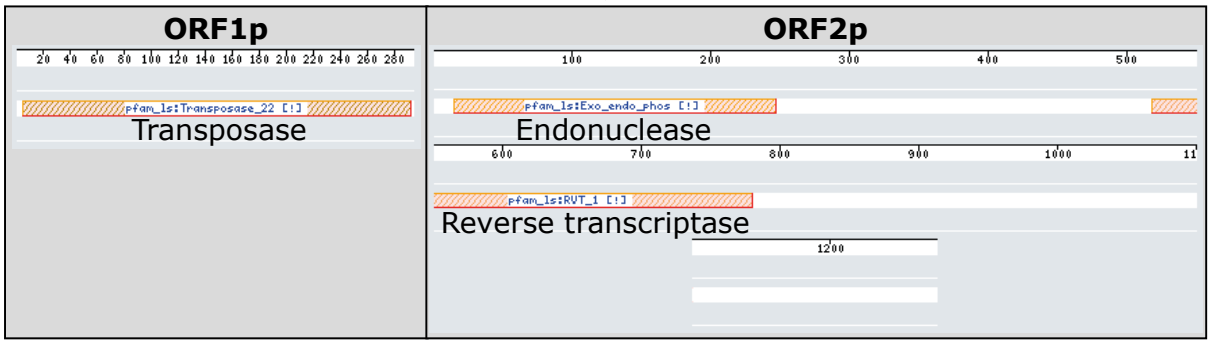
Identify loci containing full length copies of the 17 LINE-1 subfamilies in the USCS provided zebrafish reference genome



Obtain sequence of the full-length copies + 150bp genomic region up and downstream



Identify presence of the two intact ORFs using ORFfinder



Assess presence of enzymatic domains in ORF encoded proteins using Motif scan

Figure 4.2. **Identification and analysis of full-length LINE-1 copies in the zebrafish reference genome** – The figure shows the workflow used to identify and assess the potential retrotransposition competency of the full-length LINE-1 copies identified in the zebrafish reference genome. Images were captured from the UCSC genome browser, ORFinder and MotifScan.

The analyses of the RRM and CTD domains found in the ORF1p is visualised in ***Figure 4.3A and B***. The analyses of nine subdomains within the EN and RT domains, as well as the presence of the PIP, Z and C domains of the ORF2p are shown in ***Figure 4.4A, B, C, D and E***. The overall data is summarised in ***Table 4.2***, which indicates a conservation score (%), reflecting the conserved percentage at the amino acid level, of each functional domain within the ORF1p and ORF2p, as well as the overall score per LINE-1 subfamily, considering all domains together.

ZfL1-subfamily	Full-length copies (#)	Average sequence identity (%)	Copies flanked by intact TSDs (#)	Copies containing intact ORFs (#)
ZfL1-1A	8	99.6	8	6
ZfL1-1B	6	99.7	4	4
ZfL1-1D	6	99.6	5	4
ZfL1-6	7	99.8	6	6
ZfL1-7B	8	99.8	8	7
ZfL1-7C	4	99.1	0	0
ZfL1-8	5	99.8	4	4
ZfL1-10B	8	99.8	8	5
ZfL1-11A	3	99.2	1	2
ZfL1-12A	5	99.5	4	3
ZfL1-12B	2	99.9	2	2
ZfL1-13A	4	99.6	3	2
ZfL1-13B	3	99.5	1	1
ZfL1-13C	4	99.8	4	3
ZfL1-13D	6	99.7	4	5
ZfL1-16B	6	99.8	4	2
ZfL1-17B	4	99.8	4	3

Table 4.1. **Analysis of full-length ZfL1 copies found in the zebrafish reference genome for each of the 17 subfamilies** – For each subfamily, the number of full-length copies identified, the average nucleotide identity (%) of these copies, the number of elements flanked by intact TSDs, and the number of elements containing intact ORFs, is shown. Highlighted in blue are the three subfamilies later analysed in functional assays (4.3).

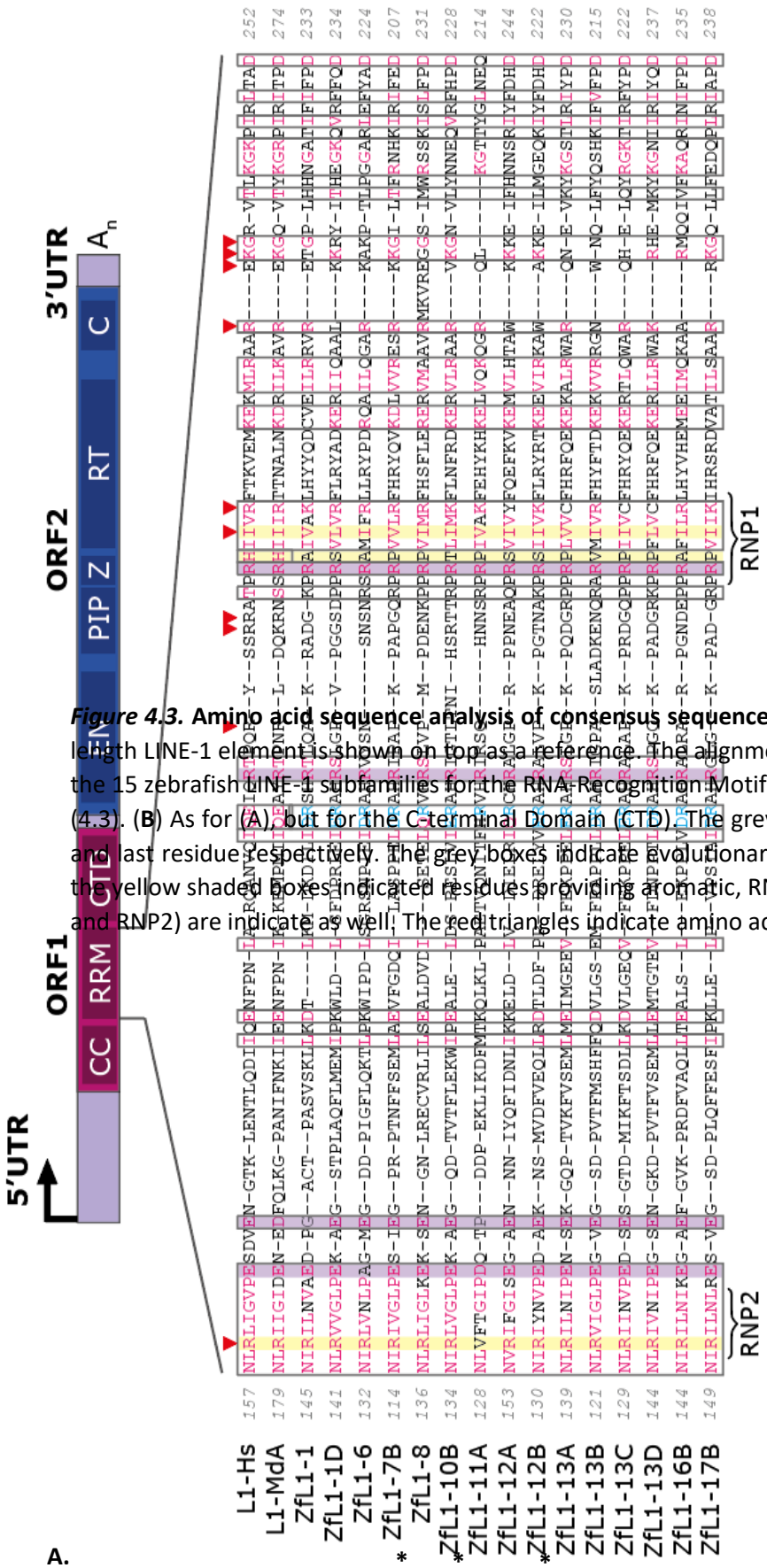


Figure 4.3. **Amino acid sequence analysis of consensus sequence encoded ORF1p of a human, mouse and th**
length LINE-1 element is shown on top as a reference. The alignment corresponds to the amino acid sequence
the 15 zebrafish LINE-1 subfamilies for the RNA-Recognition Motif in ORF1p. Indicated with an * are the three
(4.3). (**B**) As for (A), but for the C-terminal Domain (CTD). The grey numbers flanking the amino acid sequence
and last residue respectively. The grey boxes indicate evolutionary conserved regions, while the purple shade
the yellow shaded boxes indicated residues providing aromatic, RNA-binding side-chains in canonical RRMs. Th
and RNP2) are indicate as well. The red triangles indicate amino acids identified as essential for retrotransposit

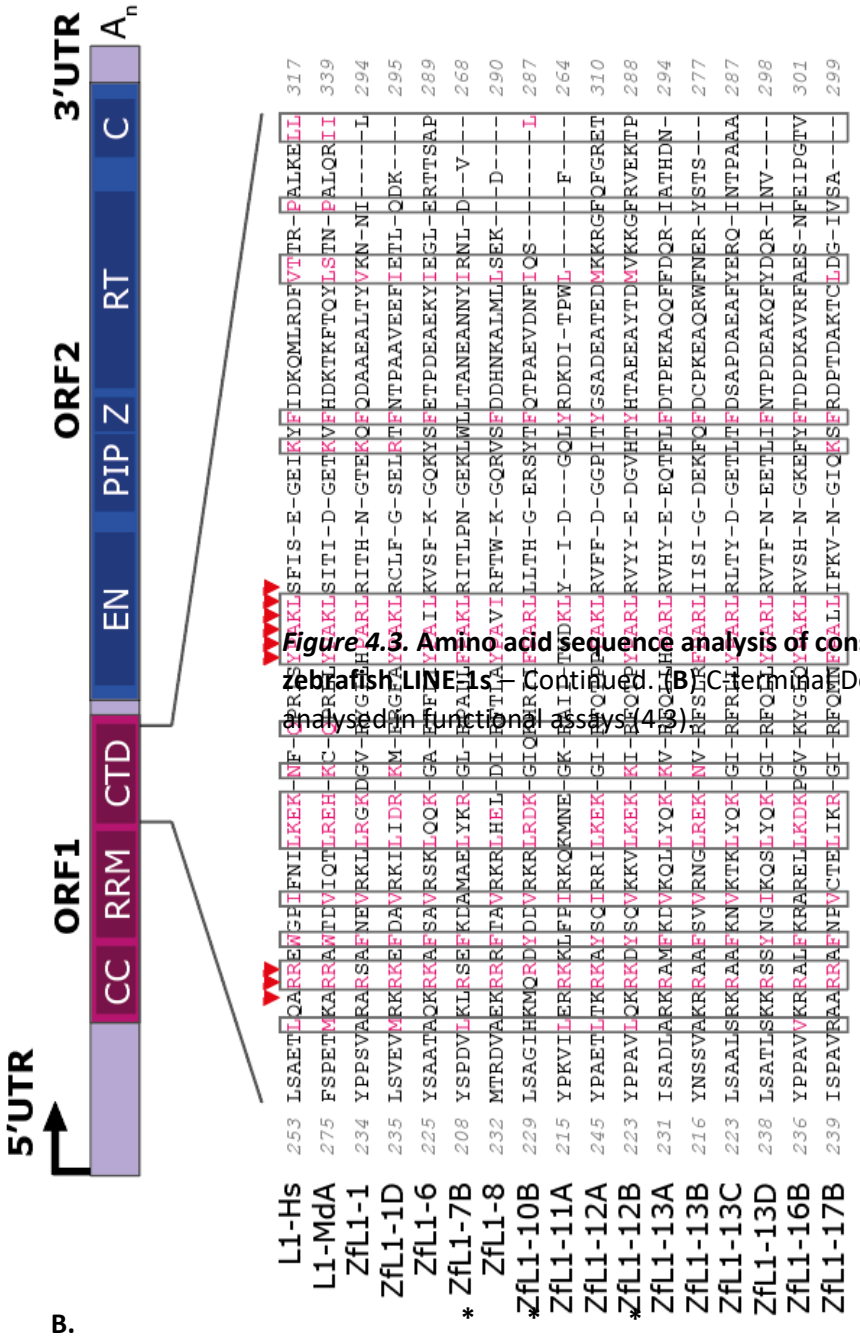


Figure 4.3. **Amino acid sequence analysis of consensus sequence encoded ORF1p of a huma**
zebrafish LINE-1s – Continued. **(B)** C-terminal Domain (CTD). Indicated with an * are the th
analysed in functional assays (4.3).

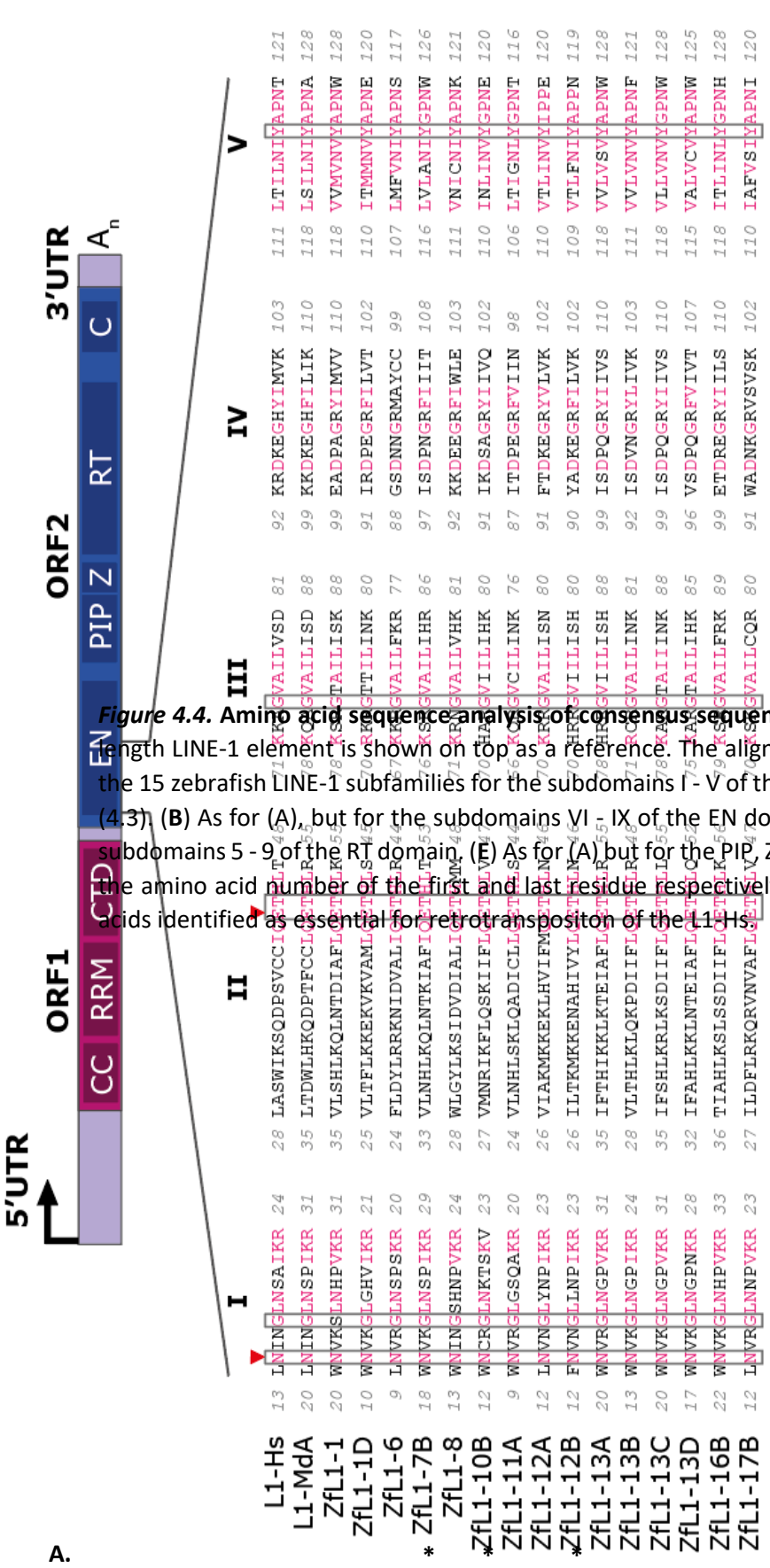


Figure 4.4. **Amino acid sequence analysis of consensus sequence encoded ORF2p of a human, mouse and**
length LINE-1 element is shown on top as a reference. The alignment corresponds to the amino acid sequen
the 15 zebrafish LINE-1 subfamilies for the subdomains I - V of the EN domain. Indicated with an * are the thr
(4.3). (**B**) As for (A), but for the subdomains VI - IX of the EN domain. (**C**) As for (A) but for the subdomains
subdomains 5 - 9 of the RT domain. (**E**) As for (A) but for the PIP, Z and C domain of the L1-ORF2p. The grey nu
the amino acid number of the first and last residue respectively. The grey boxes indicate evolutionary con
acids identified as essential for retrotranspositon of the L1-Hs.

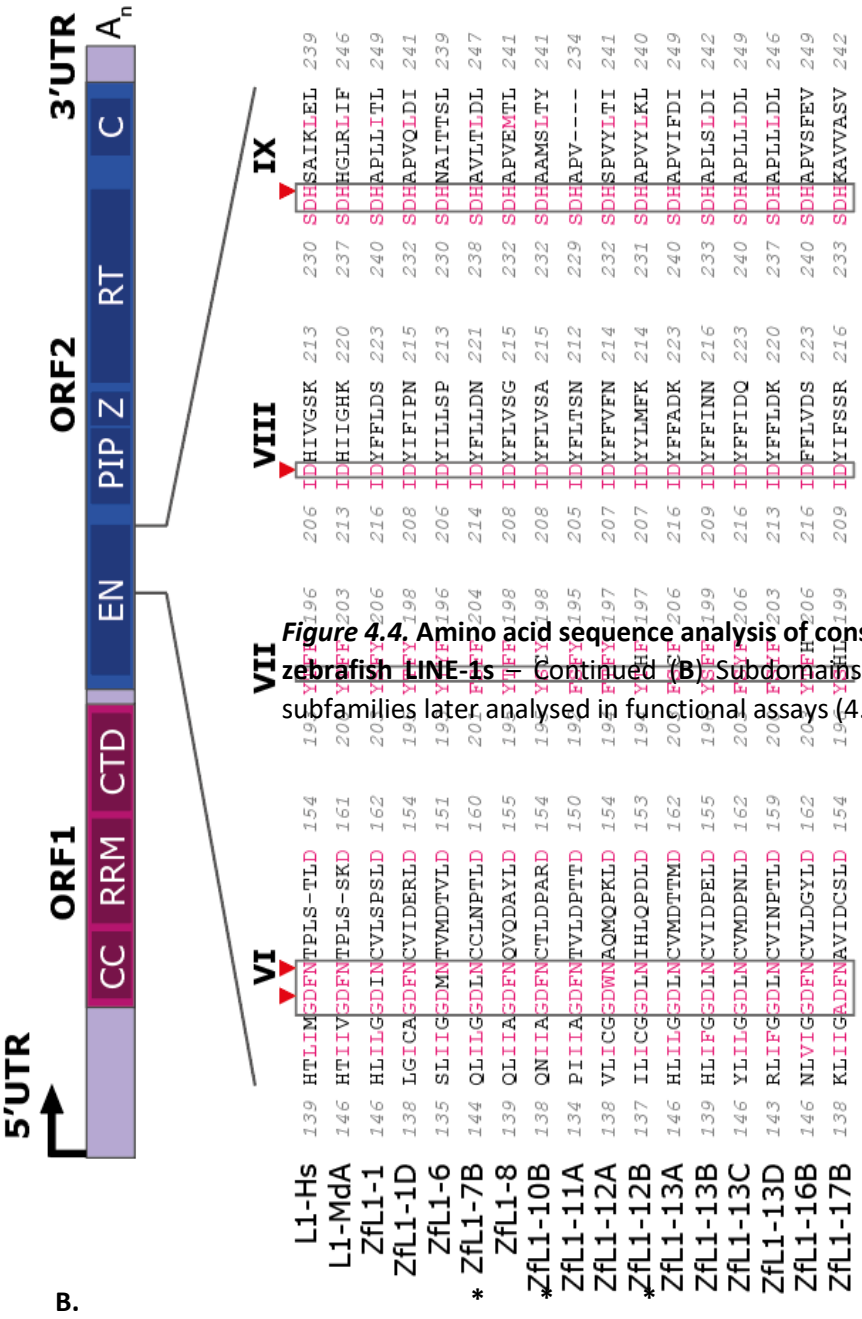


Figure 4.4. **Amino acid sequence analysis of consensus sequence encoded ORF2p of a human and zebrafish LINE-1s** – Continued (**B**) Subdomains VI - IX of the EN domain. Indicated with * subfamilies later analysed in functional assays (4.3).

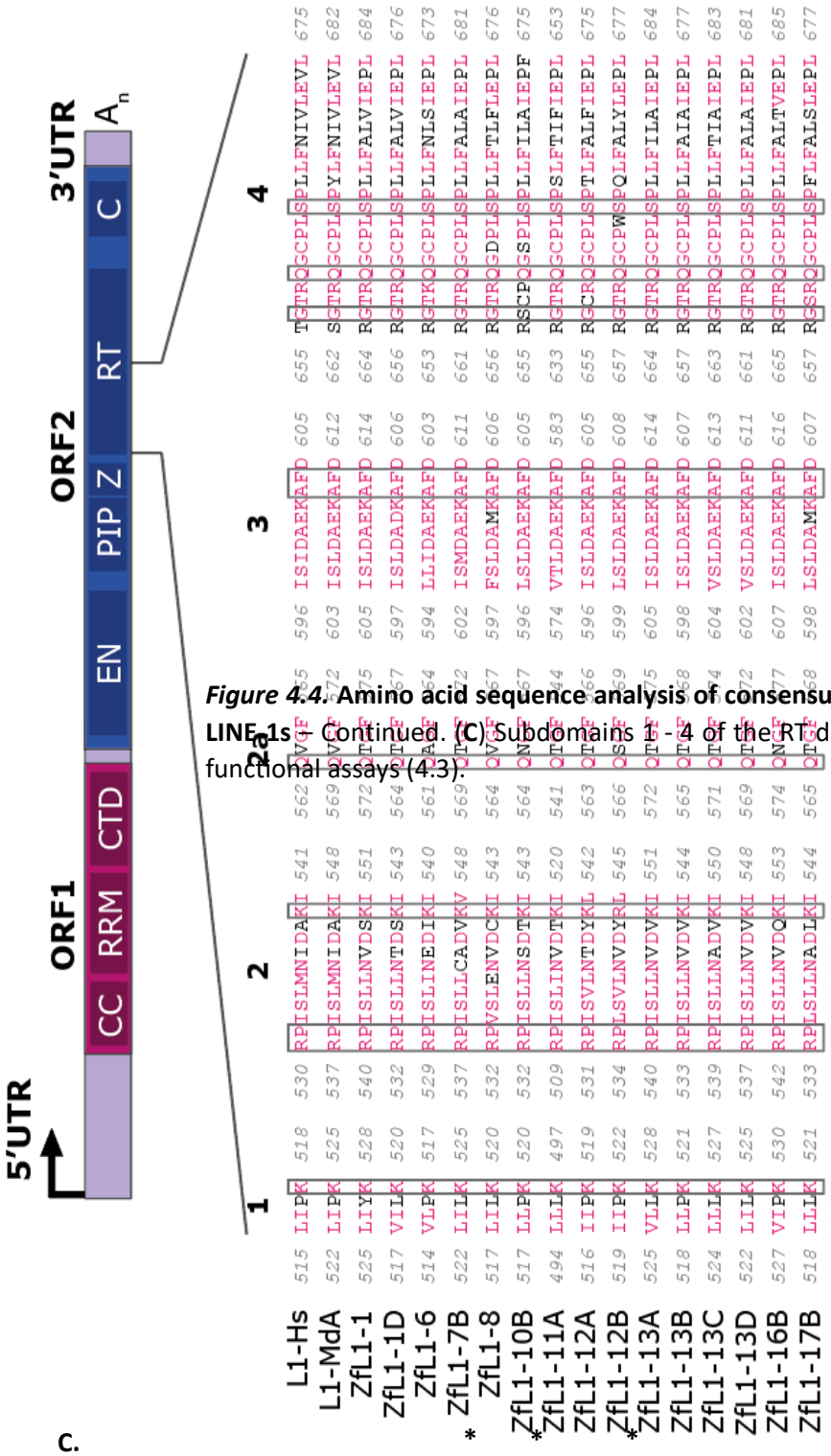


Figure 4.4. **Amino acid sequence analysis of consensus sequence encoded ORF2p of a human, mo**
LINE-1s – Continued. (**C**) Subdomains 1 - 4 of the RT domain. Indicated with an * are the three sub
functional assays (4.3).

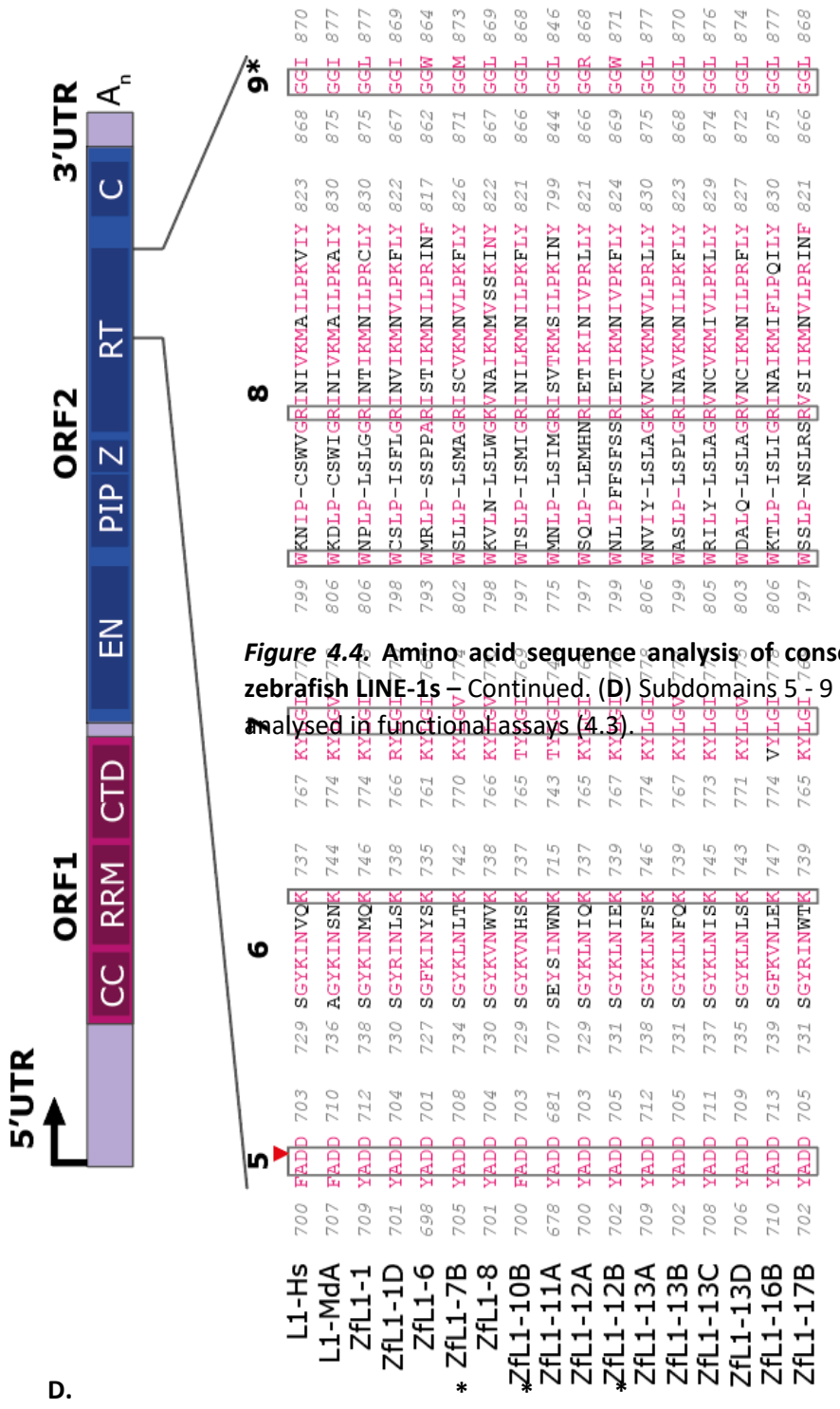


Figure 4.4. **Amino acid sequence analysis of consensus sequence encoded ORF2p of a hu**
zebrafish LINE-1s – Continued. (**D**) Subdomains 5 - 9 of the RT domain. Indicated with an * are t
analysed in functional assays (4.3).

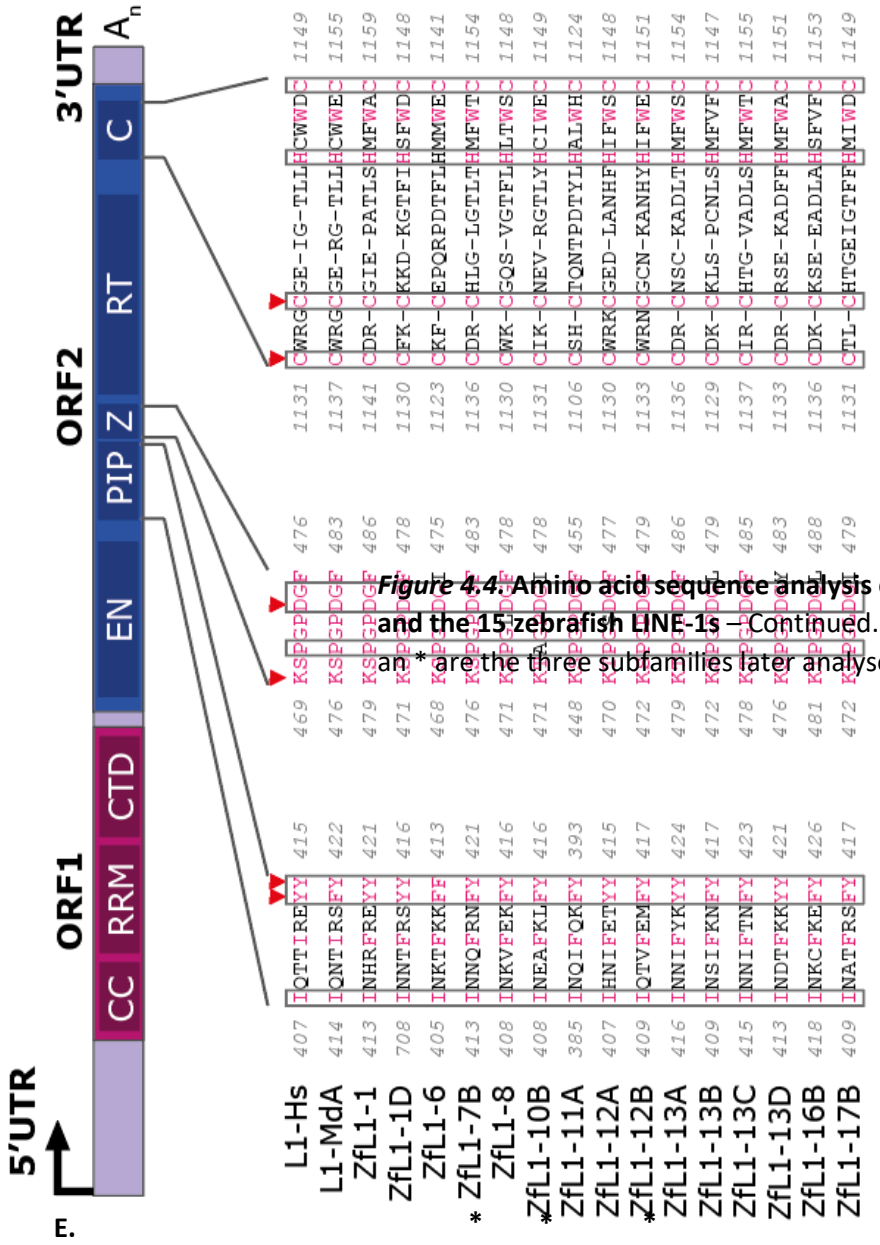


Figure 4.4. **Amino acid sequence analysis of consensus sequence encoded ORF2p of a and the 15 zebrafish LINE-1s** – Continued. (**E**) The PIP, Z and C domain of the L1-ORF2 an * are the three subfamilies later analysed in functional assays (4.3).

	Percentage conserved amino acids of analysed							
	ORF1p		***ORF2p***					***Overall score***
	RRM	*CTD*	*EN*	*PIP*	*Z*	*RT*	*C*	
L1-Hs	100%	100%	100%	100%	100%	100%	100%	100%
L1-MdA	98%	100%	100%	100%	100%	100%	100%	100%
ZfL1-1	70%	57%	86%	100%	100%	100%	100%	88%
ZfL1-1D	80%	74%	100%	100%	100%	100%	100%	91%
ZfL1-6	60%	52%	93%	100%	100%	100%	100%	82%
ZfL1-7B	88%	52%	93%	100%	100%	100%	100%	90%
ZfL1-8	80%	52%	100%	100%	100%	100%	100%	85%
ZfL1-10B	80%	65%	100%	100%	67%	95%	100%	86%
ZfL1-11A	53%	35%	100%	100%	100%	100%	100%	79%
ZfL1-12A	68%	65%	100%	100%	100%	100%	100%	86%
ZfL1-12B	65%	74%	93%	100%	100%	100%	100%	86%
ZfL1-13A	80%	48%	93%	100%	100%	100%	100%	86%
ZfL1-13B	68%	61%	93%	100%	100%	100%	100%	88%
ZfL1-13C	75%	48%	93%	100%	100%	100%	100%	87%
ZfL1-13D	83%	48%	93%	100%	100%	100%	100%	87%
ZfL1-16B	70%	57%	100%	100%	100%	100%	100%	86%
ZfL1-17B	70%	57%	100%	100%	100%	100%	100%	84%

Table 4.2. Summary of the analysis at the amino acid level for ORF1p and ORF2p encoded in the functional consensus sequence – The data in this table is a summary of the analysis of the conserved amino acids specified by Moran & Gilbert (2002) in ORF1p and ORF2p, encoded in the functional consensus sequence of the 15 ZfL1 subfamilies, compared to the human L1-Hs. The human and mouse (L1-MdA) LINE-1s are included as references. Highlighted in blue are the three subfamilies later analysed in functional assays (4.3).

Detailed analyses of the different conserved domains found in the ZfL1-ORF1p and -ORF2p revealed an overall higher conservation in the ORF2p among the ZfL1 subfamilies compared to the human (L1-Hs) and mouse LINE-1 (L1-MdA) (***Table 4.2***). Within the ZfL1-ORF1p, the RRM is the best conserved domain, although all ZfL1 subfamilies lack one of the two conserved residues of the RNP1 motif, usually providing aromatic RNA-binding side-chains in canonical RRMs (***Figure 4.3A***). However, mutation of this residue has not been shown to affect mammalian LINE-1 retrotransposition (Khazina & Weichenrieder, 2009). Only ZfL1-11A has a non-compatible amino acid substitution of a similar residue in the RNP2 motif. All four salt bridges are completely conserved in most ZfL1 subfamilies, except ZfL1-1, -6 and 11A. On the other hand, the CTD domain showed less conservation (***Table 4.2***), although most of the amino acids identified to be essential for human LINE-1 retrotransposition remain unchanged in the ZfL1 subfamilies (***Figure 4.3B***). ZfL1-ORF2p is nearly completely conserved in most ZfL1 subfamilies compared to mouse and human LINE-1s, with the EN domain showing the most

deviation (*Table 4.2*). Nonetheless, all amino acids identified as essential for human LINE-1 retrotransposition are unchanged in all the EN domains (*Figure 4.4A* - *E*).

4.4.2. ZfL1 expression along zebrafish development

To complement the above sequence analyses, we next used publicly available RNA-sequencing data at different zebrafish developmental time points to analyse ZfL1 expression. Specifically, in collaboration with a bio-informatician, we explored the expression of the 15 ZfL1 subfamilies, as well as a LINE-2, ZfL2-2, used as a positive control, at the following stages: 2-4 cell, 1K cell, dome, shield bud, 28 hpf, two and five dpf. We first analysed the data by mapping raw sequencing reads to the full-length functional consensus sequences, establishing RPKM values of ZfL1 and ZfL2-2 expression (2.4.3; *Figure 4.5A*). ZfL1-7B, -11A, and -13A subfamilies showed the highest expression along all zebrafish developmental stages, comparable to that found for the positive control, ZfL2-2. The ZfL1-17B shows the lowest expression across all developmental stages. ZfL1-1, -1D, -6, -8, -10B, -12A, -12B, -13C, -16B and -17B subfamilies were expressed at barely detectable levels in the first two developmental stages, the 2-4 and 1K cell stage. Overall, the highest measured ZfL1 expression is at the dome and shield stage.

As an internal control, the distribution of reads along the consensus sequence of each LINE was analysed (*Figure 4.5B*). Most ZfL1 subfamilies, as well as the ZfL2-2 show an even distribution of reads across their sequence, with a clear enrichment at their 3´ends. However, we noticed artefactual mapping of reads to particular regions in certain ZFL1 subfamilies: ZfL1-7B, -11A, -13A, -13B and -13D. For these subfamilies, a disproportional number of reads aligned to just a single area of the consensus, which correspond to homopolymeric tracts present in the 3' UTR and/or intergenic regions of these ZfL1s (*Figure 4.5B*). To overcome the artefactual mapping to homopolymeric regions, and to focus mainly on potentially full-length LINE RNAs, we next mapped reads to just the first 1000 bp of each of the 15 ZfL1 consensus sequences, using the LINE-2 ZfL2-2 as a control, establishing the FPKM values of ZfL1 and ZfL2-2 expression (2.4.3; *Figure 4.6A*). This resulted in a more homogenous read coverage for all ZfL1 subfamilies (*Figure 4.6B*). Whereas the ZfL1-7B, -11A and -13A subfamilies previously showed a comparable high expression across all zebrafish developmental stages, they now follow a similar pattern to the other subfamilies, with lower expression levels in the earlier stages of development (*Figure 4.6A*). ZfL1-1D and -11A subfamilies now show the highest expression, comparable to that

found for ZfL2-2 LINE-2s, particularly in the later developmental stages starting at the bud stage and dome stage respectively. ZfL1-17B still shows the lowest expression across all developmental stages. Transcripts of most subfamilies, were detected at very low levels at the 2-4 cell stage, while barely detected at the 1K cell stage, including the positive control. The developmental stage at which the highest ZfL1 expression is measured, is now more variable between subfamilies, although it was consistently higher at the later stages of development (28 hpf - 5 dpf).

To analyse the expression pattern during zebrafish development with a different approach, I used RT-qPCR and a TU × AB breeding pair from the zebrafish colony housed at IGC (Zebrafish facility). Briefly, total RNA was isolated from zebrafish embryo and larvae at the same developmental stages analysed above [2-4 cell, 1K cell, dome, shield and bud stage as well as 28 hpf, 2 dpf and 5 dpf], and was converted into cDNA using random hexamers (2.4.3.2). Two primer sets targeting the 5'end of each of the 15 ZfL1 elements and the ZfL2-2 element were tested, as well as primers against the housekeeping genes EF1-α and Act-β. To confirm the presence of a single amplicon, the corresponding melting curves were analysed, and the reactions were resolved by electrophoresis on agarose gels (2.2.1.5). The best performing primer set for each ZfL1 element was used. Of the two housekeeping genes, Act-β was expressed consistently across all developmental stages and was therefore used to calculate expression differences (***Figure 4.7***). The expression pattern of each element along development, was analysed. Since the primer efficiency was not calculated for each primer set, potential differences in amplification efficiency prohibits a side-by-side comparison of LINE expression. Similar to the RNA-seq results, the RT-qPCR shows higher expression at the later developmental stages, starting around the shield stage (***Figure 4.7***).

A.

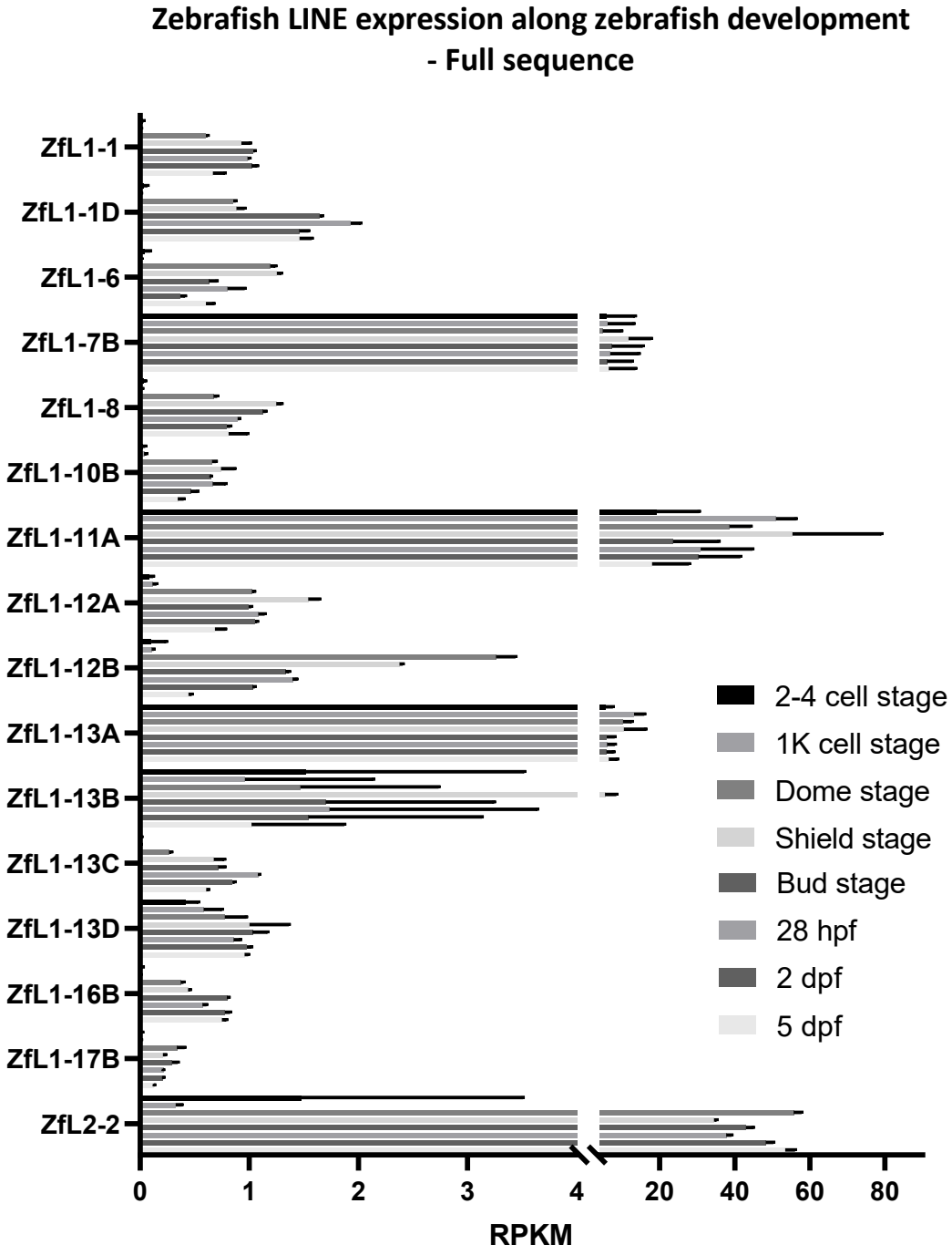


Figure 4.5. Zebrafish LINE expression along zebrafish development (full-length analysis) – (A)The graph shows the average RPKM mapped to the full sequence of the 15 ZfL1 subfamilies [ZfL1-1, -1D, -6, 7B, -8, -10B, -11A, -12A, -12B, -13A, -13B, -13C, -13D, -16B and -17B] and the ZfL2-2 element in eight different zebrafish developmental stages [2-4 cell, 1K cell, dome, shield and bud stages, 28 hpf, 2 and 5 dpf] of two biological replicates. Error bars represent standard deviation between biological replicates. (B) [in the next page] A graphic representation of the sequence coverage. A schematic of the 15 ZfL1s and the ZfL2-2. Colour code: 5'UTR, IGR and 3'UTR, purple; ORF1, red; ORF2, blue. The length scale (bp) is indicated in the centre and the mapped sequence coverage is indicated in the bottom, in dark blue.

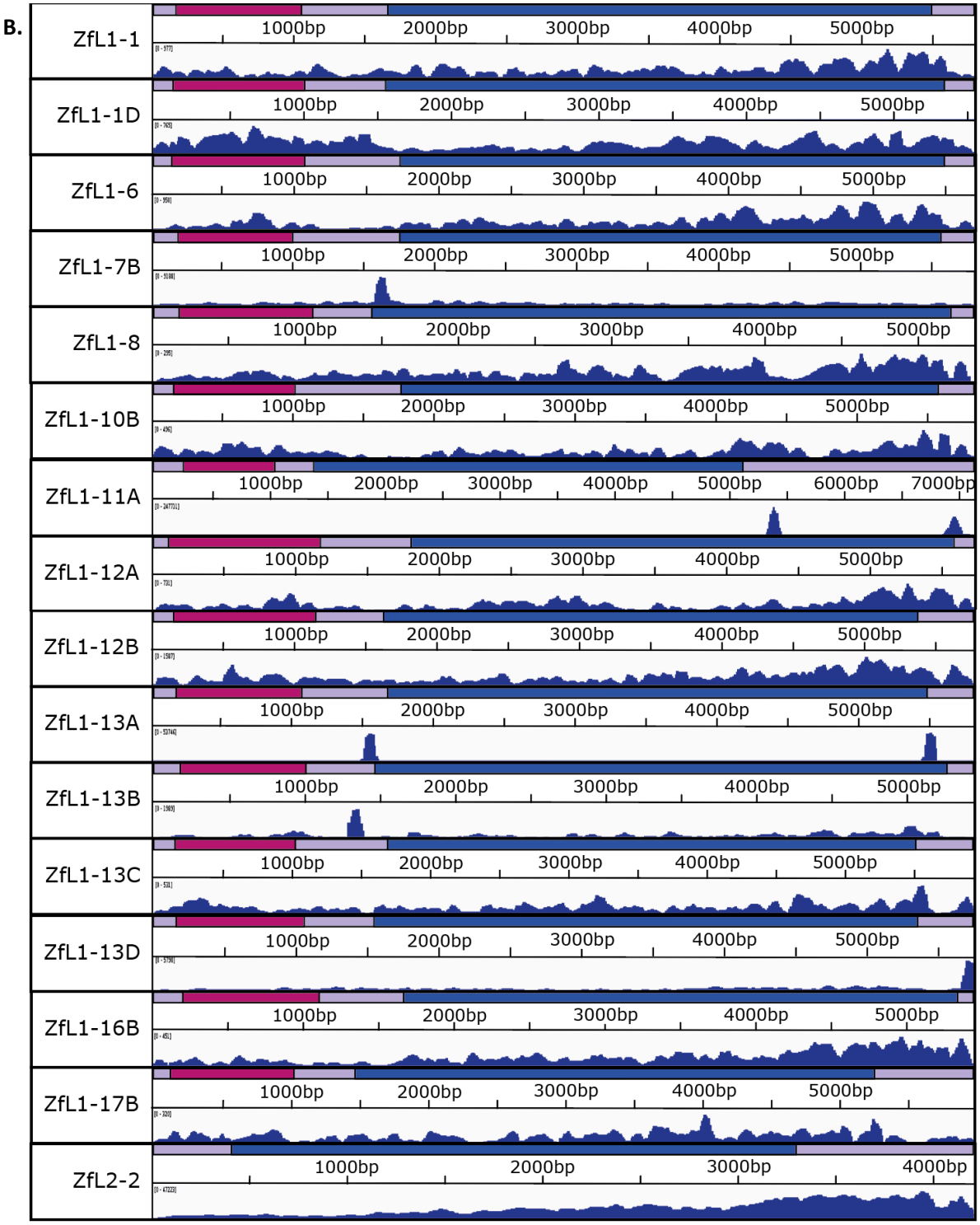


Figure 4.5. Zebrafish LINE expression along zebrafish development (full-length analysis) – Legend in the previous page.

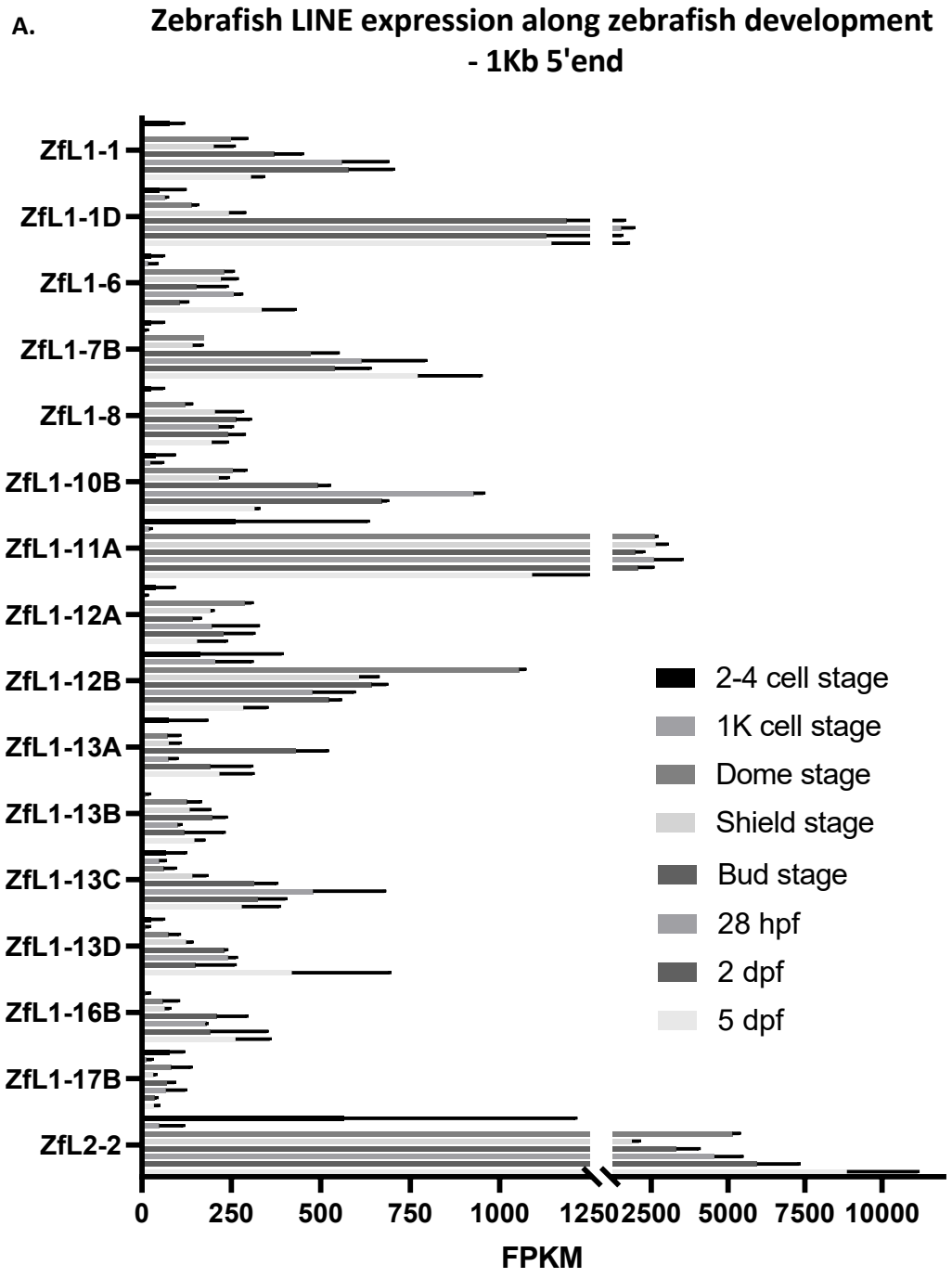


Figure 4.6. Zebrafish LINE expression along zebrafish development (1Kb 5'end) – (A) The graph shows the average FPKM mapped to the 1Kb 5'end of the 15 ZfL1 subfamilies [ZfL1-1, -1D, -6, -7B, -8, -10B, -11A, -12A, -12B, -13A, -13B, -13C, -13D, -16B and -17B] and the ZfL2-2 element in eight different zebrafish developmental stages [2-4 cell, 1K cell, dome, shield and bud stages, 28 hpf, 2 and 5 dpf], of two biological replicates. Error bars represent standard deviation between biological replicates. (B) [in the next page] A graphic representation of the sequence coverage. A schematic of the 15 ZfL1s and the ZfL2-2. Colour code: 5'UTR, IGR and 3'UTR, purple; ORF1, red; ORF2, blue. The length scale (bp) is indicated in the centre and the mapped sequence coverage is in the bottom, dark blue.

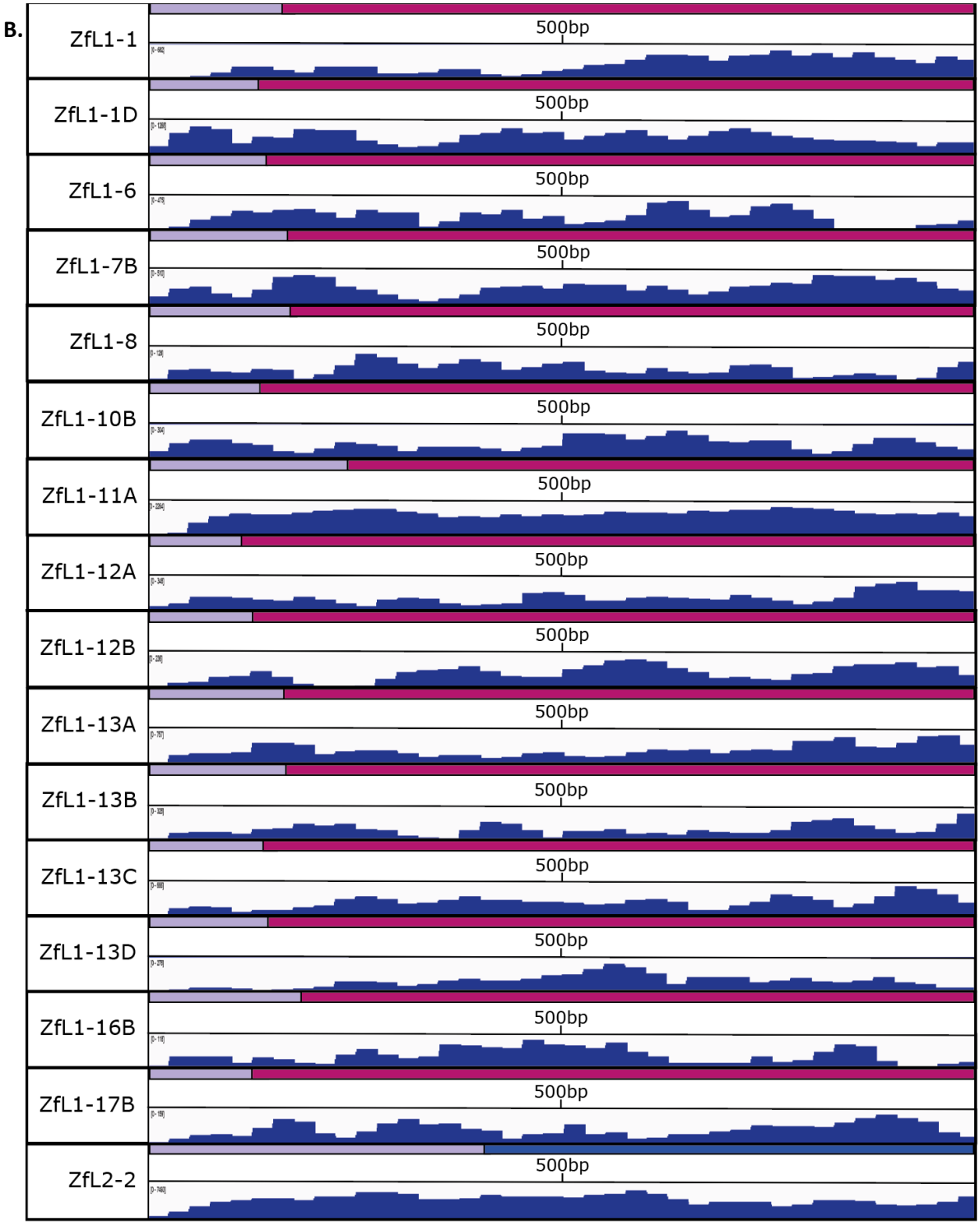


Figure 4.6. Zebrafish LINE expression along zebrafish development (1Kb 5'end) – Legend in the previous page.

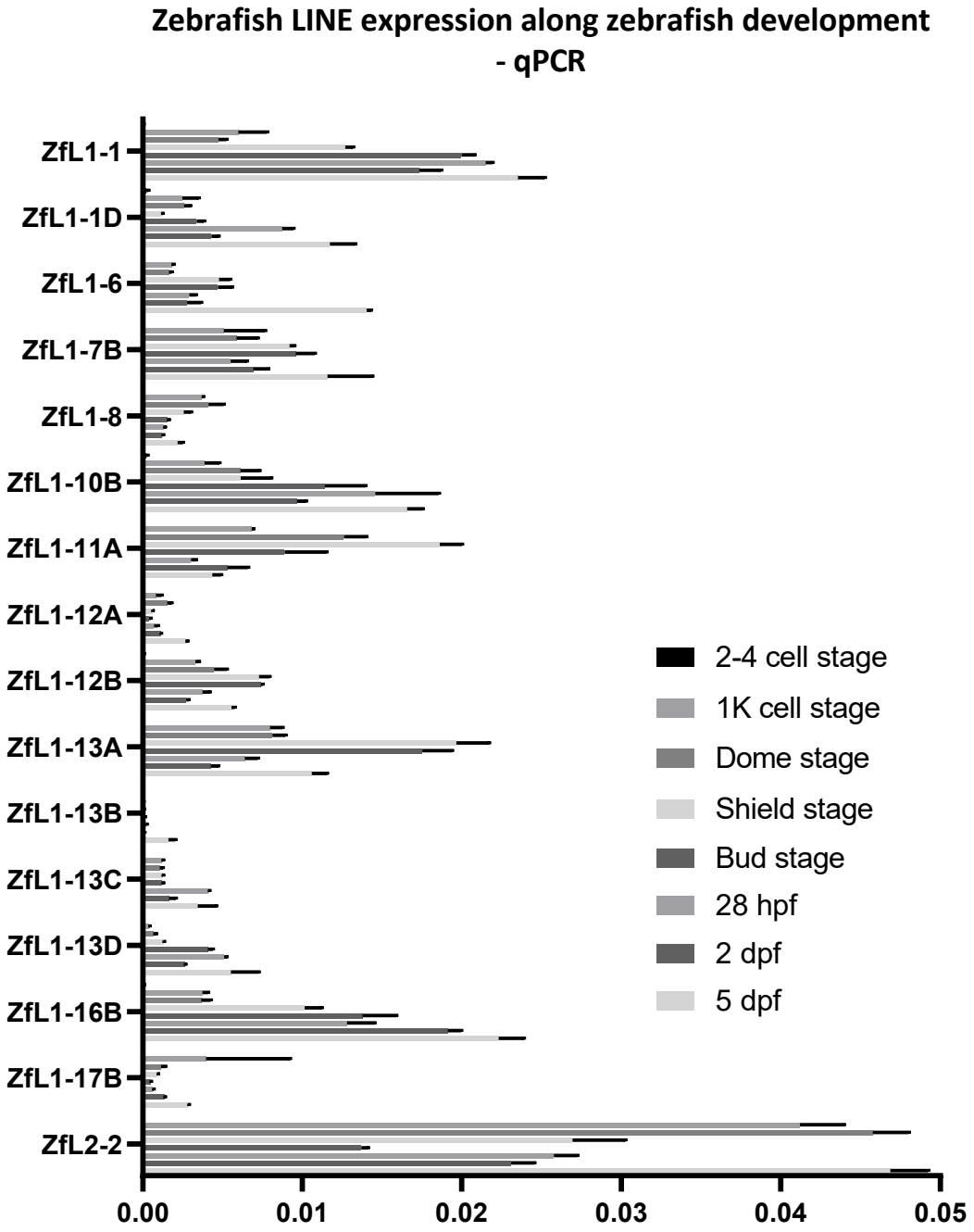


Figure 4.7. Zebrafish LINE expression along zebrafish development - qPCR – (A) The graph shows the average Fc to Act-β of the 15 ZfL1 subfamilies [ZfL1-1, -1D, -6, -7B, -8, -10B, -11A, -12A, -12B, -13A, -13B, -13C, -13D, -16B and -17B] and the ZfL2-2 element in eight different zebrafish developmental stages [2-4 cell, 1K cell, dome, shield and bud stages, 28 hpf, 2 and 5 dpf] as measured by qPCR. Error bars reflect standard deviation between three technical replicates.

4.5. Retrotransposition competence analyses of ZfL1 genomic copies from several subfamilies annotated in the zebrafish genome

To start exploring whether potentially active ZfL1s copies can be found the zebrafish genome, I selected a subset of ZfL1 subfamilies based on previous analyses: ZfL1-6, -7B, -10B, -11A, -12B, -13A and -13B (4.2). These seven ZfL1 subfamilies were selected because: a) high sequence identity; b) full-length copies with intact ORFs were annotated in the zebrafish genome reference (4.2.1); and c) all were expressed at various levels along zebrafish development (4.2.2).

In these selected subfamilies, the ZfL1 copies previously identified as potential retrotransposition competent in *in silico* analyses (4.2.1), were genotyped in the zebrafish colony housed at the IGC Zebrafish facility. The zebrafish TU strain was chosen, as this was the same strain used to assemble the zebrafish reference genome used to identify potential retrotransposition competent LINE-1 copies (GRCz10/danRer10). ZfL1 copies confirmed to be present in this strain, were amplified from the genome and cloned in vectors designed to follow retrotransposition *in vitro* and *in vivo*, using reporter based mobilization assays (***Table 2.4***: pMK013 - pMK057). In doing that, I was able to assess their activity in an exogenous context, using cultured human cells, as well as in an endogenous context, using single-cell fertilised embryos, allowing me to follow their integration during zebrafish development.

4.5.1. Isolation of potentially active LINE-1 copies from the zebrafish genome

I first determined whether the ZfL1 copies, identified in the zebrafish reference genome (4.2.1), were present in the zebrafish colony housed at the IGC Zebrafish facility using genotyping PCRs (2.2.1.2.2). Specific primers were designed to amplify the 5' and 3' junctions of these elements (2.2.1.2.1; ***Figure 4.8***; ***Table 6.4***). In the absence of the LINE-1 element, the primers specific to the flanking genome were used to amplify their chromosomal location (i.e., Empty site), allowing the determination of zygosity (***Figure 4.8***).

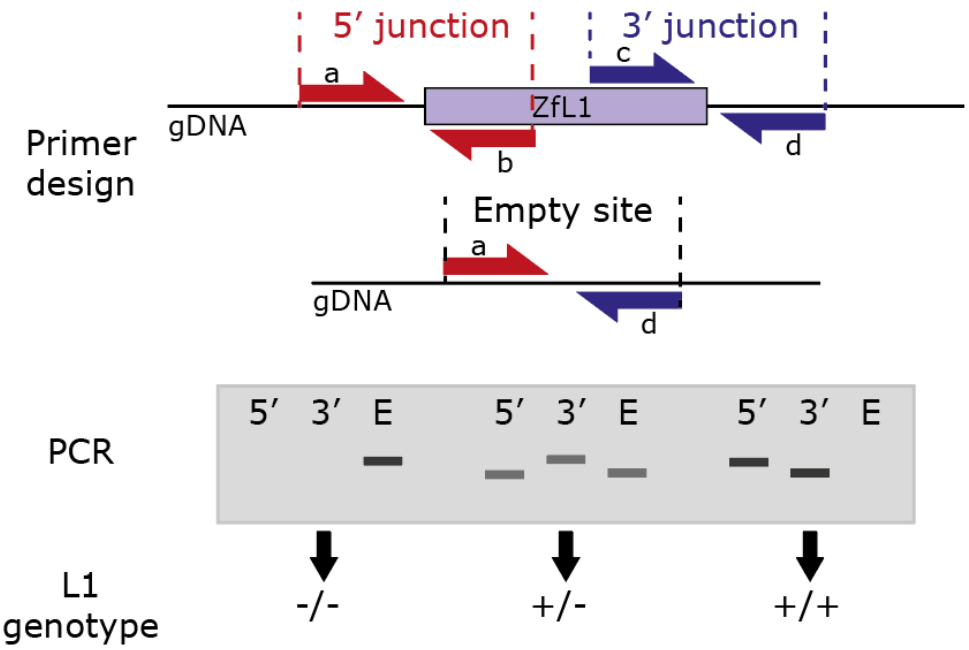


Figure 4.8. **Principle of genotyping PCR** – Schematic representation of the genotyping rationale, illustrating the relative position of primers (Primer design). Below is shown a virtual electrophoresis gel (PCR) with expected outcomes in function of their genotype (L1 genotype): homozygous absent (–/–), heterozygous (+/–) or homozygous present (+/+). Darker bands on the PCR gel indicate relative higher amplification efficiency, as the target is present twice.

Within the seven selected ZfL1 subfamilies, a total of 23 potentially active ZfL1 candidates were previously identified (4.2.1). However, the repetitive nature of flanking genomic sequences in two ZfL1-6 candidates precluded us from designing unique primers for genotyping (located on Chr16:20008305-20014003 and Chr19:43155801-43161493) (***Table 4.3***). To infer the allelic frequency of the remaining 21 potentially active ZfL1 candidates, I genotyped three fish of the TU strain housed at IGC. Due to the small colony size of this strain, these three fish were siblings. Genomic DNA from these three adult zebrafish individuals was extracted, and used as template in genotyping PCRs (2.1.6.3; 2.2.1.2.2). A non-template negative control was included with each primer pair, to control for potential cross-contamination. The PCR fragments were resolved using agarose electrophoresis and subjected to capillary DNA sequencing to confirm genotypes (2.2.1.5; 2.2.1.8; ***Figure 4.9***). The presence of a given ZfL1 candidate in the TU individuals was determined based on the validation of at least one of the junctions, 5′ or 3′.

Despite several attempts of optimisation, I failed to validate the 5′ junction of ZfL1-6 on Chr19:24218046-24223737 and ZfL1-7B on Chr4:45903129-45908934, as well as the 3′

junction of ZfL1-10B on Chr18:32301158-32307002 and ZfL1-13A on Chr7:1203268-1209061. As a result, these junctions were not sequenced by capillary DNA sequencing, and the exact sequence was not available to facilitate primer design for full length amplification of the element (see below). Additionally, I was unable to determine the status of ZfL1-13B on Chr16: 39601658-39607121 due to lack of amplification of the 5' and 3' junction, as well as of the empty site. Thus, in total I successfully genotyped 20 putative active ZfL1s in the TU colony available at the IGC, and calculated their allele frequency in this colony (*Table 4.3*). Of the 20 genotyped ZfL1 candidates, six were absent in the IGC TU zebrafish colony, while 11 were found in heterozygosity (at an allelic frequency of 16.7 – 83.3%), and three in homozygosity (*Table 4.3*; *Figure 4.9*). Thus, in total I was able to continue the work with 14 potentially active ZfL1s.

ZfL1 subfamily	Genomic location	Primers designed	Frequency
ZfL1-6	Chr6:21659365-21665057	Yes	16.7%
	Chr10:42609301-42614999	Yes	0%
	Chr11:7069121-7074820	Yes	50%
	Chr16:20008305-20014003	No	-
	Chr19:24218046-24223737	Yes	33.3%
	Chr19:43155801-43161493	No	-
ZfL1-7B	Chr4:45903129-45908934	Yes	33.3%
	Chr5:41057392-41063181	Yes	33.3%
	Chr6:18504003-18509793	Yes	16.7%
	Chr11:25254051-25259845	Yes	0%
	Chr24:31415464-31421258	Yes	83.3%
ZfL1-10B	Chr3:1795415-1801315	Yes	0%
	Chr5:42036670-42042519	Yes	16.7%
	Chr8:24020995-24026840	Yes	50%
	Chr14:37952310-37958210	Yes	0%
	Chr18:32301158-32307002	Yes	33.3%
ZfL1-11A	Chr3:868842-876012	Yes	66.7%
	Chr8:24262630-24269847	Yes	100%
ZfL1-12B	Chr19:12893436-12899204	Yes	100%
	Chr20:17401219-17406985	Yes	0%
ZfL1-13A	Chr2:7491273-7497084	Yes	0%
	Chr7:1203268-1209061	Yes	100%
ZfL1-13B	Chr16:39601658-39607121	Yes	?

Table 4.3. ZfL1 copies annotated in the zebrafish reference genome – ZfL1 subfamilies selected for further experimental analyses and their potentially retrotransposition competent copies identified previously. The table shows their genomic location, whether genotyping primers were successfully designed (Yes/No), and their allelic frequency inferred from the determined zygosity of three individuals from the TU zebrafish colony housed at the IGC Zebrafish facility.

Next, I designed a strategy to PCR amplify and clone the full-length sequence of the genotyped elements in pCEP4 and pTOL2 vectors, to test their retrotransposition potential *in vitro* and *in vivo* respectively (2.4.4.1; 2.4.4.2; ***Table 6.3***). Briefly, I used my own capillary DNA sequencing data, or data deposited at the UCSC browser, to design specific primers annealing at the 5' and 3' junction of each potentially active candidate ZfL1 genotyped (***Table 6.5***). Unique AscI and SrfI restriction sites were included in the 5' and 3' primers respectively, allowing me to clone these ZfL1s in the pCEP4 and pTOL2 vectors in an oriented manner.

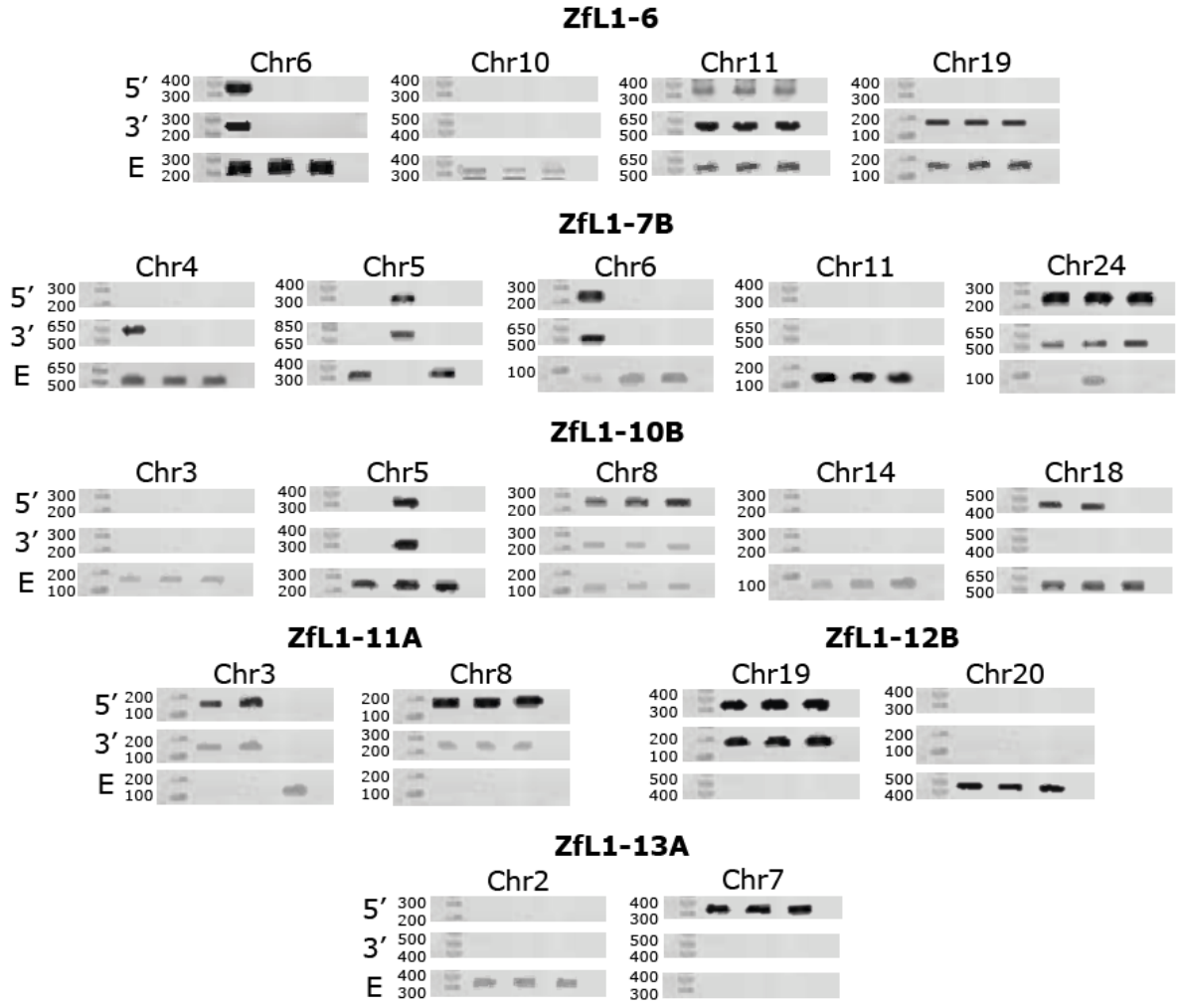


Figure 4.9. **ZfL1 genotyping PCRs** – Representative electrophoresis results from the indicated genotyping PCR of the different candidates of the six ZfL1 subfamilies [ZfL1-6, -7B, -10B, -11A, -12B, and -13A] (***Table 4.3***).

Of the 14 putative active ZfL1s confirmed to be present in the TU colony by genotyping PCR, despite several attempts using different primers, I failed to amplify the full-length sequence of five of the ZfL1s: ZfL1-6 Chr11:7069121-7074820 and Chr19:24218046-24223737,

ZfL1-11A Chr3:868842-876012 and Chr8:24262630-24269847, and ZfL1-13A Chr7:1203268-1209061. Notably, the 5' and 3' junctions of the elements ZfL1-6 Chr19:24218046-24223737 and ZfL1-13A Chr7:1203268-120906 respectively, were also not successfully amplified by genotyping PCR (***Figure 4.9***). This could indicate a discrepancy between the actual sequence of these regions and that annotated in the reference genome. Full-length amplicons of eight of the remaining nine elements, ZfL1-7B Chr4:45903129-45908934, Chr5:41057392-41063181, Chr6:18504003-18509793 and Chr24:31415464-31421258, ZfL1-10B Chr5:42036670-42042519, Chr8:24020995-24026840 and Chr18:32301158-32307002 and ZfL1-12B Chr19:12893436-12899204, were successfully obtained and fully resolved by capillary DNA sequencing. The sequences of these elements matched the sequence found in the zebrafish reference genome. For the element, ZfL1-6 Chr6:21659365-21665057, a PCR amplicon matching the expected size of the full-length element was obtained, but despite the junctions matching the reference genome annotated sequence, this element was never successfully cloned, nor was the internal sequence successfully resolved by capillary DNA sequencing. This element was therefore excluded from further investigation. Thus, the final ZfL1 candidates selected for further experimental analysis included eight copies, representing three ZfL1 subfamilies; four from the ZfL1-7B, three from the -10B and one from the -12B subfamily.

Further analyses of independent clones of PCR amplified ZfL1 elements revealed distinct mutations, suggesting they were generated during the PCR amplification step. To obtain clones with the exact sequence of the element in the zebrafish genome, I next used subcloning to remove PCR-induced mutations in these eight full-length amplified ZfL1 copies. Additionally, I removed the natural polyadenylation signal present in the 3'UTR of ZfL1 elements (mutating AATAAA to AAAAAA), as this could interfere with the retrotransposition assays used here (Belancio, Whelton, & Deininger, 2007) (***Figure 4.11***). I also used subcloning to generate additional ZfL1 elements, reconstruting the functional consensus sequence of the three subfamilies of which natural copies were successfully amplified (ZfL1-7B, -10B and -12B). In total, I generated 11 potentially active candidate ZfL1s to test for retrotransposition competence. These 11 elements were then cloned into pCEP4 and pTOL2 vectors, to test their *in vitro* and *in vivo* retrotransposition potential, respectively. Since the role of the ZfL1 3'UTR in the mobilisation of these elements is undetermined, I generated two versions of each pCEP4 and pTOL2 vector: a series where the retrotransposition indicator reporter was cloned

upstream of the ZfL1 3'UTR, and a second series containing the same reporter downstream of the 3'UTR (*Figure 4.10*).

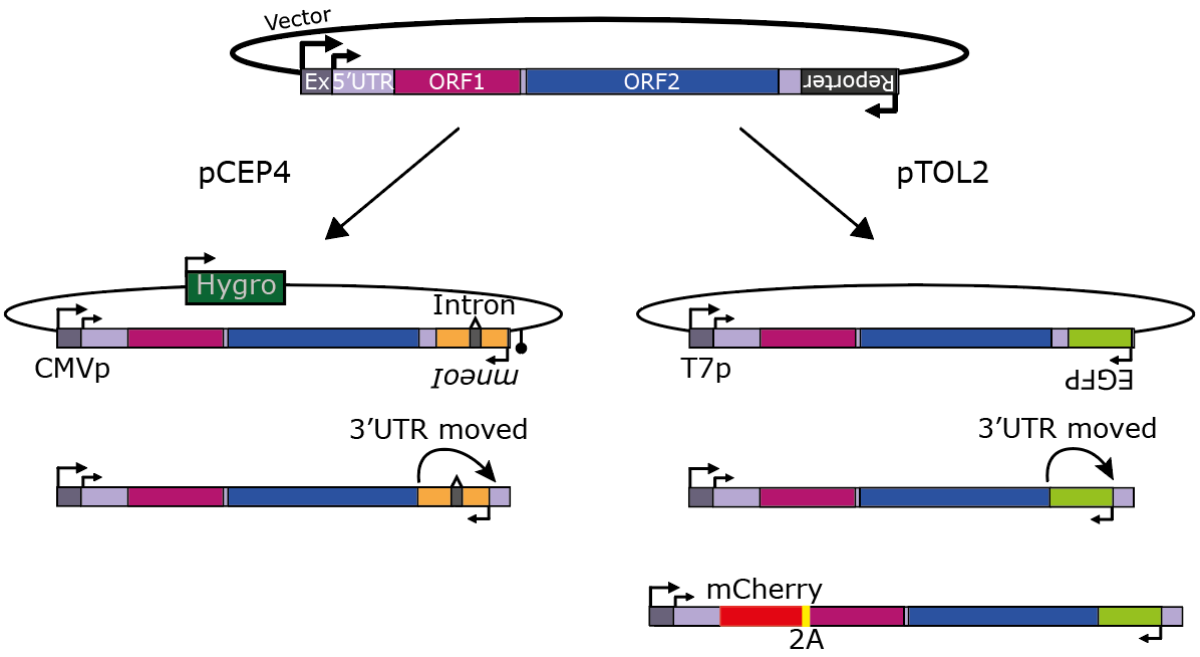


Figure 4.10. **ZfL1 retrotransposition assay vectors** – A schematic representation of a ZfL1 element in a retrotransposition reporter cassette vector (top), under control of an external promoter (Ex) and containing a reporter cassette at its 3'UTR in the anti-sense orientation (Reporter) under control of its own SA40 promoter (arrow). The ZfL1 pCEP4 vectors, containing a hygromycin selectable marker (Hygro), under control of an external human cytomegalovirus promoter (CMVp) and containing the Neomycin gene in anti-sense orientation as a retrotransposition reporter, interrupted by an intron in sense orientation (*mneoI*) ending in an SV40 late poly-adenylation signal (black lollipop) (left). The external CMVp is used to ensure transcription in human cells and to correct for expression differences among ZfL1 subfamilies. The late SV40 poly-adenylation signal mediates transcript maturation. The ZfL1 pTOL2 vectors, under control of an external T7 promoter (T7p) and containing the EGFP gene in anti-sense orientation as a retrotransposition reporter (right). The external T7p is used to *in-vitro* transcribe ZfL1 RNA from this plasmid.

Additionally, because prior to this thesis ZfL1 retrotransposition has never been tested, I created an additional series of plasmids that would allow me to detect ORF1p translation from the ZfL1 IVT RNA (i.e., using pTOL2 vectors). Briefly, these translation controls were constructed using the consensus sequence of the ZfL1-7B, -10B and -12B containing the retrotransposition indicator reporter upstream of their 3'UTR, and cloning a mCherry fluorescent marker that lacks a stop codon but contains a picornaviral 2A sequence (from the porcine teschovirus-1) (Kim et al., 2011), in-frame in the N-terminus of each ZfL1-ORF1p (***Figure 4.10; Table 2.4***: pMK058 - pMK060). The presence of the 2A sequence generates mCherry-2A

and ZfL1-ORF1p as independent proteins, allowing me to analyse the efficiency of ZfL1 translation (i.e., mCherry levels) without interfering with the enzymatic machinery encoded by ORF1p involved in retrotransposition. As an additional control, I also added the mCherry-2A reporter to the N-terminus of the single ORF of the zebrafish LINE-2 control ZfL2-2 (***Table 2.4***: pMK061, pMK062).

4.5.2. Retrotransposition potential of cloned ZfL1-7B, -10B and -12B candidates

The 47 constructed vectors (22 pCEP4, and 25 pTOL2) (4.3.1) were used to assess the retrotransposition activity of the different ZfL1 candidates (2.4.4), using *in vitro* and *in vivo* retrotransposition assays. The *in vitro* assays were performed using human HeLa cells, using DNA-based ZfL1 constructs (pCEP4 vectors) or IVT RNAs from ZfL1 constructs (pTOL2 vectors). In both assays, I exploited antisense retrotransposition indicator reporters that confer resistance to neomycin/G418 (pCEP4) or that activate expression of the fluorescent marker EGFP (pTOL2) (***Figure 4.11A***). The *in vivo* assays were performed using fertilised single-cell zebrafish embryos, in combination with IVT RNAs derived from ZfL1 constructs that were tagged with an antisense EGFP retrotransposition indicator reporter, activating expression of EGFP after a round of retrotransposition (pTOL2) (***Figure 4.11B***).

4.5.2.1. Results from *in vitro* retrotransposition assays

The 22 pCEP4-based vectors containing the different ZfL1 candidates (four elements from the ZfL1-7B subfamily, three from the -10B, one from the -12B subfamily, as well as the reconstructed consensus of these three subfamilies, containing the reporter up-/downstream of their 3´UTRs) were used to perform a colony retrotransposition assay in HeLa cells (2.4.4.4.1.1; ***Table 2.4***, pMK036 - pMK057). In parallel, HeLa cells were transfected with an active human LINE-1 also tagged with *mneoI* (L1.3-WT) or with an inactive allelic human LINE-1 containing a missense mutation in its RT domain (L1.3-RT^{-}), as positive and negative controls respectively (***Table 2.3***, JM101 and JM105). To increase the sensitivity of the retrotransposition assays, the transfected HeLa cells were first subjected to hygromycin selection (marker present in the backbone of pCEP4 vector; ***Figure 4.10***), to ensure the proportion of transfected HeLa cells was nearly 100%. This experiment was repeated three times with different batches of

purified vector DNAs (2.2.1.1). As expected, numerous G418-resistant foci were observed in HeLa cells transfected with the active human LINE-1 (WT), while no retrotransposition was detected when cells were transfected with the inactive RT-mutant human LINE-1 allele (RT^{-}) (*Figure 4.12*). None of the HeLa cells transfected with the ZfL1 candidates presented G418-resistant foci, regardless of the position of the reporter cassette up- or downstream of their 3'UTRs (*Figure 4.12*).

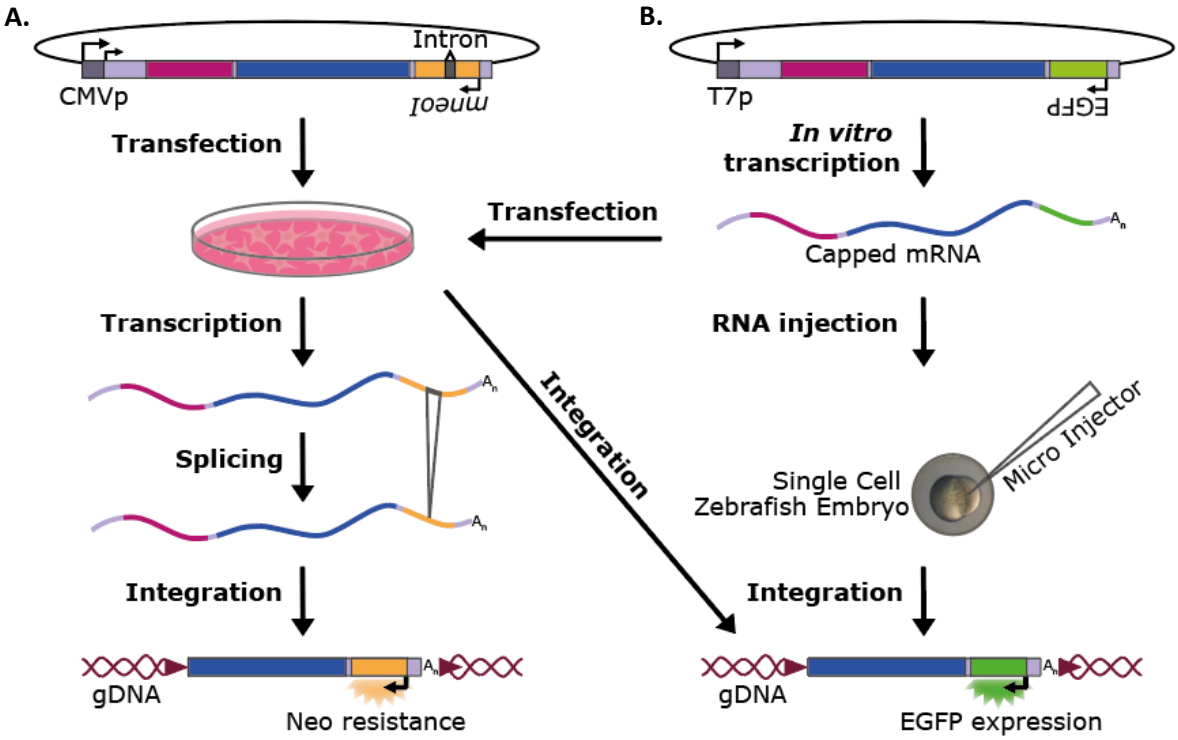


Figure 4.11. Rationale of ZfL1 retrotransposition assays – (A) Rationale of the LINE-1 retrotransposition assay using transfected DNA plasmids. reporter construct tagged with a reporter gene providing neomycin (Neo) resistance (yellow) in antisense orientation and interrupted by an intron in sense orientation (*mneoI*). The vector is transfected into cultured cells where the element, including reporter, is transcribed from the external CMVp or the internal endogenous ZfL1 promoter (arrows). When the retrotransposition intermediary RNA undergoes splicing, the intron interrupting the reporter gene is removed, leading to the generation of a functional reporter gene upon integration into the cellular genomic DNA (gDNA), allowing selection with neomycin/G418 for cells harbouring a retrotransposition event. (B) The working principle of the LINE-1 reporter construct tagged with an EGFP reporter gene (green) in antisense orientation. The vector is linearized and IVT from the external T7p. The resulting capped mRNAs are injected in single cell zebrafish embryos or transfected into cultured cells. Upon integration of the LINE-1 element into the gDNA, EGFP expression can be observed by microscopy or FACS, and the presence of the integrated EGFP gene can be confirmed by PCR.

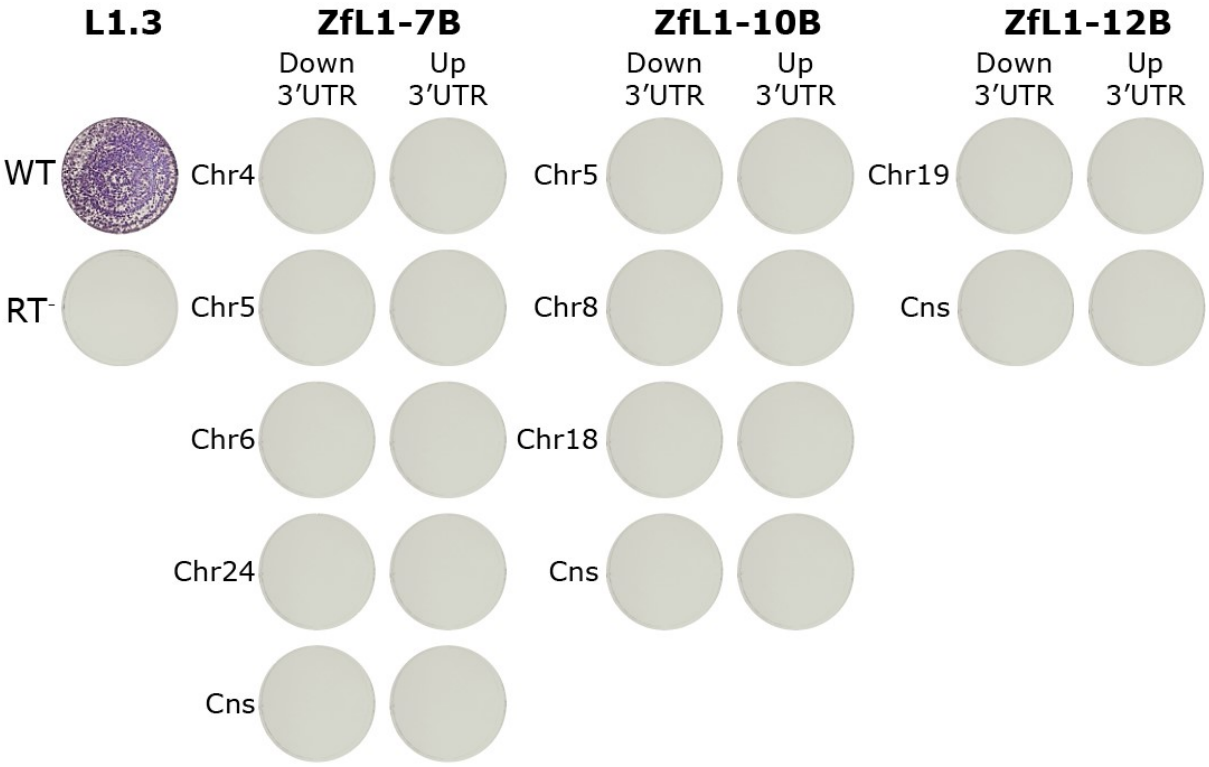


Figure 4.12. ZfL1 retrotransposition colony assay – Pictures show representative colony retrotransposition assay results of the 25 different ZfL1 copies tested. These ZfL1s belong to the ZfL1-7B (4 plus consensus), -10B (3 plus consensus) or -12B (1 plus consensus) subfamilies, and contained the retrotransposition indicator reporter downstream - (Down 3'UTR) or upstream (Up 3'UTR) of the L1 3'UTR. Results from positive and negative control transfected cells, human L1.3-WT and RT-, are shown in the left side. The chromosome where each ZfL1 copy was located is indicated left to each picture. The consensus version is labelled as Cns.

Because the sequence of LINE-1s are known to possess multiple SA and SD sites, the introduction of canonical SA and SD sites in the *mneoI* reporter cassette might lead to aberrant alternative splicing, producing dysfunctional LINE-1 RNAs (1.3.1). To rule out that splicing might interfere with our capability to detect retrotranspositon of ZfL1s in human HeLa cells, I performed *in vitro* retrotransposition assays using IVT RNA from the pTOL2 vectors, which use EGFP as a reporter and does not rely on splicing (***Figure 4.11***). Thus, IVT RNAs from the 22 pTOL2 vectors (four elements from the ZfL1-7B subfamily, three from the -10B, one from the -12B subfamily, as well as the reconstructed consensus of these three subfamilies, containing the reporter up-/downstream of their 3'UTRs) were transfected in HeLa cells (2.4.4.3; 2.4.4.4.1.2; ***Table 2.4***, pMK014 - pMK035). In parallel, HeLa cells were transfected with RNA derived from an active ZfL2-2, also tagged with EGFP (ZfL2-2 - WT) or with an inactive allelic ZFL2-2 RNA containing a missense mutation in its RT domain (ZfL2-2 - RT^{-}), acting as positive

and negative controls respectively (***Table 2.3***, ZfL2-2-WT; ***Table 2.4***, pMK013). EGFP-expressing cells were quantified using FACS, 24h and 48h after transfection. This experiment was repeated three times with different batches of IVT RNAs (2.4.4.3). As expected, EGFP-expressing cells were readily detected in cells transfected with IVT RNAs from ZfL2-2-WT at both time points examined and no EGFP-expressing cells were detected in cells transfected with the negative control (ZfL2-2-RT-) (***Figure 4.13***). However, no EGFP-expressing cells were detected in cells transfected with any of the 22 tested ZfL1 IVT RNAs, regardless of the position of the EGFP reporter cassette up- or downstream of the ZfL1 3'UTR (***Figure 4.13***). In sum, the 11 ZfL1 candidates analysed here do not retrotranspose at detectable levels in cultured human HeLa cells, under any of the experimental conditions used in this thesis.

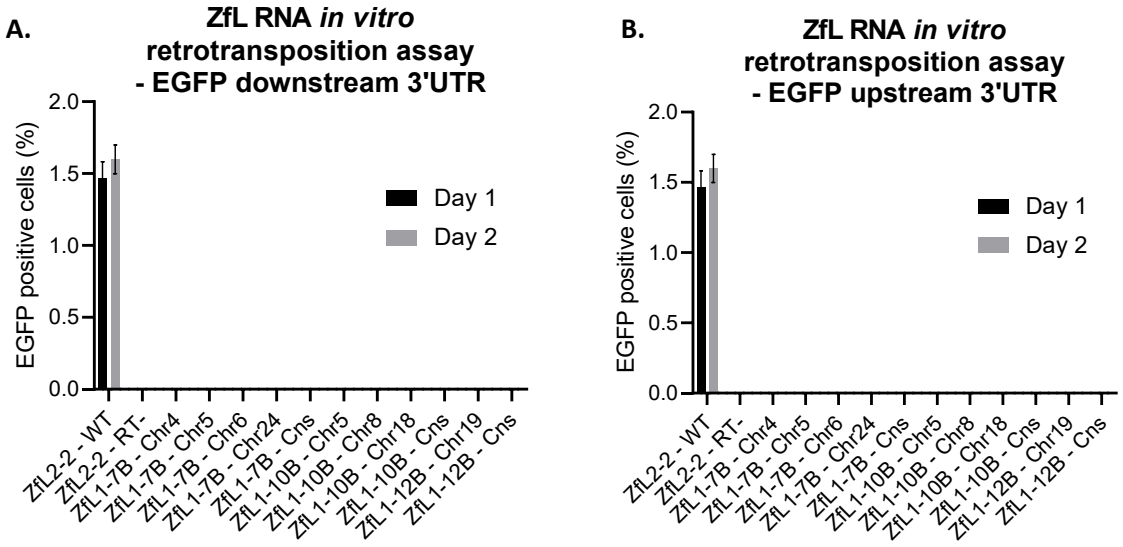


Figure 4.13. ZfL1 RNA *in vitro* retrotransposition assay – Each graph plots the percentage of EGFP-expressing cells detected 24h (Day 1) or 48h (Day 2) after transfection with the indicated antisense-EGFP tagged IVT LINE RNA. The results from constructs containing the EGFP reporter cassette (A) downstream and (B) upstream of the 3'UTR of ZfL1s; this doesn't apply to Zfl2-2s. Error bars represent the standard deviation of three biological replicates.

4.5.2.2. Results from *in vivo* retrotransposition assays

The *in vitro* retrotransposition assays revealed that, unlike the ZfL2-2, ZfL1s do not show retrotransposition activity in cultured human HeLa cells, under the experimental conditions tested in this thesis (4.3.2.1). To explore the activity of the different ZfL1 candidates in their natural context, I next analysed their retrotransposition potential in an *in vivo* retrotransposition assay. To do this, IVT RNAs from the 22 pTOL2 vectors (4 elements from the

ZfL1-7B subfamily, three from the -10B, one from the -12B subfamily, as well as the reconstructed consensus of these three subfamilies, containing the reporter up-/downstream of their 3´UTRs) were injected in single-cell fertilised zebrafish embryos (2.4.4.3; 2.4.4.4.2; *Table 2.4*, pMK014 - pMK035). The injected embryos were then monitored using a fluorescent stereomicroscope, to assess retrotransposition (EGFP expression) one, two and three days post injection (dpi). After the third analysis (3 dpi), genomic DNA from the injected larva was extracted and analysed by PCR using EGFP primers, to confirm the presence of integrated EGFP reporter copies.

This assay was previously optimised in our lab to assess the retrotransposition activity of the ZfL2 elements (Widmann et al., in preparation). Injecting 1 nL containing 0.05 ng of IVT ZfL2 RNA in the yolk of the fertilised single-cell zebrafish embryo consistently showed reproducible results in terms of embryo viability and ZfL2 retrotransposition activity (on average >85% of injected embryos had EGFP-expressing cells). To also optimise this assay for ZfL1 elements, allowing me to inject the maximum amount of IVT RNA (measured in ng) with minimal effect on embryo viability and morphology, I performed trial injections with different amounts of IVT RNAs from the ZfL1s and ZfL2-2. Specifically, I performed 1 nL injections containing 0.05 ng, 0.1 ng or 0.2 ng of IVT RNAs from the pTOL2 vectors containing ZfL2-2 - WT and ZfL1-7B, -10B and -12B (*Table 2.3*, ZfL2-2-WT; *Table 2.4*, pMK018, pMK022 and pMK024). The eggs of each clutch were divided over the different IVT RNAs to be injected, as well as the uninjected control. The survival rate of injected embryos was assessed 24 h post injection (hpi), and normalised to the survival rate of the uninjected embryos (deformed embryos were considered unviable). This experiment was repeated twice, with different batches of IVT RNAs, including 30 embryos per condition (2.4.4.3). High amounts of ZfL1 RNA injections were found to be less toxic to the developing zebrafish embryo than ZfL2-2 RNA injections, as indicated by their higher survival rate. Injection of 0.1 ng ZfL2-2 RNA resulted in a significant reduction of embryo survival rate (as expected based on previous experiments performed in our lab), while this amount was tolerated for ZfL1 RNAs (*Figure 4.14*). Injections of 0.2 ng ZfL1 RNA did result in significant decrease in embryo viability (*Figure 4.14*). Therefore, injections for the following experiments were performed, using 0.05 ng RNA injections for the ZfL2-2 element, and 0.1 ng of the ZfL1 elements, in a volume of 1 nL per injection. Notably, although injections with 0.05 ng of ZfL2-2-WT IVT RNA did not significantly affect embryo viability, embryos showed a delay

in development compared to those injected with ZfL1 IVT RNAs and uninjected (***Figure 4.15***; ***Figure 4.16***).

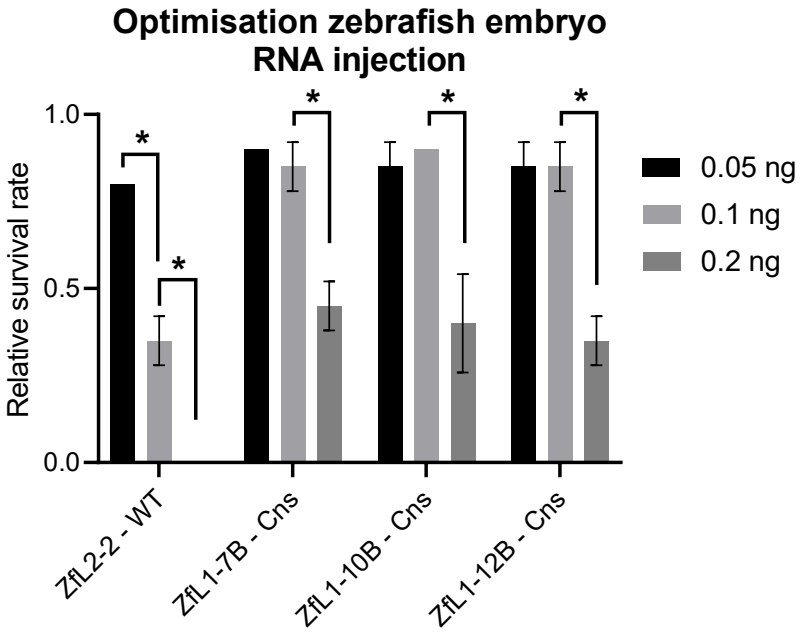


Figure 4.14. **Optimisation of zebrafish embryo RNA injection** – The graph plots the survival rate of zebrafish embryos injected with 0.05, 0.1 and 0.2 ng IVT RNAs from ZfL2-2 - WT, ZfL1-7B - Cns, ZfL1-10B - Cns and ZfL1-12B - Cns, relative to uninjected embryos of the same clutch. N = 30 for each RNA injection per experiment. Error bars represent the standard deviation of two independent experiments. Data was compared using a multiple unpaired *t*-test (0.05 ng *vs* 0.1 ng, and 0.1 ng *vs* 0.2 ng; * = p-value < 0.05).

As optimised, the *in vivo* retrotransposition assay was performed by injecting 1 nL containing 0.05 ng IVT ZfL2-2-WT RNA, or 0.1 ng IVT ZfL1 RNA in the yolk of fertilised single-cell zebrafish embryos (***Table 2.3***, ZfL2-2-WT; ***Table 2.4***, pMK018, pMK022 and pMK024). Not all ZfL1 RNAs were injected on the same day, as the amount of available embryos was dependent on the spawning behaviour and efficiency of the zebrafish mating pairs. However, in general, from each clutch 10% of embryos were injected with the positive control, ZfL2-2-WT, 80% were injected with ZfL1 RNAs, while 10% remained uninjected, acting as negative controls. This experiment was repeated three times using different batches of IVT RNAs, including approximately 50 embryos per RNA (2.4.4.3). As expected, we readily detected EGFP signal in embryos injected with ZfL2-2-WT, at the three time points analysed, without antibody staining (on average, >85% of injected embryos showed EGFP signal) (***Figure 4.15***; ***Figure 4.16***). In contrast, no EGFP signal was detected in uninjected embryos (***Figure 4.15***; ***Figure 4.16***). The

embryos injected with IVT ZfL1 RNAs also showed no EGFP signal in any of the days analysed, regardless of the location of the antisense EGFP in their 3'UTRs (*Figure 4.15*; *Figure 4.16*).

Although previous optimisation in our lab showed that the most consistent results, in terms of embryo viability and ZfL2 retrotransposition, were obtained when IVT RNA injections took place in the yolk, it is generally accepted that RNAs are most active when injected directly into the embryonic cell. Since ZfL1 IVT RNAs are larger than the ZfL2-2 IVT RNAs (~8Kb vs ~5Kb), which could potentially difficult RNA migration towards the developing embryonic cell, I decided to also inject ZfL1 IVT RNAs in the cytoplasm of the embryonic cell. This experiment was repeated two times, using different batches of IVT RNAs, including about 30 embryos per RNA (2.4.4.3). However, as with injections in the yolk, embryos injected with the ZfL1 IVT RNAs in the cytoplasm of the cell did not show any EGFP signal, regardless of the location of the antisense EGFP in their 3'UTRs (data not shown).

To confirm the microscopy results, I isolated genomic DNAs from the injected larva at 3 dpi, to genotype the presence of integrated EGFP (i.e., inserted as a result of retrotransposition) (2.2.1.6.3; 2.2.1.2.2). As expected, PCR analyses confirmed the presence of EGFP in gDNAs of embryos injected with the positive control ZfL2-2 IVT RNA (ZfL2-2-WT RNA injected embryos), while no amplification was detected for uninjected control embryos. There was also a lack of EGFP amplification for ZfL1 IVT RNA injected fish, regardless of the position of the EGFP reporter up- or downstream of the 3'UTR. These results strongly suggest that retrotransposition of the 11 ZfL1 candidates tested here does not occur, or is very low in their natural context (below the detection limit of the in vivo retrotransposition assay used here), at least during early embryogenesis and with the experimental set-up used in this thesis.

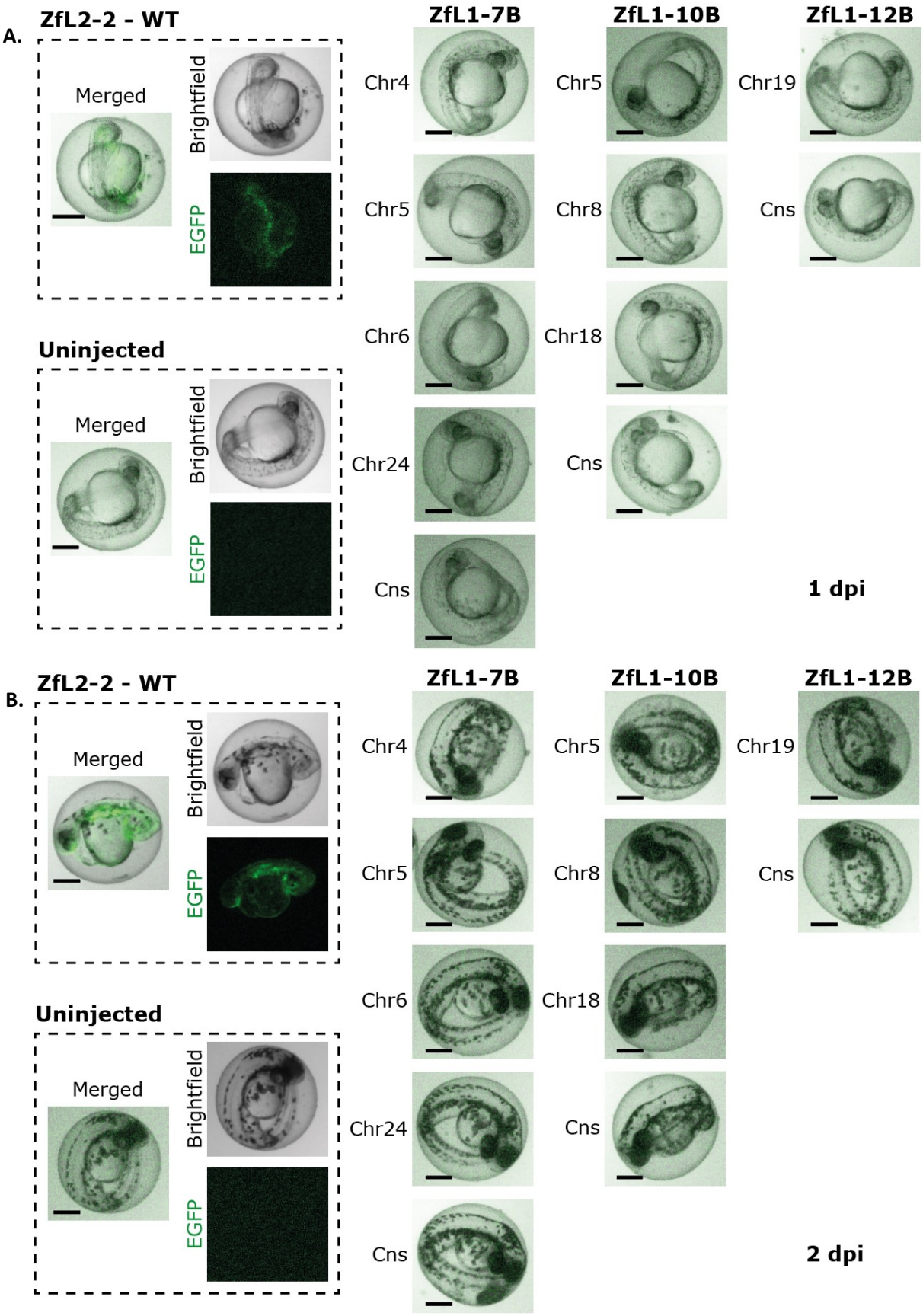


Figure 4.15. ZfL1 *in vivo* retrotransposition assay, EGFP report downstream of 3'UTR – Legend in the next page.

C.

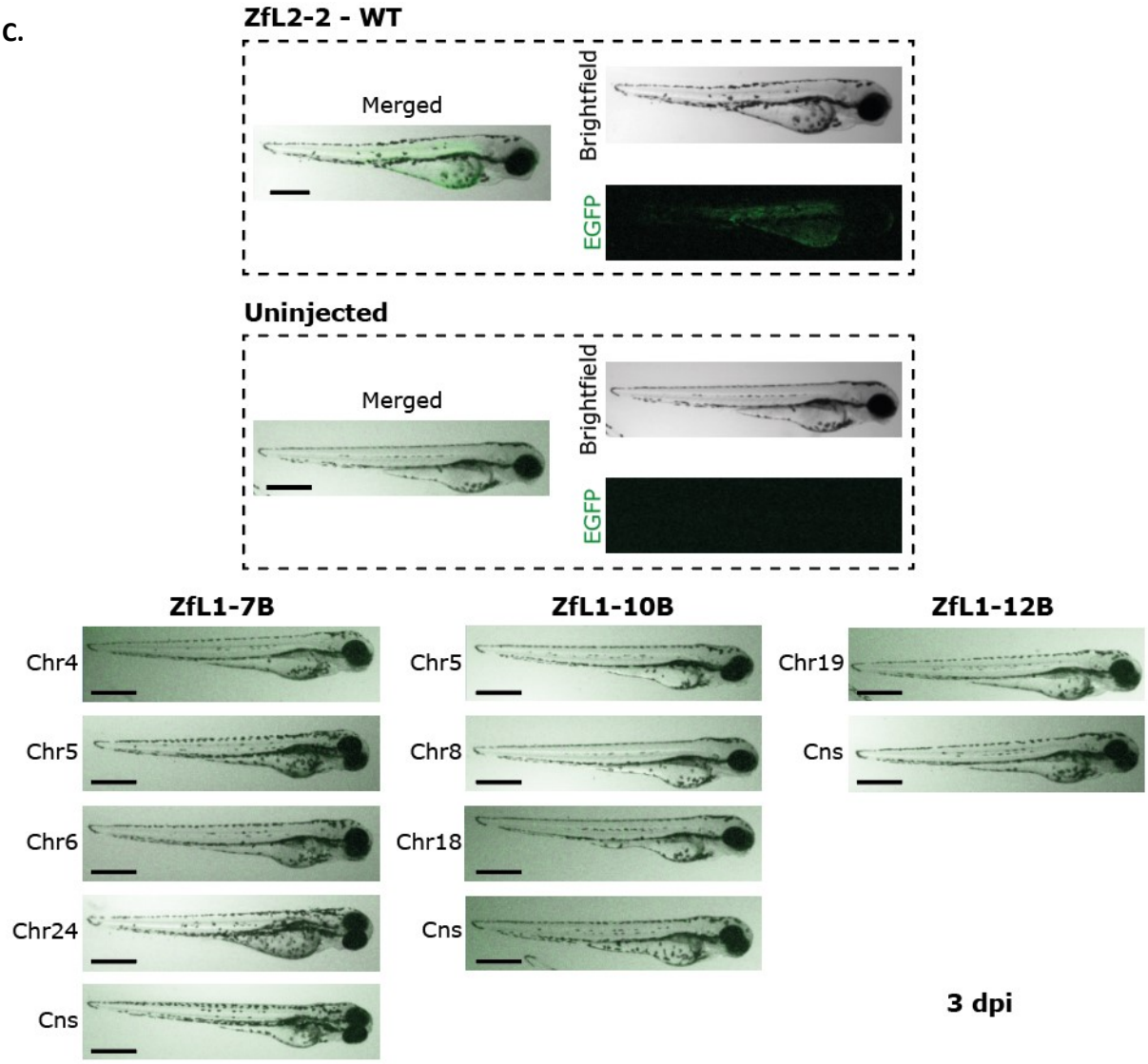


Figure 4.15. ZfL1 *in vivo* retrotransposition assay, EGFP report downstream of 3'UTR – [A and B in the previous page] Each picture shows representative data from retrotransposition assays conducted with the indicated ZfL1 element and ZfL2-2-WT (positive control), as well as uninjected embryos (negative control). The images are the merge of the EGFP channel with a brightfield image of the ZfL1 IVT RNA injected embryos. The EGFP channel and brightfield capture, as well as the merged image is shown for the positive and negative control as a reference. Images were captured 1 dpi (A), 2 dpi (B) and 3 dpi (C). Scale bars: 250 µm.

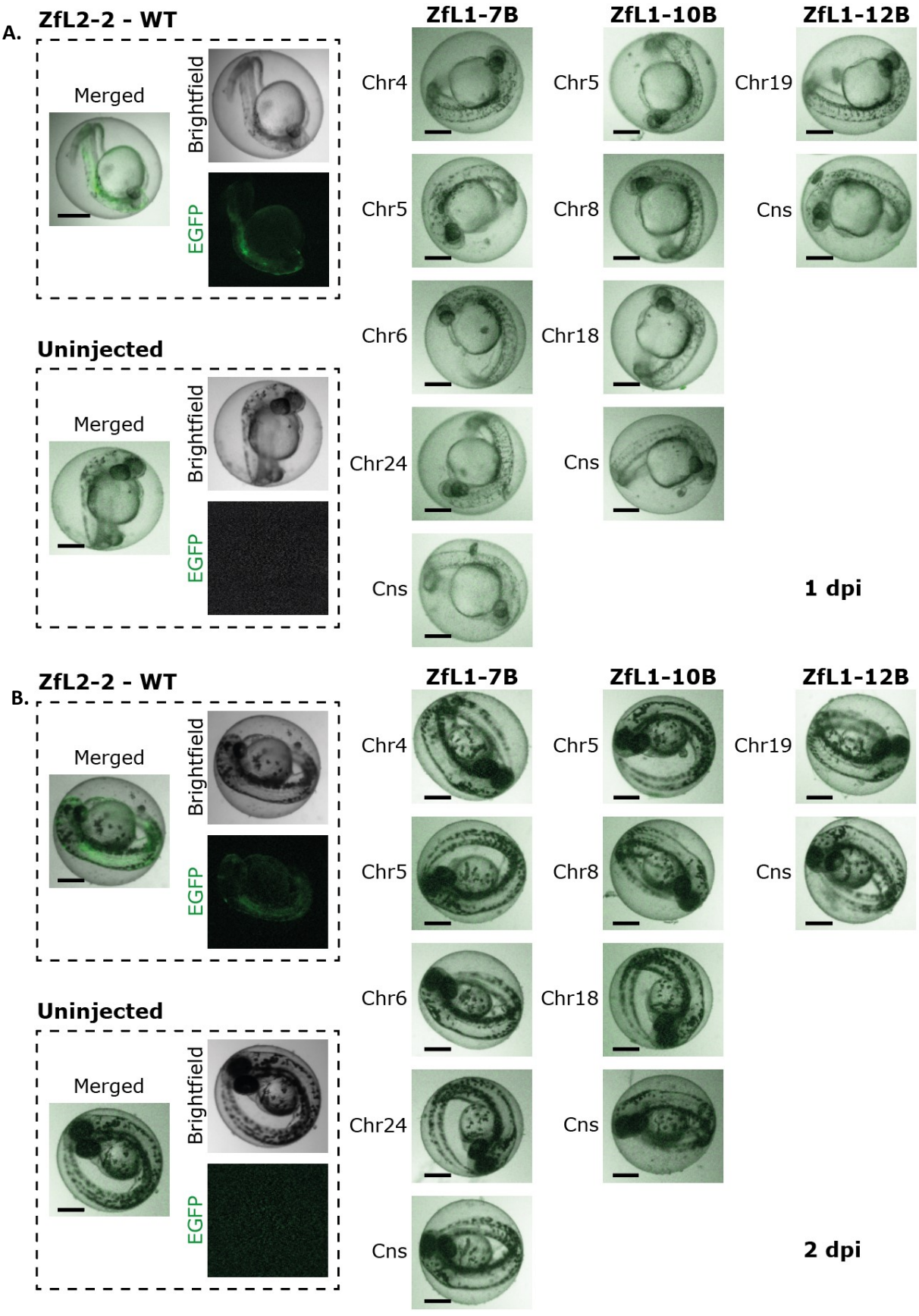


Figure 4.16. ZfL1 *in vivo* retrotransposition assay, EGFP report upstream of 3'UTR – Legend in the next page.

C.

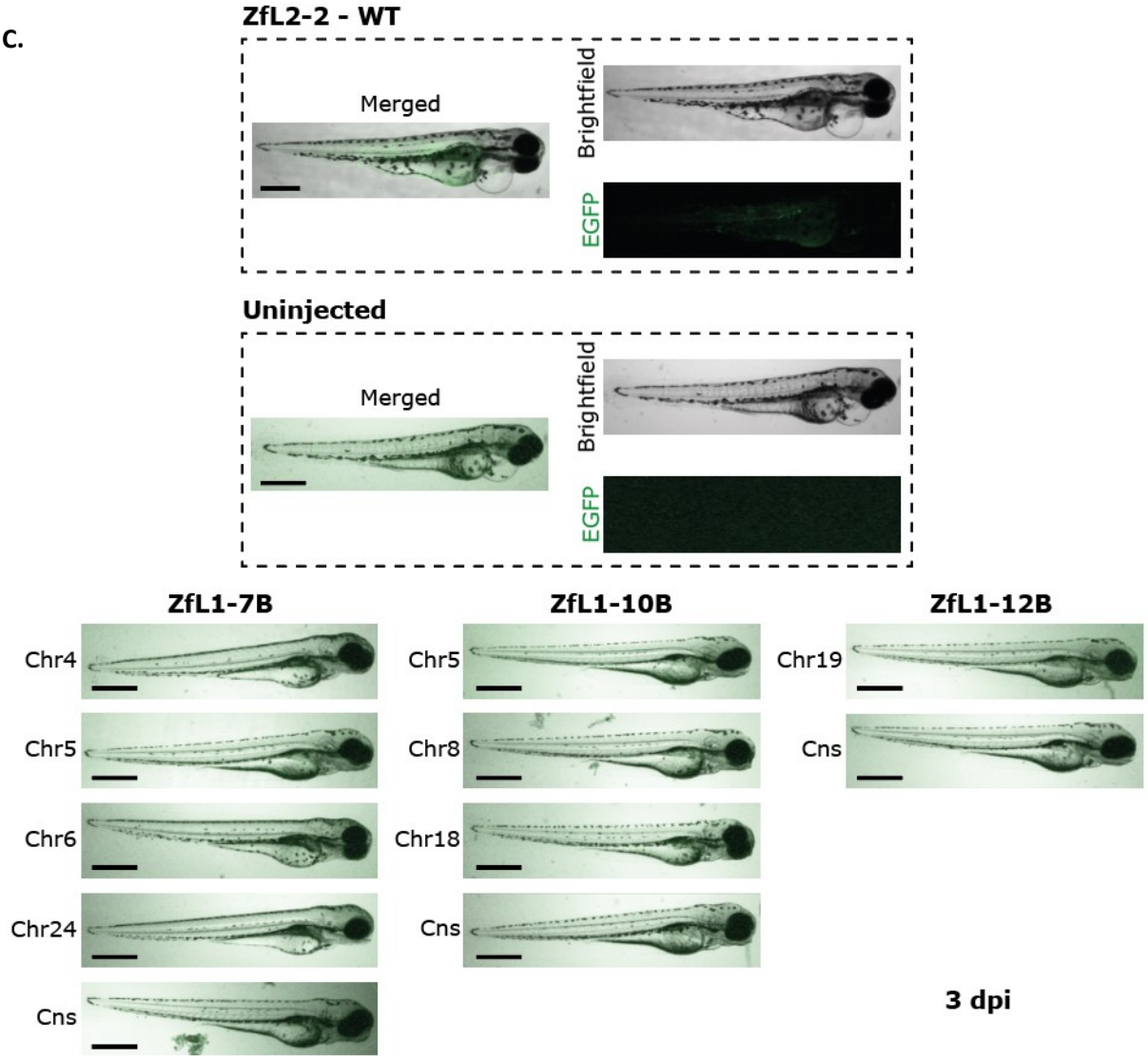


Figure 4.16. ZfL1 *in vivo* retrotransposition assay, EGFP report upstream of 3'UTR – [A and B in the previous page] Each picture shows representative data from retrotransposition assays conducted with the indicated ZfL1 element and ZfL2-2-WT (positive control), as well as uninjected embryos (negative control). The images are the merge of the EGFP channel with a brightfield image of the ZfL1 IVT RNA injected embryos. The EGFP channel and brightfield capture, as well as the merged image is shown for the positive and negative control as a reference. Images were captured 1 dpi (A), 2 dpi (B) and 3 dpi (C). Scale bars: 250 µm.

4.5.3. Efficient translation of the ZfL1-ORF1p from IVT RNAs from the ZfL1 subfamilies ZfL1-7B, 10B and 12B

To analyse whether lack of *in vivo* ZfL1 retrotransposition activity is related to poor translation of injected IVT RNAs, I constructed the pTOL2 vectors containing the ZfL1 consensus candidates with a mCherry gene fused to the N-terminus of their ZfL1-ORF1, using a self-processing 2A sequence between both ORFs (***Table 2.4***, pMK058 - pMK060). As positive and negative controls, I fused the same mCherry-2A reporter to the N-terminus of the single ORF of the Zfl2-2-WT and ZfL2-2-RT^{-} (***Table 2.4***, pMK061 and pMK062). The ORF1 from ZfL1s, as well as the single ORF from ZfL2-2s, is expected to be translated by a canonical cap-dependent scanning mechanism; thus mCherry expression would act as a reporter of ZfL1-ORF1p translation (and of ORF-Zfl2-2 translation) (Alisch et al., 2006). To note, the porcine teschovirus-1 2A autoproteolytic peptide used here has been shown to be functional in zebrafish, and allows the synthesis of mCherry and a second ORF as independent polypeptides from the same RNA (Kim et al., 2011). IVT RNAs were injected into the yolk of single-cell zebrafish embryos, as described previously for the *in vivo* retrotransposition assays (2.4.4.4.2; 4.3.2.2), and the injected embryos were then monitored using a fluorescent stereomicroscope. This allowed me to simultaneously explore IVT RNA translation (mCherry expression) and retrotransposition (EGFP expression) *in vivo*, 6 and 24 hpi (***Figure 4.17***). A limitation from these analyses is that they are qualitative in nature, not quantitative; in fact, there are several variables that can influence the intensity of Cherry/EGFP, such as the number of IVT RNA copies injected (RNA was quantified in ng, not taking into account size differences). Additionally, several aspects of the RNA injection are hard to control, such as the exact injection site of the yolk in relation to the embryonic cell, as well as differences in injected RNA arising during the injection process as a consequence of organic matter building up in the needle with each injection. This experiment was repeated two times using different batches of IVT RNA, including about 30 embryos per RNA (2.4.4.3).

I consistently found that at 6 hpi, mCherry is detectable under the microscope, without antibody staining, in the developing embryos injected with the ZfL1-7B, -10B and -12B consensus IVT RNA (on average, >90% of injected embryos showed mCherry signal), but not for those injected with ZfL2-2-WT, or uninjected embryos (***Figure 4.17A***). Consistent with

previous observations from our lab, no detectable EGFP signal was observed in the embryos injected with ZfL2-2-WT IVT RNA at 6 hpi. As expected based on the previous results in this thesis, no EGFP signal was detected in the embryos injected with ZfL1 IVT RNAs, or uninjected embryos. Interestingly, 24 hpi the mCherry protein can still be observed under the microscope in the developing embryos injected with ZfL1-7B, -10B and -12B consensus RNA (on average, >90% of injected embryos showed mCherry signal), but no EGFP signal was detected (***Figure 4.17B***). Surprisingly, while no detectable mCherry signal was observed in embryos injected with mCherry fused ZfL2-2-WT IVT RNAs, I readily detected EGFP-expressing cells (on average, >85% of injected embryos showed EGFP signal), indicative of Zfl2-2 retrotransposition, further suggesting that Zfl2-2 RNAs were translated. As expected, no mCherry or EGFP signal was detected on uninjected embryos

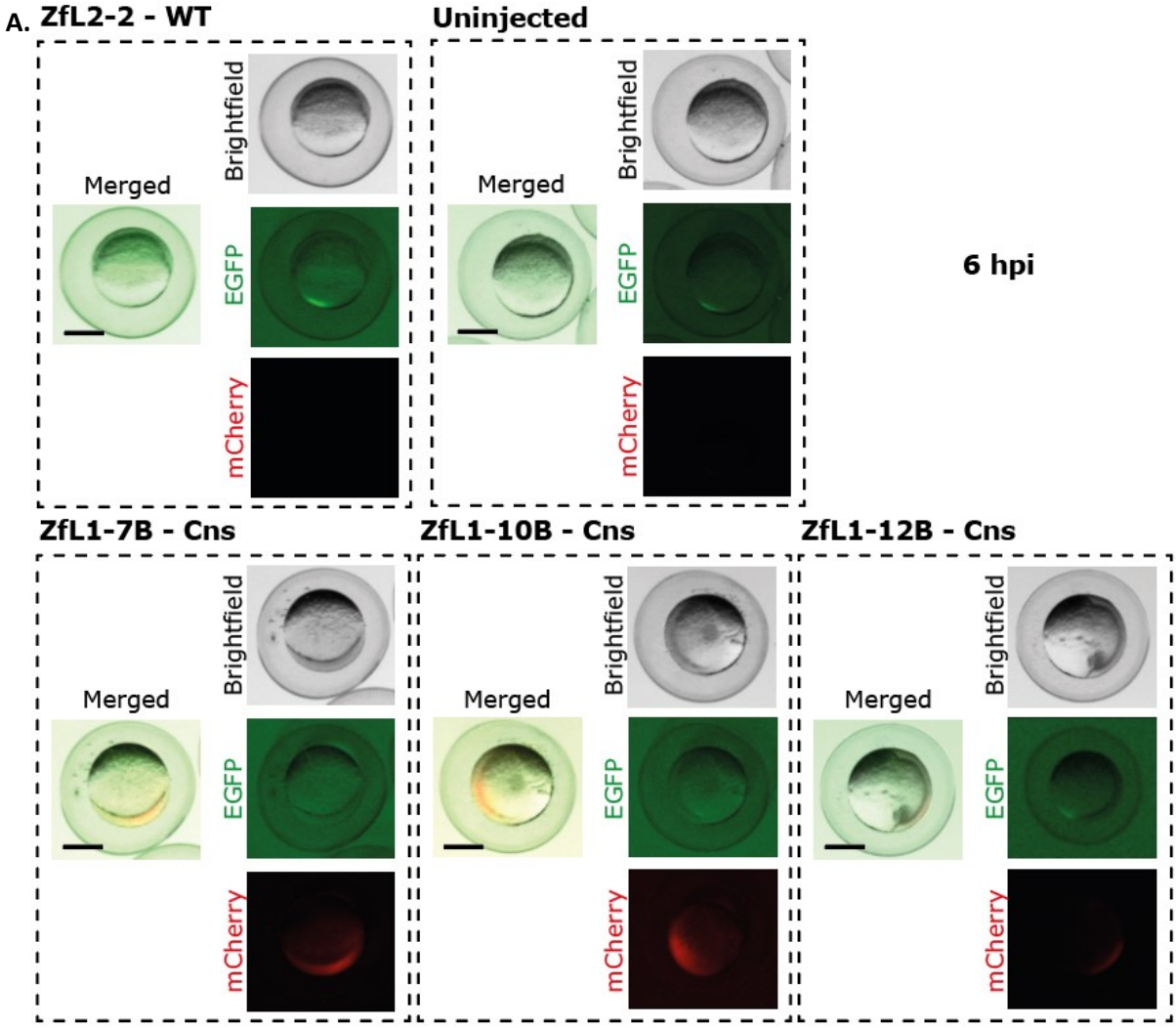


Figure 4.17. Translation of IVT RNA *in vivo* – Legend in the next page.

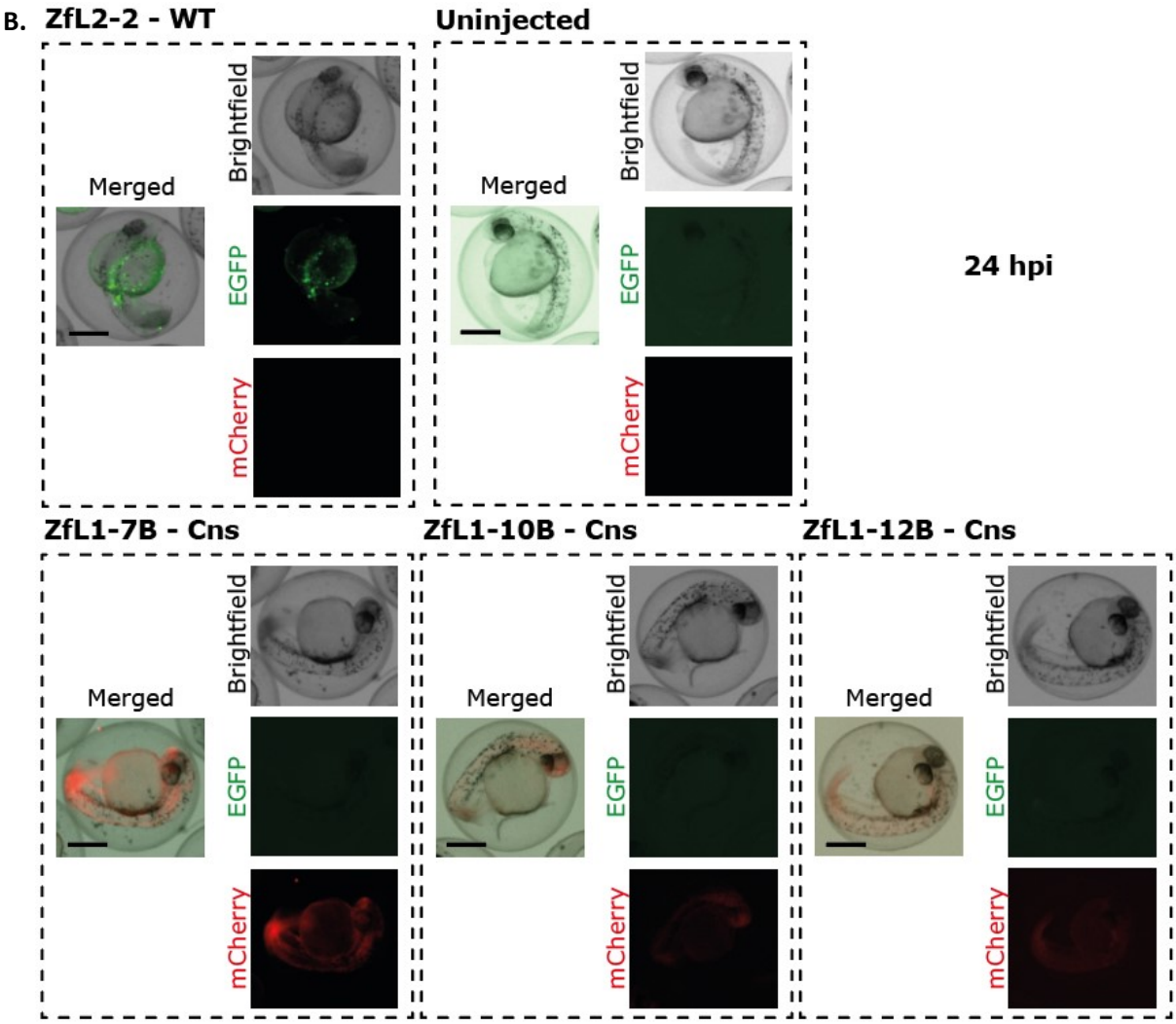


Figure 4.17. Translation of IVT RNA *in vivo* – [A in the previous page]. Each picture shows representative data from microscopy/retrotransposition assays conducted with the indicated ZfL1 element and ZfL2-2-WT (positive control), as well as uninjected embryos (negative control). The images show the EGFP channel and brightfield capture, as well as the merged image of the positive and negative control, as well as the different ZfL1 elements, ZfL1-7B, 10B and 12B consensus (Cns). Images were captured (A) 6 hpi and (B) 24 hpi. Scale bars: 250 µm.

To support the above findings in a different cellular context, and to better quantify zebrafish LINE translation from IVT RNAs, I transfected these mCherry-tagged IVT RNAs in HeLa cells, and analysed mCherry and EGFP expression levels using FACS, as described previously for the *in vitro* retrotransposition assays (2.2.3.4; 2.4.4.4.1.2; 4.3.2.1). mCherry fused ZfL2-2-RT^{-} IVT RNA was included as a negative control for retrotransposition. Transfected HeLa cells were analysed one and two days post transfection, and the percentage mCherry and EGFP expressing cells was determined (***Figure 4.18***). Additionally, the intensity of EGFP and mCherry

was also scored (***Figure 4.19***). This experiment was repeated two times using different batches of IVT RNAs. Consistent with the *in vivo* data, we found that cells transfected with the three tested ZfL1 constructs (ZfL1-7B, -10B or -12B consensus), showed mCherry expression one and two days post transfection (***Figure 4.18A***), but no EGFP-expressing cells were detected (***Figure 4.18B***). Unlike in the *in vivo* assays, HeLa cells transfected with mCherry-ZfL2-2-WT IVT RNAs showed significant mCherry expression one and two days post transfection (***Figure 4.18A***), as well as EGFP expression (***Figure 4.18B***). Using a Turkey's multiple comparison test to compare the % mCherry expressing cells detected in the cells transfected with the different IVT RNA showed no significant difference. There was however an overall significant decrease in the % mCherry expressing cells detected two days post transfection compared to one day post transfection, as indicated by a 2-tailed paired t-test ($p < 0.01$).

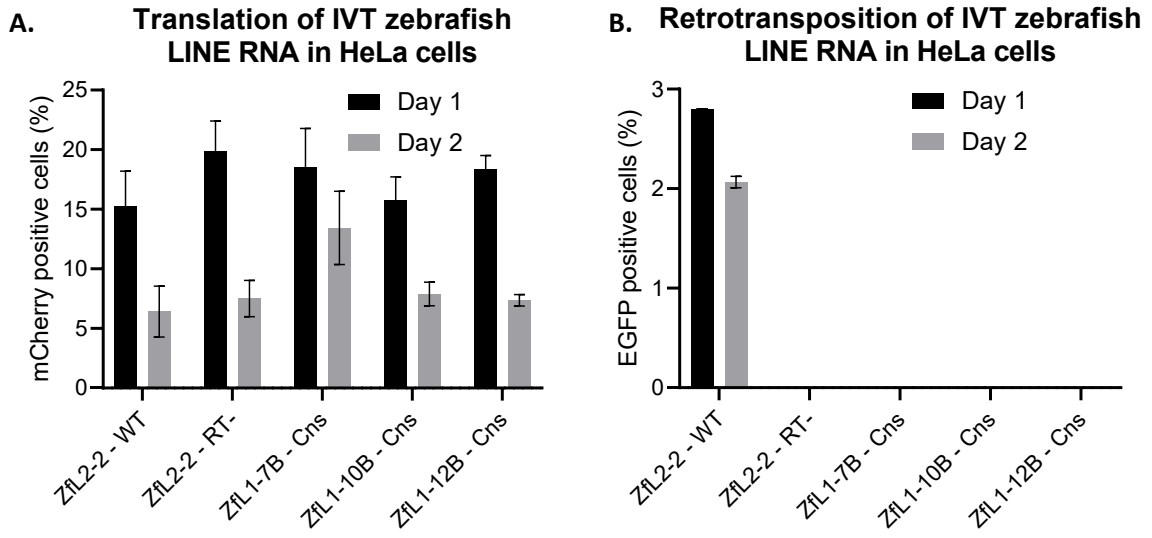


Figure 4.18. Translation and retrotransposition of IVT zebrafish LINE RNA in HeLa cells – The graphs show FACS analyses of the percentage of (A) mCherry and (B) EGFP expressing cells in transfected HeLa cells, one (Day 1) and two days (Day 2) post transfection. Cells were transfected with IVT RNAs from ZfL2-2-WT and ZfL2-2-RT^{-}, as well as ZfL1-7B, -10B and -12B consensus, all tagged with mCherry-2A fused to the N-terminus of the single ORF of ZfL2-2 or to the N-terminus of ZfL1-ORF1p. Error bars represent the standard deviation of two biological replicates.

Next, I analysed the mean fluorescent intensity of the mCherry-expressing cells for each IVT RNA (***Figure 4.19***). Using a Turkey's multiple comparison test to compare the mean fluorescent intensity of mCherry measured in the mCherry positive cells transfected with the different IVT RNAs, indicated a significantly stronger mCherry signal in the cells transfected with IVT RNAs from ZfL1-7B and -10B compared to IVT RNAs from the ZfL2-2-WT ($p < 0.05$).

Additionally, there was an overall increase in mCherry signal intensity from day 1 to day 2 post transfection, as indicated by a 2-tailed paired t-test ($p < 0.05$). These data, together with the previous result showing that there is no difference in % mCherry positive cells between the different IVT RNAs, suggests that the lack of detected mCherry signal *in vivo* (i.e., zebrafish injected with the ZfL2-2-WT IVT RNA) (***Figure 4.17***) is due to less efficient translation leading to a weaker mCherry signal. Since the translation of the ZfL1-ORF1p and ZfL2-2-ORFp should occur by analogous cap-dependent scanning mechanisms, these were unexpected findings. However, there is a possibility that potential differences in RNA stability of the different constructs influenced these results, which was not controlled for in this experimental setup.

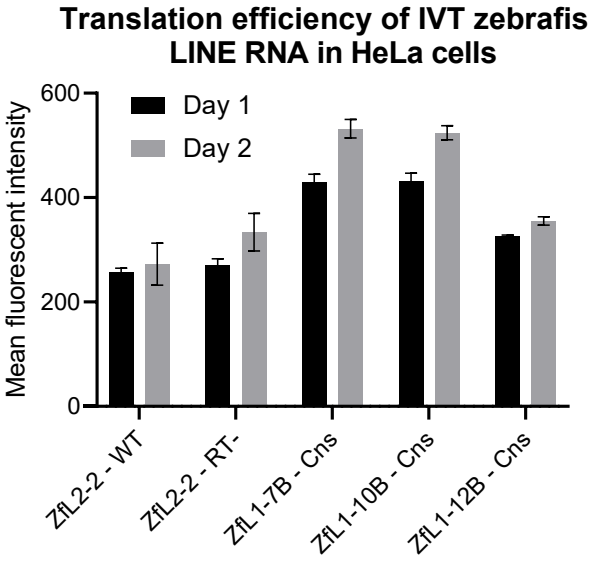


Figure 4.19. Translation efficiency of IVT zebrafish LINE RNA in HeLa cells – The graph shows the FACS analysis of the mean fluorescent intensity of mCherry expressing HeLa cells one (Day 1) and two days (Day 2) post transfection. Cells were transfected with IVT RNAs from ZfL2-2-WT and ZfL2-2-RT^{-}, as well as ZfL1-7B, -10B and -12B consensus, all tagged with mCherry-2A, fused to the N-terminus of the single ORF of ZfL2-2 or fused in the N-terminus of ZfL1-ORF1p. Error bars represent the standard deviation of two biological replicates.

This prompted me to take a closer look at the mechanism underlying the less efficient translation of ZfL2-2 RNA. To identify which characteristic of the ZfL2-2-ORFp that might be the cause of the decreased translation, I build a series of ZfL2-2 constructs containing debilitating mutations in; i) the EN (ZfL2-2 - EN-); ii) and RT (ZfL2-2 - RT-) domains; iii) two conserved tail-binding regions (TBRs) located in the ORF (ZfL2-2 - TBR2m); iiii) and the conserved stem-loop found in the 3'UTR (ZfL2-2 - WT - NSL) (***Table 2.4***, pMK062 - pMK065). Additionally I generated a construct containing a STOP codon following the 2A sequence, preventing translation of the

single ORF (ZfL2-2 - STOP) (*Table 2.4*, pMK066). As above, I produced IVT RNAs from these constructs, and transfected them in HeLa cells. the percentage and intensity of mCherry was analysed by FACS one and two days post transfection, and these data was used as a surrogate of translation efficiency, as described in previous in vitro retrotransposition assays (2.2.3.4; 2.4.4.3; 2.4.4.4.1.2).

A Turkey's multiple comparison test to compare % mCherry positive cells showed a significantly higher level in the IVT ZfL2-2 - STOP RNA transfected cells compared to the ZfL2-2 - WT one day after transfection ($p < 0.01$) (*Figure 4.20A*). Comparing the mean fluorescent intensity of mCherry of the HeLa cells transfected with the different IVT RNAs using the same test, showed the levels were significantly higher in the HeLa cells transfected with the IVT ZfL2-2 - STOP RNA compared to all other IVT RNAs, both one and two days post transfection ($p < 0.05$) (*Figure 4.20B*). As mentioned for the previous experiment, there is a possibility that potential differences in RNA stability of the different constructs influenced these results, which was not controlled for in this experimental setup.

A.

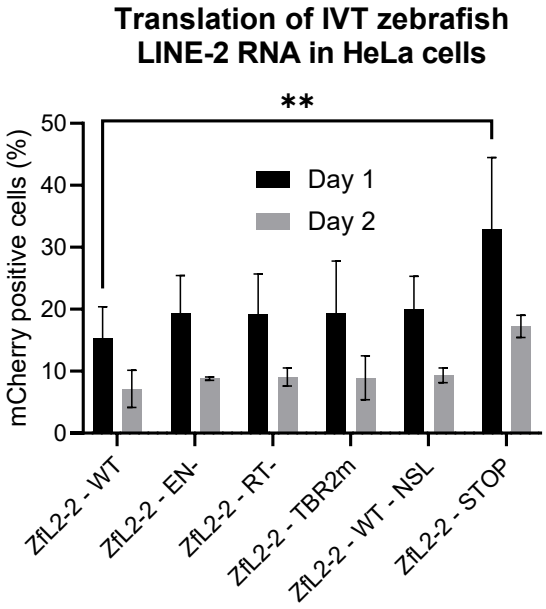


Figure 4.20. Translation efficiency of IVT zebrafish mutant LINE RNAs in HeLa cells – The graph shows the FACS analysis of the (A) % mCherry positive cells, and (B) [in the next page] mCherry mean fluorescence one (Day 1) and two days (Day 2) post transfection. HeLa cells were transfected with IVT RNAs from ZfL2-2-WT, ZfL2-2-EN^{-}, ZfL2-2-RT^{-}, ZfL2-2-TBR2m, ZfL2-2-WT-NSL and ZfL2-2-STOP all tagged with mCherry-2A, fused to the N-terminus of the single ORF of ZfL2-2. Error bars represent the standard deviation of two biological replicates. Turkey's multiple comparison test was used to compare the different transfected HeLa cells one day (Day 1) and two days (Day 2) post transfection (* = p-value < 0.05; ** = p-value < 0.01).

B.

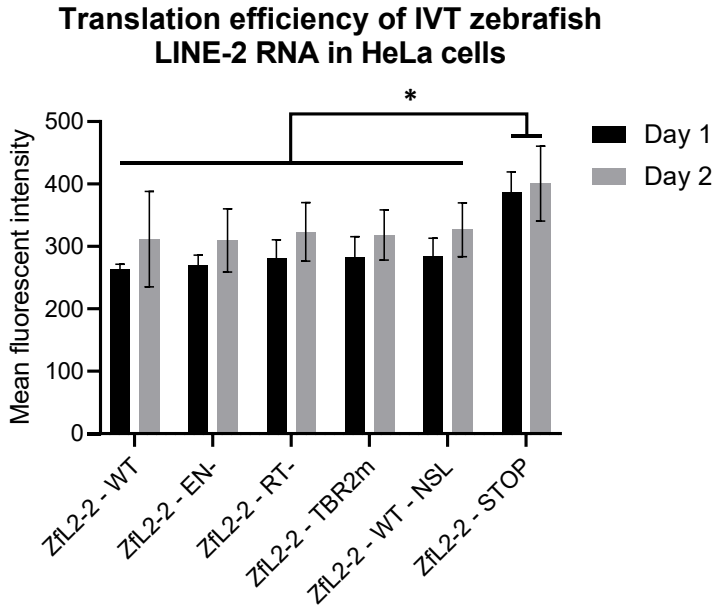


Figure 4.20. Translation efficiency of IVT zebrafish LINE RNAs in HeLa cells – Legend in the previous page.

Discussion

Animal models have been of great value to help elucidate specific cellular niches and contexts in which LINE-1s are active (1.4). The opportunity to study endogenous LINEs in animal models that recapitulate the behaviour and function of human LINE-1s, significantly adds to its value, as there is no need to include artificial/engineered sequences, potentially biasing LINE-1 models and conclusions. Mice have proven to be one of these valuable animal models, even though not all aspects of mouse LINE-1 biology match that of the human LINE-1s, and they can be time consuming and costly to work with (4.1.1). This section of my thesis aims to analyse 17 potentially active LINE-1 subfamilies present in the zebrafish genome (4.1.3), in order to validate this animal model (worth due to its easy manipulation and associated to lower costs) for studying LINE-1 biology, like its contribution to human pathologies. Prior to this thesis, the retrotransposition competence and expression of these 17 zebrafish LINE-1 subfamilies has not been explored, and remained unconfirmed. Here I aim to answer these questions through a series of *in silico*, *in vitro* and *in vivo* experiments. In this chapter, I will discuss the outcomes of these experiments to address the aims and objectives set out (4.1.4).

4.6. *In silico* analysis of the sequence and expression of the ZfL1 subfamilies

To determine whether any of the 17 annotated zebrafish LINE-1 subfamilies have the potential to be currently active in the zebrafish genome, I first performed an *in silico* analysis to make an informed decision about which subfamilies to select for experimental validation. I first examined the full-length copies present in the zebrafish reference genome, identifying the presence of intact ORFs and the hallmarks of canonical retrotranspostion (4.2.1); then, I analysed the conserved domains of the polypeptides encoded in the consensus sequence of each subfamily, built from the intact full-length copies previously identified (4.2.1); and finally, I assessed the expression levels of the different subfamilies during early embryogenesis/gastrulation (4.2.2). ZfL1-7C and -1B were excluded as potential candidates due to lack of copies containing intact ORFs and the absence of a poly-A tract at their 3'ends respectively (***Table 4.1***). I narrowed down the remaining 15 subfamilies using a more detailed analysis of the amino acid sequence of their ORFps (ORF1p and ORF2p; ***Figure 4.3***; ***Figure 4.4***), as well as considering their expression profile *in vivo* (***Figure 4.5***; ***Figure 4.7***), although this proved complicated for various reasons. The amino acid sequence of ORF1p and ORF2p encoded in the different subfamilies were similarly conserved, with few non-conserved substitutions spread across the different domains. Since there was no clear way to predict the severity and functional impact of the different substitutions, exclusion based on this analysis was not straight forward. Regarding the expression data, initially only the analysis of the RNA-seq data using the full-length ZfL1 sequence was performed. This meant there was an increased likelihood of including reads originating from partial ZfL1 sequences present in other cellular RNAs, as well as the phenomenon later observed when the mapping of the reads was visualised (read mapping to repetitive regions) (***Figure 4.5B***). Based on the available data the seven subfamilies, ZfL1-6, -7B, -10B, -11A, -12B, -13A and -13B, were chosen to include subfamilies varying in the conserved domains containing non-compatible substitutions (***Figure 4.3***; ***Figure 4.4***; ***Table 4.2***), and detected expression levels (***Figure 4.5***; ***Figure 4.7***). However, upon re-analysis of the RNA-seq data using only the first 1Kb of each element subfamily, the ZfL1-1D appeared as one of the highest expressed subfamilies (***Figure 4.6A***). Since this subfamily is represented by four full-length copies containing intact ORFs in the reference genome (***Table 4.1***) and showed a high conservation in the encoded polypeptide motifs in its

consensus sequence (***Figure 4.3***; ***Figure 4.4***; ***Table 4.2***), it could indicate a strong promising candidate for future research.

4.7. Genomic copies from ZfL1-7B, -10B and -12B, and their consensus sequences, show no significant retrotransposition activity *in vitro* or *in vivo*

Of the seven subfamilies chosen based on the *in silico* analysis, I successfully isolated eight full-length copies representing three subfamilies. These copies, including the consensus sequences of the three subfamilies were tested in an *in vitro* and *in vivo* retrotransposition assay, using transfection of HeLa cells (***Figure 4.12***; ***Figure 4.13***) and zebrafish embryo injections (***Figure 4.15***; ***Figure 4.16***) respectively. Importantly, in these assays, the reporter cassette was assayed both before and after the 3'UTR of the element. An important caveat of the engineered LINE-1 retrotransposition assay, *in vitro* and *in vivo*, is the position of the antisense reporter marker within the 3´UTR of the LINE-1. In fact, while initial studies demonstrated that the 3´UTR of human LINE-1s is not strictly required for retrotransposition in HeLa (Moran et al., 1996), recent studies suggest it might have a role on the efficiency of the retrotransposition process (Sahakyan et al., 2017). Additionally, recognition of the 3'UTR by the LINE-1 machinery before the reporter cassette could result in lack of integration of the reporter in the genome, leading to an inability to detect retrotransposition events. Work in our lab with the stringent-type ZfL2-2 showed that the presence of the ORFp recognition signal (the 3'UTR stem-loop) before the reporter cassette precludes the reporter from integration (data not shown). Because the role of the 3´UTR of ZfL1s is currently unknown, I decided to generate two allelic vector series, which contain the antisense reporter marker before and after the 3'UTR of each ZfL1 element. However, none of the ZfL1 copies, nor their artificially generated consensus sequences, showed retrotransposition activity in any of the experimental settings tested in this work.

The fact that the ZfL1 subfamilies selected for experimental validation contained non-compatible substitutions in different conserved domains of their amino acid sequence allows speculation about the importance of these domains for ZfL1 functionality. ZfL1-7B and -10B subfamilies contain a non-conserved substitution in the CTD of their ORF1p (R262S and R261Q respectively; ***Figure 4.3B***) in a set of residues previously identified as being essential for human

LINE-1 retrotransposition (R261/R262). These two amino acids were shown to be critically involved in ORF1p stability and thus influence the assembly of L1 RNPs (Doucet et al., 2010; Kulpa & Moran, 2005). Other ZfL1 subfamilies do show conservation of these residues, suggesting it might also be important for the formation of the ZfL1 RNPs. However, the effect of the mutation of these residues was only analysed in pair, meaning we do not know the effect of single mutations on LINE-1 retrotransposition. The ZfL1-7B and -10B subfamilies only have a single residue affected. Furthermore, since these substitutions are present in the consensus sequence of these subfamilies (as well as every copy analysed in this thesis), these elements have been able to retrotranspose with these alterations. This suggests either that the natural activity level of these subfamilies is extremely low, below the detection levels of our assay, or that other alterations are responsible for their lack of activity. The ZfL1-7B subfamily contains additional non-conserved substitutions in conserved residues in the EN domain, and the ZfL1-10B subfamily in the RT and Z domain. However, the importance of these residues in human LINE-1 retrotransposition has not been established. Finally, ZfL1-12B shows no alterations in any residues identified as important for human LINE-1 retrotransposition, but does contain several in the CTD, RRM and EN domain.

To confirm that absence of observed retrotransposition in the assays utilising the IVT RNAs were not due to poor translation of ZfL1 RNAs, I engineered a novel reporter design consisting of a mCherry cassette fused to the N-terminus of ZfL1-ORF1p by a self-processing 2A sequence. Notably, this configuration has not been used before in the context of a retrotransposition assay. As a control, I added the same mCherry-2A to the N-terminus of the single ORF of LINE-2 Zfl2-2 elements (i.e., ZfL2-2-WT mCherry construct). The retrotransposition of the ZfL2-2-WT mCherry construct serves as a proof of principle for this design, which can be used to track RNA translation without interfering with retrotranposition, allowing these two processes to be tracked independently (***Figure 4.17***; ***Figure 4.18***; ***Figure 4.19***). Using these mCherry constructs, I firstly found that the cap-dependent translation of the IVT RNA of the three ZfL1s used here was very efficient *in vivo* and *in vitro*. Indeed, consensus ZfL1-7B and -10B RNAs were significantly more translated than the retrotransposition-competent ZfL2-2-WT RNA (***Figure 4.19***). Interestingly, while ORF1p from the three ZfL1 were translated, I noticed that ZfL1-12B was translated at a lower level than the other two ZfL1s tested here. The reason behind the lower translation of the ZfL1-12B ORF1p warrants

additional research. For instance, since all the constructs maintained the L1 5'UTR upstream of mCherry, variations of the Kozak consensus surrounding the ATG start codon across the ZfL1 subfamilies and/or RNA structures, could influence transcription levels. However, the result of this experiment can also be influenced by differences in RNA stability, which was not controlled for. Since the RNA is *in vitro* transcribed, and therefore equally capped and poly-adenylated, RNA stability would be sequence-dependent. Secondary and/or tertiary RNA structures as well as RNA modifcations and cellular factors influencing RNA stability in a sequence-dependent manner, as described for other transposable elements, are examples that could also contribute to the detected differences (1.2.2; Boo & Kim, 2020).

4.8. Translation efficiency of mono- *vs* bicistronic RNAs

It is important to note that the mCherry translation that we are detecting in the ZfL1s is cap-dependent, and coupled to the synthesis of a functional ORF1p at the same levels, as is expected due to the self-processing nature of the 2A sequence located between mCherry and ORF1. LINE-1 research in mammals have shown that the ORF1p is abundantly translated in a cap-dependent manner, while translation of the second ORF is much less efficient and likely relies on a re-initiation event occurring after a ribosome completes ORF1p translation (Alisch et al., 2006) (1.2.1). In fact, ORF2p translation is so low that it has been speculated that a functional LINE-1 ribonucleoparticle contains one or two ORF2p molecules and it is undetectable in cellular extracts without enrichment methods (Mita et al., 2018). The mCherry system assayed here only reports cap-dependent translation of ORF1p and, at least when assayed in human LINE-1, it minimally disturbs ORF2p translation as inferred from the retrotransposition levels (data not shown). Therefore, we can conclude that the cap-dependent translation of ORF1 is unaffected in our system, but since we do not detect ZfL1 retrotransposition, we cannot determine if the ORF2 in these elements is being translated. Indeed, the long and the ambiguous sequence of the IGR of ZfL1s, compared to the strikingly short IGR in mammalian LINE-1s, could have resulted in the attenuation of the activity of these elements along evolution.

Another interesting result obtained with the mCherry system concerns the lower level of translation of ZfL2-2 RNAs observed *in vitro* and *in vivo*. This could be directly linked to the fact that this element belongs to an evolutionary older branch of LINEs that has not acquired

the additional ORF1, containing only a single ORFp encoding the required EN and RT activities. From our knowledge about bicistronic LINEs, it is reasonable to think that monocistronic LINEs might need fewer rounds of translation in order to generate a functional RNP, where the single ORFp binds the RNA of the LINE in *cis*, catalysing retrotransposition. Since all the RNAs transfected here are capped, its readiness for translation is expected. However, the increase in mCherry translation detected when the synthesis of the ORFp is abolished in the ZfL2-2-STOP construct, strongly suggests that either the formation of the ORFp-RNA ribonucleoparticle and/or its ability to translocate to the nucleus could be sequestering ZfL2-2 RNA from translation. This is consistent with the similar levels of translation detected for the ZfL2-2-WT, -EN^{-} and -RT^{-} RNAs, since mutations in these functional domains are not expected to affect RNP formation (***Figure 4.20B***). The lack of mCherry increase in ZfL2-2-TBR2m and -NSL may indicate that disrupting the stem-loop-TBR2 interaction is not sufficient to impair RNP formation *in vivo*. The ZfL2-2 ORFp could have more than one RNA binding domain and/or more than one ZfL2-2 RNA binding sites, as it has been described for the monocistronic R2 retrotransposon in arthropods (Jamburuthugoda & Eickbush, 2014). Interestingly, mCherry in fish embryos is nearly undetectable when using mCherry-2A chimeric constructs, which could suggest that in their natural environment, ZfL2-2 RNAs achieve retrotransposition with even lower translation than in HeLa cells. As mentioned before, an important draw back of this experimental set-up is that it does not account for differences in RNA stability between the constructs. Although there is a low possibility that point mutations would lead to such changes, and the fact that an increase in translation was only observed in the construct we expected to show higher translation, additional experiments to assess RNA stability (e.g. southern blot) should be considered when further exploring this research question.

In sum, our reporter system can be a useful tool for future research to clarify the translation stoichiometry of monocistronic and bicistronic LINEs.

Chapter 5

Conclusions

5. Conclusions

5.1. RNase H involvement in LINE-1 retrotransposition

To contribute to the knowledge of how the human LINE-1 is regulated, in this thesis I explored the importance of the PIP motif of RNase H2 and L1-ORF2p on RNase H2's ability to facilitate LINE-1 retrotransposition, building from previous work by our team (Benitez-Guijarro et al., 2018). My work corroborates previous findings that RNase H2 is needed to support efficient LINE-1 retrotransposition (Benitez-Guijarro et al., 2018), and that the PIP motif of L1-ORF2p is important for its retrotransposition (Taylor et al., 2013). In this thesis, I describe for the first time, that the PIP motif of RNase H2 does not mediate the processes that underlie its function in human LINE-1 retrotransposition. Additionally, I found that very low levels of RNase H2 (<15% of WT) are sufficient to support WT levels of LINE-1 retrotransposition, at least in HeLa cells and using engineered LINE-1s. Finally, I suggest an alternative function of the PIP motif found in L1-ORF2p. Based on the conservation between L1-ORF2p EN domain and human APE2, I propose that L1-ORF2p could have acquired its EN domain from an APE2 ancestral gene, which also provided the PIP motif. Therefore, the role of the L1-ORF2p PIP motif in retrotransposition might mirror the role of the PIP motif in the cellular function of APE2. Future research could take a closer look at cellular factors known to be coordinated by PCNA together with APE2, through their PIP motifs, and their potential involvement in regulating LINE-1 retrotransposition.

Limitations of this study include the fact that the RNase H2 cellular activity against RNA:DNA hybrids, which is suggested to facilitate LINE-1 retrotransposition, was not measured in the different cell lines generated. Instead, the way to check RNase H2 functionality was by measuring its ability to excise misincorporated ribonucleotides within short DNA substrates. Notably, RNASEH1 also possess activity against RNA:DNA hybrids and its overexpression has been shown to partially rescue LINE-1 retrotransposition. Although absence of RNase H2 protein was also confirmed by western blot, the endogenous RNASEH1 present in the cells could be a confounding factor in this study and could contribute to the observed variation between cell lines. Furthermore, HeLa cells were chosen as the parental cell line used to generate the RNase H2 cellular models, and the impact of their already known unstable

karyotype and intrinsic genetic variability could have become more prevalent as a result of the clonal expansions, which could have influenced my results.

5.2. Active LINE-1 elements in the zebrafish genome

To explore the value of zebrafish as a model in LINE-1 research, I assessed the retrotransposition competence of several previously described zebrafish LINE-1 subfamilies (Boissinot & Sookdeo, 2016). The three families analysed in this thesis, ZfL1-7B, -10B and -12B, did not show retrotransposition activity *in vitro* nor *in vivo*, using the experimental designs tested here. Later re-analyses of zebrafish LINE-1 expression at different zebrafish developmental stages, did reveal another promising ZfL1 subfamily, ZfL1-1D (the retrotransposition competence of this subfamily was not analysed in this thesis). Exploring the expression of IVT RNAs, I engineered a novel reporter design, allowing the analyses of LINE-1 translation and retrotransposition independently. Using this system, in combination with a previously described active zebrafish LINE-2, ZfL2-2, I show that this design does not interfere with retrotransposition. This reporter system can be used in future research investigating the translation stoichiometry of monocistronic and bicistronic LINEs.

Limitations of this study are related to the repetitive nature of LINE-1 sequences. This characteristic of LINE-1 makes it very difficult to map short reads to specific elements, especially younger ones that share more similarity to each other. This makes the reference genome an unreliable source for LINE-1 copies, since older elements that have accumulated more mutations are more likely to be correctly annotated. The fact that the zebrafish genome has undergone an additional round of whole-genome duplication and shows the highest reported repeat content in vertebrates so far (Howe et al., 2013b), contributes to the difficulty in its annotation; in fact, this also affects transcriptomic analyses of RNA-seq data. Unspecific mapping of reads to repetitive and homopolymeric regions within the LINE-1 sequences led to an altered representation of LINE-1 expression along zebrafish development. The use of only the first 1 Kb sequence of the LINE-1 is a solution to this problem. Unfortunately, this data was only obtained later in the project and did not contribute to the selection of the ZfL1 subfamilies for further experimental analysis.

The issue with the annotation of LINE-1 copies in the zebrafish genome could have potentially led to an underestimation of the copies of each subfamily, enriching for mainly older copies that are less likely to have retained retrotransposition competence. To overcome this caveat, I generated constructs with the consensus sequence of each subfamily, in order to eliminate disabling mutations that would have accumulated independently in the individual elements analysed. However, the consensus sequences were also not functional in our experimental conditions; notably, I was able to confirm that mRNAs were functional and properly translated in fish egg/embryos, by detecting cap-dependent ORF1 translation. However, ORF2 is also essential for retrotransposition and its translation is highly inefficient, with different LINEs employing different strategies to initiate its translation (re-initiation events *vs* presence of an IRES among others). The naturally low abundance of ORF2p has complicated endogenous ORF2p detection in a variety of models tested (Ardeljan et al., 2020). Remarkably, the intergenic region in these elements is abnormally longer than in other bicistronic L1s, and is very rich in poorly conserved homopolymeric stretches. Initial predisposition of poor fidelity in the replication of this area could have led to the current configuration, progressively disabling the translation by re-initiation or the existence of a structural domain that could facilitate translation in all subfamily members (i.e. and IRES). Although it is uncertain if the addition of flag/tags to ORF2p would affect its functionality, these approaches could be use in the future to inform about the correct generation of ORF2p and/or allow a system to troubleshoot this issue. As an alternative, *in vitro* methods developed to measure RT activity of ORF2p in WCEs, such as LEAP (Kopera et al., 2016), could be used to further test ORF2p translation.

Appendices

6. Appendices

6.1. Oligonucleotides

Name	Sequence (5'-3')	Description
RNH2B-CRISPR-F	TCTGTCTTTCTCACTTAGGGTTGG	Check CRISPR targeting of human *RNASEH2B* exon 3 (product size: 749 bp).
RNH2B-CRISPR-R	GACACACAGGAGATGACTTAACCT	
RNH2B-CRISPR-seqF	TGAGGATTATGGAGCTGGAAAACT	Forward sequencing primer on 749bp product.

Table 6.1. Oligonucleotides for RNASEH2B CRISPR KO validation.

Name	Sequence (5'-3')	Description
ecoliRH1-5NheI	ATAGCTAGCGGCTCGGGCTCGGGCTCGGGCTCGGGCTCGCTTAAACAGGTAGAAATTTTCACC	Amplify *E. coli RNASEHI* from vectors for cloning (***Table 2.3***; pMAR673 and pMAR683), removing start codon, incorporating unique cloning sites (NheI and PacI) and GC rich linker sequence (products size: 509 bp). Products cloned in pGEM-T-ORF2C-OR (***Table 2.3***).
ecoliRH1-3-Pac1	GTTAATTAAACTTCAACTTGGTAGCCTG	
hRH1-5NheI	ATAGCTAGCGGCTCGGGCTCGGGCTCGGGCTCGGGCTCGGAGACTTCGTCGTCGTCTACAC	Amplify catalytic core of human *RNASEH1* from vectors for cloning (***Table 2.3***; pMAR676 and pMAR684), incorporating unique cloning sites (NheI, PmeI and PacI) and GC rich linker sequence (products size: 516 bp). Products cloned in pGEM-T-ORF2C-OR (***Table 2.3***).
hRH1-3-Pme1Pac1	GTTAATTAAGTTTAAACTCAGTCTTCCGATTGTTTAGCTCC	
M13F	GTAAAACGACGGCCAG	Colony PCR of above cloning, together with primer jORF2Hs (products size: 1109 bp and 1129, for *E. coli* and human respectively).
jORF2Hs	GATATCATCTCACACCAGTTAG	Colony PCR (see above), and forward sequencing primer of colony PCR.
jORF2Is	GACTGGATTAAGAAAATGTG	Second forward sequencing primer of colony PCR.

Table 6.2. Oligonucleotides for cloning of LINE-1 ORF2p-RNase H chimeric constructs.

Name	Sequence (5'-3')	Product size	Description
qPCR_L1.1_169F	ATGGCGGGCAAACTACGTAA	89bp	Targeting ZfL1-1 5'end.
qPCR_L1.1_257R	GGCTCGGACATCATGCTAGG		
qPCR_L1.1_13F	GAGCAGGCGATGGAGTAAGG	173bp	
qPCR_L1.1_185R	CGTAGTTTGCCCGCCATTTG		
qPCR_L1.1B_15F	TCGGCAACATGTGAAGACGT	81bp	Targeting ZfL1-1B 5'end.
qPCR_L1.1B_95R	CTGGTCGAGAATGCGTTTGTG		
qPCR_L1.1B_8F	GGTTTGGTCGGCAACATGTG	179bp	
qPCR_L1.1B_186R	GCTCGAGAAAGTCCAGAAGCA		

Continues in the next page.

Name	Sequence (5’-3’)	Product size	Description
qPCR_L1.1D_31F	CAGCCTAAACTGTAGCTCCGC	114bp	Targeting ZfL1-1D 5’end.
qPCR_L1.1D_144R	GGTCTCGGGCATGTTCTCTG		
qPCR_L1.1D_19F	TGGAGAAGATGGCAGCCTAAAC	139bp	
qPCR_L1.1D_157R	GTGGAAAATCCTGGGTCTCGG		
qPCR_L1.6_223F	CATCGTCAGCCATGCAATCG	179bp	Targeting ZfL1-6 5’end.
qPCR_L1.6_401R	TCGACGCCATGTTCTTACCC		
qPCR_L1.6_253F	GGAAGCCTTGCGTGAAATGG	147bp	
qPCR_L1.6_399R	GACGCCATGTTCTTACCCAGA		
qPCR_L1.7B_796F	TCCTAAAGCTGCGCAGTGAAT	141bp	Targeting ZfL1-7B 5’end.
qPCR_L1.7B_936R	CTTCGTTAGCTGTCAGCAGCC		
qPCR_L1.7B_784F	ACAGCCCAGATGTCCTAAAGC	149bp	
qPCR_L1.7B_932R	GTTAGCTGTCAGCAGCCATAGT		
qPCR_L1.7C_143F	CGCGTTACCATGTCAAAGGC	132bp	Targeting ZfL1-7C 5’end.
qPCR_L1.7C_274R	AAAATCGCTTGACAGGCCCT		
qPCR_L1.7C_2F	ACTTCCGGTTATGAGGCGTG	150bp	
qPCR_L1.7C_151R	GGTAACGCGATCAAATTGACTGC		
qPCR_L1.8_30F	AGGGCTGAAGACTAGCTCCG	73bp	Targeting ZfL1-8 5’end.
qPCR_L1.8_102R	GTTAACAGTGGGGTGCTAACTTACT		
qPCR_L1.8_1F	GAGTCACGTGGGACCGTTG	70bp	
qPCR_L1.8_70R	AATAAATTAACGAGGTCAAGCCGGA		
qPCR_L1.10B_10F	GGTTATGGCGAGGTGCTGAG	153bp	Targeting ZfL1-10B 5’end.
qPCR_L1.10B_162R	CTTTTCTGGTACCTTTGCGTGAC		
qPCR_L1.10B_1F	TGGACTTCCGGTTATGGCGA	72bp	
qPCR_L1.10B_72R	GGATTGATCGCAAGTTAGCAGC		
qPCR_L1.11A_5F	TGATGGCTGAGTAACGCGAG	134bp	Targeting ZfL1-11A 5’end.
qPCR_L1.11A_138R	GGGCGAAACACACGGAGAAA		
qPCR_L1.11A_33F	AGACCTGAGTGCGAGATCCA	113bp	
qPCR_L1.11A_145R	TTTTGATGGGCGAAACACACG		
qPCR_L1.12A_42F	TTTTGTCTCCCGTTCGGTCG	196bp	Targeting ZfL1-12A 5’end.
qPCR_L1.12A_237R	CCTCTTGTCCGCCATTACGT		
qPCR_L1.12A_141F	GGCAAAGGAAAACAGCTTGGG	92bp	
qPCR_L1.12A_232R	TGTCCGCCATTACGTTCTCC		
qPCR_L1.12B_27F	TGGCAGCATAGTGCATCGTC	167bp	Targeting ZfL1-12B 5’end.
qPCR_L1.12B_193R	TTGTTGCCCCACTATCCACAAG		
qPCR_L1.12B_247F	CGTCAACAGCCTAGTAACCCTG	94bp	
qPCR_L1.12B_340R	ACTGGCGGTTGTTGTCCTTC		
qPCR_L1.13A_27F	ACGACGCCTATCAACAGAGC	136bp	Targeting ZfL1-13A 5’end.
qPCR_L1.13A_163R	TGCTCATCATGTAGTCTCGACAC		
qPCR_L1.13A_32F	GCCTATCAACAGAGCTCCCG	149bp	
qPCR_L1.13A_180R	TTGCGGCTGATTTATTCTTGCTC		

Continues in the next page.

Name	Sequence (5'-3')	Product size	Description
qPCR_L1.13B_15F	GCATGTGAGTGGGTAGACGC	134bp	Targeting ZfL1-13B 5'end.
qPCR_L1.13B_148R	AACTGTAACCGTCTGATGCCC		
qPCR_L1.13B_33F	GCACTTGAGGTTAGCTCCCG	114bp	
qPCR_L1.13B_146R	CTGTAACCGTCTGATGCCCG		
qPCR_L1.13C_124F	GACGTGGATGCAAAATGCCG	80bp	Targeting ZfL1-13C 5'end.
qPCR_L1.13C_203R	AGCCAGCTTGCTGCTTATCG		
qPCR_L1.13C_34F	CGTGCGTATGAGCCTCCTG	115bp	
qPCR_L1.13C_148R	GAACCCGGCATTTTGCATCC		
qPCR_L1.13D_213F	GCAAGTTAGCAACCATGGCC	129bp	Targeting ZfL1-13D 5'end.
qPCR_L1.13D_341R	CTTCCCTGATACTAACACGCTGT		
qPCR_L1.13D_23F	AGGAGTAAGGTGCGTGGAAGA	185bp	
qPCR_L1.13D_207R	ATGCTCGTGGTAGATCTCGTTG		
qPCR_L1.16B_229F	CTTGTGTGGAAGAAACGCCTG	95bp	Targeting ZfL1-16B 5'end.
qPCR_L1.16B_323R	TGAGTGAAGTTCTGCACGGAG		
qPCR_L1.16B_27F	TAGCACGCACGTCAAGACAG	158bp	
qPCR_L1.16B_184R	GCCGGCATAATAAGAAAACTGGAC		
qPCR_L1.17B_125F	GTTTGATCATGCCCAAGCCG	84bp	Targeting ZfL1-17B 5'end.
qPCR_L1.17B_208R	CCTCGCTGGAGTTAGTGCAA		
qPCR_L1.17B_30F	TCGCGTTCTTTTACCTCTCCC	146bp	
qPCR_L1.17B_175R	CTCGAGAATGGTCCGGCTTG		
qPCR_L2.2_866F	CGGCCCTTGTTTTGACTCT	63bp	Targeting ZfL1-1B 5'end.
qPCR_L2.2_929R	AACCCCAAAGCACTAGCTGA		
EF1a-FW	CTTCTCAGGCTGACTGTGC	358bp	Targeting the EF1α gene
EF1a-RV	CCGCTAGCATTACCCTCC		
b-act-FW	GCCTGACGGACAGGTCAT,	94bp	Targeting the Actβ gene
b-act-RV	ACCGCAAGATTCCATACCC		

Table 6.3. Oligonucleotides used for validation of ZfL1 RNA-seq by qPCR.

Name	Sequence (5'-3')	Description
L1.6R	CTTCGATTGCATGGCTGACG	Reverse primer at 5'end of ZfL1-6 subfamily.
L1.6F	TGTGCGTATAATGGGGTGGG	Forward primer at 3'end of ZfL1-6 subfamily.
L1.6-Chr6F	GCCTTTGGTCTTTAAATTCACCC	Forward and reverse primer targeting zebrafish genome 5' and 3' of the ZfL1-6 copy at Chr6:21659365-21665057, respectively.
L1.6-Chr6R	AGAAAACGGTTGATGGTCGC	
L1.6-Chr10F	GCAAGCCCAGATAGTCAAGC	Forward and reverse primer targeting zebrafish genome 5' and 3' of the ZfL1-6 copy at Chr10:42609301-42614999, respectively.
L1.6-Chr10R	TGTGCTGGCAGTTTGTTACA	
L1.6-Chr11F	GCCTAATTACCCTAACCTGCC	Forward and reverse primer targeting zebrafish genome 5' and 3' of the ZfL1-6 copy at Chr11:7069121-7074820, respectively.
L1.6-Chr11R	CAGACGAAGGAAGGAGTGGT	

Continues in the next page.

Name	Sequence (5’-3’)	Description
L1.6-Chr19F	ACAGGAATAACACATGCGGC	Forward and reverse primer targeting zebrafish genome 5’ and 3’ of the ZfL1-6 copy at Chr19:24218046-24223737, respectively.
L1.6-Chr19R	GCCCCTCATGACAAACGAGA	
L1.7BR	TCCACCTGTTTGCCCTTTCC	Reverse primer at 5’end of ZfL1-7B subfamily.
L1.7BF	CTTTGCTTGCTAGGCGTGTC	Forward primer at 3’end of ZfL1-7B subfamily.
L1.7B-Chr4F	GACGCAGCACTGTTCAGTAC	Forward and reverse primer targeting zebrafish genome 5’ and 3’ of the ZfL1-7B copy at Chr4:45903129-45908934, respectively.
L1.7B-Chr4R	CACAAGTCACATGAGTCCAACA	
L1.7B-Chr5F	TGTTCAGCTACCAACTTTTCCA	Forward and reverse primer targeting zebrafish genome 5’ and 3’ of the ZfL1-7B copy at Chr5:41057392-41063181, respectively.
L1.7B-Chr5R	CACCTTTAAGAAGCGCACAAA	
L1.7B-Chr6F	GATGATGACACTTCCACGGC	Forward and reverse primer targeting zebrafish genome 5’ and 3’ of the ZfL1-7B copy at Chr6:18504003-18509793, respectively.
L1.7B-Chr6R	AGTTGCATAGTTCCCGATAAAGA	
L1.7B-Chr11F	TGGTACGGTACTGTTCGCTT	Forward and reverse primer targeting zebrafish genome 5’ and 3’ of the ZfL1-7B copy at Chr11:25254051-25259845, respectively.
L1.7B-Chr11R	GGAATCCAAGAACACTAAGGGG	
L1.7B-Chr24F	ACTTGGATAGGTGGCAAGACT	Forward and reverse primer targeting zebrafish genome 5’ and 3’ of the ZfL1-7B copy at Chr24:31415464-31421258, respectively.
L1.7B-Chr24R	GCCTTTGACCCTTGACCTTG	
L1.10BR	GCCATGGCGTCGTTAGTTTC	Reverse primer at 5’end of ZfL1-10B subfamily.
L1.10BF	GAATGGGTTGTTGGGAGGGA	Forward primer at 3’end of ZfL1-10B subfamily.
L1.10B-Chr3F	AGACGGCTGAAAATACACTCA	Forward and reverse primer targeting zebrafish genome 5’ and 3’ of the ZfL1-10B copy at Chr3:1795415-1801315, respectively.
L1.10B-Chr3R	CGTGGGGTTTGTGCGTAC	
L1.10B-Chr5F	CTTGCCACAACAAAGGAGCT	Forward and reverse primer targeting zebrafish genome 5’ and 3’ of the ZfL1-10B copy at Chr5:42036670-42042519, respectively.
L1.10B-Chr5R	CAGCGGATGATTAAACGACTCT	
L1.10B-Chr8F	ATCAGTGCTTGCTCATGTGA	Forward and reverse primer targeting zebrafish genome 5’ and 3’ of the ZfL1-10B copy at Chr8:24020995-24026840, respectively.
L1.10B-Chr8R	TTTGAATCGCCGACAAGTCC	
L1.10B-Chr14F	ACAGGGCTATCATTAACTTTGTG	Forward and reverse primer targeting zebrafish genome 5’ and 3’ of the ZfL1-10B copy at Chr14:37952310-37958210, respectively.
L1.10B-Chr14R	ATTAAGCAAAAGAAAGCAGGTCA	
L1.10B-Chr18F	CCAAAATCGAGCAGCCCTAC	Forward and reverse primer targeting zebrafish genome 5’ and 3’ of the ZfL1-10B copy at Chr18:32301158-32307002, respectively.
L1.10B-Chr18R	CGCCTCTGAAGAATCACTGC	

Continues in the next page.

Name	Sequence (5'-3')	Description
L1.11AR	GATCTCGCACTCAGGTCTCG	Reverse primer at 5'end of ZfL1-11A subfamily.
L1.11AF	ACACACTCACGCACATACCA	Forward primer at 3'end of ZfL1-11A subfamily.
L1.11A-Chr3F	ATCTGGGCTCTGACTGTTGT	Forward and reverse primer targeting zebrafish genome 5' and 3' of the ZfL1-11A copy at Chr3:868842-876012, respectively.
L1.11A-Chr3R	TCCTCTGTTTCTGCTGCTCA	
L1.11A-Chr8F	AAATGAGCGTCTCCCTGTGT	Forward and reverse primer targeting zebrafish genome 5' and 3' of the ZfL1-11A copy at Chr8:24262630-24269847, respectively.
L1.11A-Chr8R	AGTTTAAACACTTACACTCTCTGA	
L1.12BR	ATGCTGCCATTCCCTCACTC	Reverse primer at 5'end of ZfL1-12B subfamily
L1.12BF	CCAATGTACCTTGAAATGCCGT	Forward primer at 3'end of ZfL1-12B subfamily.
L1.12B-Chr19F	AGCCAACCAATCACATTCAACA	Forward and reverse primer targeting zebrafish genome 5' and 3' of the ZfL1-12B copy at Chr19:12893436-12899204, respectively.
L1.12B-Chr19R	GCCATGAAACCTCCAACACT	
L1.12B-Chr20F	AGCCAGTATCTCCCTTTGCA	Forward and reverse primer targeting zebrafish genome 5' and 3' of the ZfL1-12B copy at Chr20:17401219-17406985, respectively.
L1.12B-Chr20R	CATGTCGCATCCTGTGTTGT	
L1.13AR	GCGTCGTCTTCTCACATTGC	Reverse primer at 5'end of ZfL1-13A subfamily.
L1.13AF	TAAATTCCCAACCCCACCCC	Forward primer at 3'end of ZfL1-13A subfamily.
L1.13A-Chr2F	CACTTTCAAACGCCACCACT	Forward and reverse primer targeting zebrafish genome 5' and 3' of the ZfL1-13A copy at Chr2:7491273-7497084, respectively..
L1.13A-Chr2R	ACTCTGTCTTGTGGGAAGCT	
L1.13A-Chr7F	GTACGCAGAATCACACGCAT	Forward and reverse primer targeting zebrafish genome 5' and 3' of the ZfL1-13A copy at Chr7:1203268-1209061, respectively.
L1.13A-Chr7R	CCAAGCTAATGTGACATGGGG	
L1.13BR	TAACCGTCTGATGCCCGATG	Reverse primer at 5'end of ZfL1-13B subfamily.
L1.13BF	GGGGATTGTTTTGTTTTGTGGC	Forward primer at 3'end of ZfL1-13B subfamily.
L1.13B-Chr16F	CGCAATTGGGGATCGGATAA	Forward and reverse primer targeting zebrafish genome 5' and 3' of the ZfL1-13A copy at Chr16:39601658-39607121, respectively.
L1.13B-Chr16R	CCCAGTGGAGTACGAAAGGT	

Table 6.4. Oligonucleotides for genotyping PCRs of ZfL1 copies in the zebrafish genome.

Name	Sequence (5'-3')	Description
AscI_L1.6-Chr6F	CTAGGCGCGCCAGGATTGCTGGACTTCCGG	Amplify full-length ZfL1-6 element Chr6:21659365-21665057, incorporating unique cloning sites (AscI and SrfI at 5' and 3' end, respectively).
SrfI_L1.6-Chr6R	TGAGCCCGGGCGCAGCAATCCTCTTATTTTTTTTTG	
AscI_L1.6-Chr11F	AGAGGCGCGCCGAAAAATATCTAGTAGACTTCCGG	Amplify full-length ZfL1-6 element Chr11:7069121-7074820, incorporating unique cloning sites (AscI and SrfI at 5' and 3' end, respectively).
SrfI_L1.6-Chr11R	GACGCCCGGGCCTATTTTACTAGATATTTTCTTTTTTTTTG	

Continues in the next page.

Name	Sequence (5'-3')	Description
AscI_L1.6-Chr19F	TAAGGCGCGCCAAAAATGAAAAATACAACTTCCGGTG	Amplify full-length ZfL1-6 element Chr19:24218046-24223737, incorporating unique cloning sites (AscI and SrfI at 5' and 3' end, respectively).
SrfI_L1.6-Chr19R	AGGGCCCGGGCGAAGCGTGTATTTTTCATTTTTTTTTTTTG	
L1.6_3212R	CCTTTGGTTATGGCCTTGTT	Reverse primer to amplify 5' end only of ZfL1-6 subfamily with *locus* specific forward primers.
AscI_L1.7B-Chr4F	GCAGGCGCGCCACTTGATTTCACAAGACTTCCGG	Amplify full-length ZfL1-7B element Chr4:45903129-45908934, incorporating unique cloning sites (AscI and SrfI at 5' and 3' end, respectively).
SrfI_L1.7B-Chr4R	TTTGCCCGGGCATGTAGTGAAATCAAGTCTTTTTTTTTTTT	
AscI_L1.7B-Chr5F	ATAGGCGCGCCGAAATTGTTCGAGTCTTCCGGTT	Amplify full-length ZfL1-7B element Chr5:41057392-41063181, incorporating unique cloning sites (AscI and SrfI at 5' and 3' end, respectively).
SrfI_L1.7B-Chr5R	TAAGCCCGGGCTAAAATGAACAATTTCTATTATTTTTTTTTTTTTTTA	
AscI_L1.7B-Chr6F	GCTGGCGCGCCTTTATCGGGGGACTTCCGGT	Amplify full-length ZfL1-7B element Chr6:18504003-18509793, incorporating unique cloning sites (AscI and SrfI at 5' and 3' end, respectively).
SrfI_L1.7B-Chr6R	TAAGCCCGGGCATAGTTCCCGATAAAGATTTTTTTTTTTTT	
AscI_L1.7B-Chr24F	AGAGGCGCGCCAATGTAGACTTCCGGTTATGTCG	Amplify full-length ZfL1-7B element Chr24:31415464-31421258, incorporating unique cloning sites (AscI and SrfI at 5' and 3' end, respectively).
AscI_L1.7B-Chr24F_2	GGAGGCGCGCCAAGACTTAAGGTAATGTAGGACTTC	
SrfI_L1.7B-Chr24R	GTTGCCCGGGCTACATTACCTTAAGTCTTTTTTTTTTTTTAT	
SrfI_L1.7B-PolyAR	TTATGGCCCGGGCAACATTTTTCCATTTTAC	Reverse primer removing poly(A) signal in 3'UTR of ZfL1-7B and incorporating unique cloning site (SrfI).
AcsI_L1.10B-Chr5F	AATGGCGCGCCTGCCAGGGCTGGACCTTCC	Amplify full-length ZfL1-10B element Chr5:42036670-42042519, incorporating unique cloning sites (AscI and SrfI at 5' and 3' end, respectively).
SrfI_L1.10B-Chr5R	TTTGCCCGGGCTCTCAGCCCTGGCATTTTTTTTTTTTTTTG	
AcsI_L1.10B-Chr8F	CTCGGCGCGCCAACAATAACACAGTGGACTTCCG	Amplify full-length ZfL1-10B element Chr8:24020995-24026840, incorporating unique cloning sites (AscI and SrfI at 5' and 3' end, respectively).
SrfI_L1.10B-Chr8R	TCGGCCCGGGCAGTTTACACTGTGTTATTGTTTTTTTTTG	
AscI_L1.10B-Chr18F	TAAGGCGCGCCAAATAATTATTTACGGCGACTTCCG	Amplify full-length ZfL1-10B element Chr18:32301158-32307002, incorporating unique cloning sites (AscI and SrfI at 5' and 3' end, respectively).
SrfI_L1.10B-Chr18R	GATGCCCGGGCTGAGCAAAATGTAAATAATTATTTTTTTTTGTA	
SrfI_L1.10B-PolyAR	TATTAGCCCGGGCAATTTGCATATCACAAC	Reverse primer removing poly(A) signal in 3'UTR of ZfL1-10B and incorporating unique cloning site (SrfI).
AscI_L1.11A-Chr3F	GACGGCGCGCCCCTGATTCGACTGATGGCTG	Amplify full-length ZfL1-11A element Chr3:868842-876012, incorporating unique cloning sites (AscI and SrfI at 5' and 3' end, respectively).
SrfI_L1.11A-Chr3R	AGGGCCCGGGCGTTTCTGCTGCTCATCTTCTTTTTT	

Continues in the next page.

Name	Sequence (5'-3')	Description
AscI_L1.11A-Chr8F	CTCGGCGCGCCATTCTCAGAGACT GATGGCTG	Amplify full-length ZfL1-11A element Chr8:24262630-24269847, incorporating unique cloning sites (AscI and SrfI at 5' and 3' end, respectively).
SrfI_L1.11A-Chr8R	TTTGCCCGGGCTTTTTTTTGTTAAAC ACAAACATGATTTTTTT	
AscI_L1.12B-Chr19F	AATGGCGCGCCAATAATTATTAAAA ATGCTATAAAAAGACCG	Amplify full-length ZfL1-12B element Chr19:12893436-12899204, incorporating unique cloning sites (AscI and SrfI at 5' and 3' end, respectively).
SrfI_L1.12B-Chr19R	TAGGCCCGGGCTTTTAGTTTTTTTTT ATAGCATTTTTTTTTTTTTG	
L1.12B_3181R	ATCAGATTGGGCATGAGGTC	Reverse primer to amplify 5' end only of ZfL1-12B subfamily with *locus* specific forward primers.
SrfI_L1.12B-PolyAR	TTTATGGCCCGGGCGAACATTCAAT TGAG	Reverse primer removing poly(A) signal in 3'UTR of ZfL1-12B and incorporating unique cloning site (SrfI).
HpaI_AscI_T7p_SalIR	GTTGTTAACTTTGGCGCGCCCTATA GTGAGTCGTATTAGTCGACCCT	Remove excessive nucleotides, resulting from cloning steps, between pTOL2 T7 promoter and ZfL1 element.
pTOL2_ScreenF	CTGCGGCGAGTTCTAGCTG	
NotI-EcoRV-AscI-SrfI-ClaI-BstZ17IF	AGAGCGGCCGCGATATCGGCGCGC CATAGCCCGGGCATCGATGTATAC CCG	Oligos used to generate DNA fragment containing restriction sites (from 5' to 3': NotI, EcoRV, AscI, SrfI, ClaI and BastZ17I) to create pCEP4 and pTOL2 vectors containing the restriction sites AscI and SrfI to clone ZfL1 elements.
NotI-EcoRV-AscI-SrfI-ClaI-BstZ17IR	CGGGTATACATCGATGCCCGGGCT ATGGCGCGCCGATATCGCGGCCGC TCT	

Table 6.5. Oligonucleotides used to clone full-length zebrafish LINE elements in retrotransposition vectors.

Name	Sequence (5'-3')	Description
L1.6_226F	CGTCAGCCATGCAATCGAAG	. Stepping sequencing primers to sequence entire elements of the ZfL1-6 subfamily.
L1.6_517R	GCGATCTTCCAAATCCGTGC	
L1.6_498F	GCACGGATTTGGAAGATCGC	
L1.6_1034F	GAGGCAATGGAGGAAACCGA	
L1.6_1542F	GCCACAGACGCTCGTTATTG	
L1.6_2128F	TGGAGGAGACATGAACACGG	
L1.6_2670F	GTGGAGAATTGGAGCGGGAA	
L1.6_3121F	TCCAGGTATTGACGGCATCC	
L1.6_3524F	GATCGCTTGGAATGGGGCTA	
L1.6_4010F	GAGCTGAGCCCAACCTTACA	
L1.6_4535F	TTTCCTTCCAGAACCACCTG	
L1.6_4972F	CCACGCTTCCACAAACCCTA	
L1.6_5566F	TGTGCGTATAATGGGGTGGG	
L1.7B_272F	CTCTCTGGCCGCGGATTTTA	Stepping sequencing primers to sequence entire elements of the ZfL1-7B subfamily.
L1.7B_420R	TCTCCAGGTGTTCAAGGTGC	
L1.7B_585F	GGCGATCAGATTTTGGCGTC	
L1.7B_1059F	TCCGTAACACTCATGCCACT	
L1.7B_1549F	GCGGGTTGGTGTTTGTAGTT	
L1.7B_1907F	CGTTTTCCAGTAAGTCACGAGG	

Continues in the next page.

Name	Sequence (5'-3')	Description
L1.7B_2450F	CTCCTAGGTCTTTCTGGCGC	Stepping sequencing primers to sequence entire elements of the ZfL1-7B subfamily.
L1.7B_2777R	TTCGTTGTCAGCAGGTCAAA	
L1.7B_2910F	ACCCACAACTGACCCACTTC	
L1.7B_3507F	CATGGACGCTGAAAAGGCTT	
L1.7B_4079F	TCACACATGATTTTCGCCGC	
L1.7B_4457F	TCCTCTGCTTTCCCATCACG	
L1.7B_4972F	GCAGTTCTGGACAGGGTTCA	
L1.7B_5431F	AGACCTAACAGAACATTTGCCA	
L1.10B_237F	GGCCCTGACGAAGATGATGA	Stepping sequencing primers to sequence entire elements of the ZfL1-10B subfamily.
L1.10B_476R	TTCTAGCGTGCTCACTCTGG	
L1.10B_553F	GACTACCGGAGAAAGCGGAG	
L1.10B_951F	TACACCTTCCAAACTCCGGC	
L1.10B_1526F	AAGTGAGAGACAGACGGCTC	
L1.10B_1917F	TCTGCCCCATATACCACCCA	
L1.10B_2415F	TCGGACCATGCAGCTATGTC	
L1.10B_3004F	ACCCCAGACTTGTTGAACAGA	
L1.10B_3495F	AGAGGAGCGAAAGACACAGC	
L1.10B_3868F	GCTCTATTCCTGCACTCCTGG	
L1.10B_4435F	CCCGCTCTATCACACCACAA	
L1.10B_4991F	CGAGGCATGTAGGAGAGCTC	
L1.10B_5540F	ACAGGCTTAGGAATATGGGAAGT	
L1.12B_16F	GAGTGAGGGAATGGCAGCAT	Stepping sequencing primers to sequence entire elements of the ZfL1-12B subfamily.
L1.12B_417R	CCACGTTTTGCAATCGGGTT	
L1.12B_532F	TACAACGTGCCTGAAGATGC	
L1.12B_1030F	ACGTCGTGGGAAACTGTCTC	
L1.12B_1450F	TCTCCAGGGTTATTGTCGTCC	
L1.12B_1906F	AGGGTATTCTGATCTGTGGGG	
L1.12B_2417F	CAAGAAACTCCGGCAACAGG	
L1.12B_2981F	GGCCCGGTCCTTTAATTGGA	
L1.12B_3005F	CATCTTGGCCCGGTCCTTTA	
L1.12B_3434F	CGGCCACCTCTCAAACTCAT	
L1.12B_4030F	TTTCGCGCTTCATTTGGCAG	
L1.12B_4222R	GGGGCATGATAGTGTGGGTC	
L1.12B_4428F	TGGGCAAACAAGGGAATCACT	
L1.12B_5040F	TTTGGCTGTGTGGATCCAGA	
L1.12B_5447F	TGACATGTGTTCTAACCCTGGA	
ZFL2-2_884F	TTCCTTCTGTCTCTCAACATCCA	Stepping sequencing primers to sequence entire elements of the ZfL2-2 family.
ZFL2-2_1688F	CAGGCTAGGGTAACCCCAC	
ZFL2-2_2482F	ACTAAACTTCTCTGACCACATTTCT	
CMV_5Rv	TGACGTCAATGGAAAGTCCCT	Sequencing primer to sequence downstream of the retrotransposition cassettes

Table 6.6. Oligonucleotides used to sequence full-length zebrafish LINE elements.

Name	Sequence (5'-3')	Description
Srf_L1.7B_5518R	AGTGCCCGGGCTCAGTCTAG ATCCAGCTGGAG	Reverse primer removing 3'UTR of ZfL1-7B and incorporating unique cloning site (SrfI).
BamHI_L1.7B_3UTRF	CTAGGATCCTCACCTCAACTT TTGATCACAATAGTAC	Amplify 3'UTR of the ZfL1-7B, incorporating unique cloning sites (BamHI and HpaI at the 5' and 3' end, respectively) for cloning downstream of retrotransposition cassettes in pCEP4 and pTOL2 vectors.
HpaI_L1.7B_3UTRR	TTAGTTAACTTTTTTTTTTTTT TTATGATTTATTTAACATTTT TCCATTTTACAAACATATCAA AC	
SrfI_L1.10B_5564R	TCCGCCCGGGCTAAGCCTGT TCTATTTGTAGTAG	Reverse primer removing 3'UTR of ZfL1-10B and incorporating unique cloning site (SrfI).
BamHI_L1.10B3UTRF	CAGGGATCCGAATATGGGA AGTATCATGGTCTA	Amplify 3'UTR of the ZfL1-10B, incorporating unique cloning sites (BamHI and HpaI at the 5' and 3' end, respectively) for cloning downstream of retrotransposition cassettes in pCEP4 and pTOL2 vectors.
HpaI_L1.10B_3UTRR	TTAGTTAACTTTTTTTTTTTT TTTGTACAATATATTTATTAC TTTTTTCAATTTGCATATCAC AACAGATTACA	
SrfI_L1.12B_5342R	GCAGCCCGGGCTATTCATTT ATAGCAATAAAGTCAGG	Reverse primer removing 3'UTR of ZfL1-12B and incorporating unique cloning site (SrfI).
BamHI_L1.12B_3UTRF	AATGGATCCAAGATGATGCT CTGATAGATAAATATATG	Amplify 3'UTR of the ZfL1-12B, incorporating unique cloning sites (BamHI and HpaI at the 5' and 3' end, respectively) for cloning downstream of retrotransposition cassettes in pCEP4 and pTOL2 vectors.
HpaI_L1.12B_3UTRR	TTAGTTAACTTTTTTTTTTTT TTGCAACTTTATGTTTATTTT GAACATTCAATTGAGTACAA TCATAC	

Table 6.7. Oligonucleotides used to move ZfL1 3'UTRs behind retrotranposition inidicator cassettes.

Name	Sequence (5'-3')	Description
L1.7B_78R	GTTTGTAAAGAAGAAATAAAAAGTTAAAGG	Build ZfL1-7B consensus sequence.
L1.7B_49F	CCTTTAACTTTTTATTTCTTCTTTACAAAC	
L1.7B_1567R	CAACTACAAACACCAACCCGC	
L1.7B_2471R	GCGCCAGAAAGACCTAGGAG	
L1.7B_4089R	AGCGGCGAAAATCATGTGTGA	
L1.7B_5668R	GATAAAAAAAGTGAATAAATAAATTAAAATAAAAAAG	
L1.7B_5632F	CTTTTTTATTTTAATTTATTTATTCACTTTTTTTATC	
L1.10B_254R	TCATCATCTTCGTCAGGGCC	Build ZfL1-10B consensus sequence.
L1.10B_1330R	GATATAAATAAAAGAACAAATCAAACAGTAC	
L1.10B_1300F	GTACTGTTTGATTTGTTCTTTTATTTATATC	
L1.10B_1611R	GGAAAGAAAAAAAAAACACACCTTATAC	
L1.10B_1585F	GTATAAGGTGTGTTTTTTTTTCTTTCC	
L1.10B_2433R	GACATAGCTGCATGGTCCGA	
L1.10B_5009R	GAGCTCTCCTACATGCCTCG	
L1.12_927R	GTGAACCCCGTCCTCATAGT	Build ZfL1-12B consensus sequence.
L1.12_908F	ACTATGAGGACGGGGTTCAC	
L1.12_1027R	GTTTGAGTTGTTCCATTAACGATTC	
L1.12_1003F	GAATCGTTAATGGAACAACTCAAAC	

Continues in the next page.

L1.12_1508R	ACTTATGTTGGTTTAGTATTTCAATATCT	Build ZfL1-12B consensus sequence.
L1.12_1480F	AGATATTGAAATACTAAACCAACATAAGT	
L1.12_1783R	CGACCTTCTTTGTCAGCATATTC	
L1.12_1761F	GAATATGCTGACAAAGAAGGTCG	
L1.12_2140R	GTTTTAAACATAAGGTAATAATCAATCC	
L1.12_2113F	GGATTGATTATTACCTTATGTTTAAAAC	
L1.12_4943R	CCAAAAGATATGATAATGGTTTGCTT	
L1.12_4918F	AAGCAAACCATTATCATATCTTTTGG	

Table 6.8. Oligonucleotides used to build ZfL1 consensus sequences.

Name	Sequence (5'-3')	Description
Cherry_F	ATGGTGAGCAAGGGCGAGG	Amplify mCherry sequence from standard mCherry containing vector.
2AR	AGGTCCAGGGTTCTCCTCC	
pTOL2_F	AGCACTTTGGGAATTCTGGA	Forward primer upstream of the pTOL2 cloning site, used to generate 5'end of ZfL1s containing mCherry flap.
CherryFlap_L1.7BR	CCTCGCCCTTGCTCACCATGTTAAG TGAAAATTTAATACTACGTACT	Reverse primer incorporating 19 bp of mCherry sequence 5' of the ZfL1-7B-ORF1p with pTOL2F.
2AFlap_L1.7BF	GGAGGAGAACCCTGGACCTTCGAA AGGAAAGGGCAAACAGG	Forward primer incorporating 19 bp of 2A sequence at the 5' of the ZfL1-7B-ORF1p with L1.7B_2471R.
CherryFlap_L1.10BR	CCTCGCCCTTGCTCACCATTTTTCTT CCTTATGCTTTTATATTTGTG	Reverse primer incorporating 19 bp of mCherry sequence 5' of the ZfL1-10B-ORF1p with pTOL2F.
L1.10BFlap_2A_R	CTTTTCTGGTACCTTTGCGTGAAGG TCCAGGGTTCTCCTCC	Forward primer incorporating 19 bp of 2A sequence at the 5' of the ZfL1-10B-ORF1p with L1_10B_2433Rv.
CherryFlap_L1.12BR	CCTCGCCCTTGCTCACCATTGTTGC CCCACTATTATCGGTT	Reverse primer incorporating 19 bp of mCherry sequence 5' of the ZfL1-12B-ORF1p with pTOL2F.
2AFlap_L1.12BF	GGAGGAGAACCCTGGACCTGAGAC GAAAAAAAAGAGCAGGAAA	Forward primer incorporating 19 bp of 2A sequence at the 5' of the ZfL1-12B-ORF1p with L1.12_2140R.
CherryFlap_L2_R	CCTCGCCCTTGCTCACCATAATAAA AGTGCACTGACAACAAAAAATAA	Reverse primer incorporating 19 bp of mCherry sequence 5' of the ZfL2-2-ORFp with pTOL2F.
2AFlap_L2_F	GGAGGAGAACCCTGGACCTTGTTT TCTAATTCCTGTTGTTACTAAC	Primers incorporating 19 bp of 2A sequence at the 5' of the ZfL2-2-ORFp.
ZFL2-2_2143R	CCTTGAGGTACCCCAGTGTT	

Table 6.9. Oligonucleotides used to build zebrafish LINE mCherry constructs.

Name	Sequence (5'-3')	Description
2AFlap_L2_StopF	GGAGGAGAACCCTGGACCTTGATT TCTAATTCCTGTTGTTACTAAC	Forward primer incorporating 19 bp of 2A sequence and STOP codon at the 5' of the ZfL2-2-ORFp with ZFL2-2_2143R

Continues in the next page.

Name	Sequence (5’-3’)	Description
ZFL2-2_TBR2m_1179R	GTTTTGAGGCATGCTCACGGAGAG CATCCGAGAGCGCGGGTGCAGG	Generate ZfL2-2-TBR2m construct.
ZFL2-2_TBR2m_1160F	CCGTGAGCATGCCTCAAAACTTCG	
ZFL2-2_2225_RTmF	CTGCTATGCTTATGACACCCAG	Generate ZfL2-2-RTm construct.
ZFL2-2_2246_RTmR	CTGGGTGTCATAAGCATAGCAG	
ZFL2-2_376_ENmF	CTCTAACTGCGACCTGGTTG	Generate ZfL2-2-ENm construct.
ZFL2-2_395_ENmR	CAACCAGGTCGCAGTTAGAG	

Table 6.10. Oligonucleotides used to build ZfL2-2 mutant constructs.

Name	Sequence (5’-3’)	Description
AmpR_R	AGTTGGCCGCAGTGTTATCA	Amplify pTOL2 downstream of restriction site used to linearize vector for IVT (product size: 243 bp).
pTOL_R	TCCTTGAGAGTTTTCGCCCC	
EGFP_F	CTGGGGCACAAGCTGGAGTA	Amplify EGFP (product size: 221 bp).
EGFP_R	TCTTTGCTCAGGGCGGACTG	

Table 6.11. Oligonucleotides validate integrated EGFP in zebrafish genome as result of retrotransposition of IVT RNA.

6.2. Vector maps

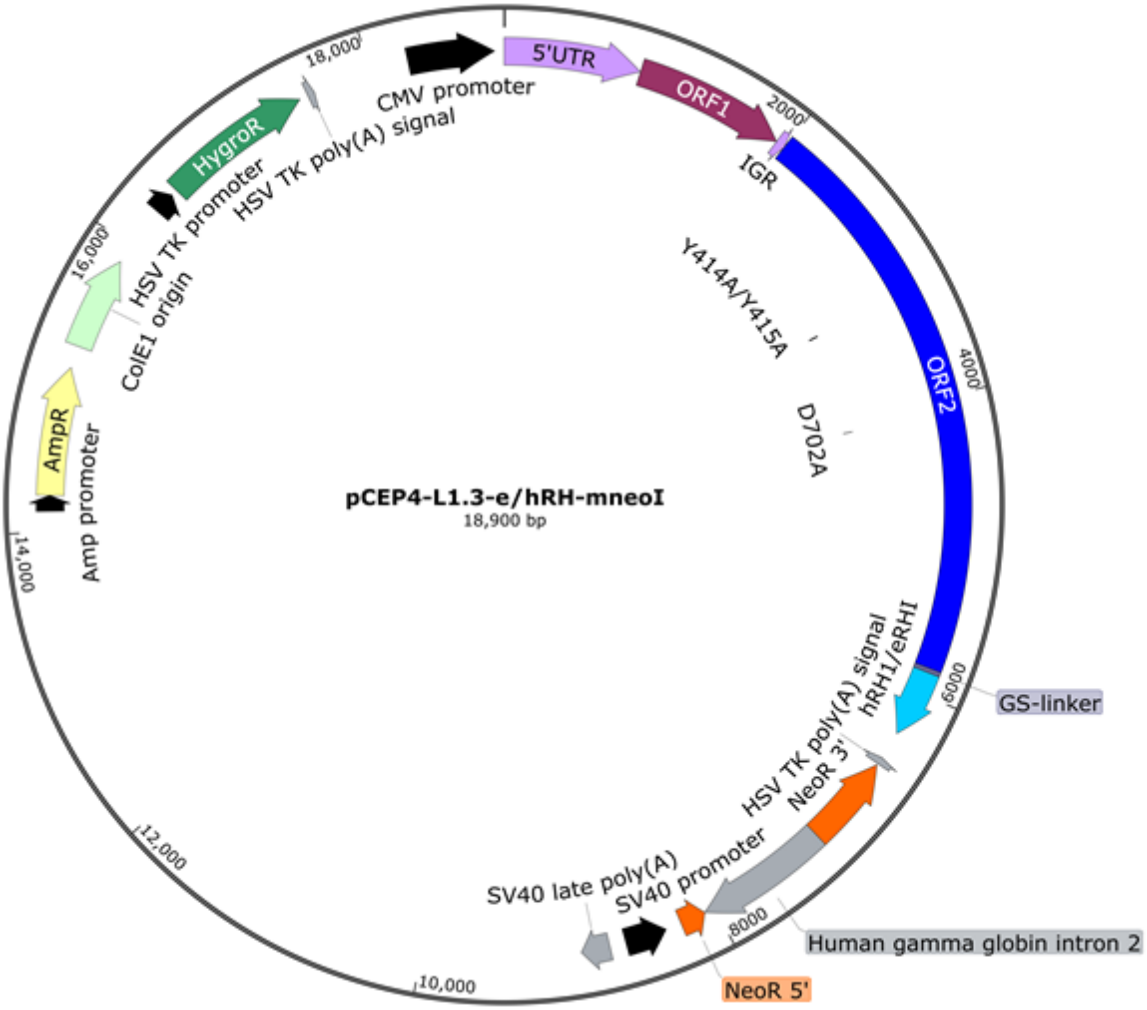


Figure 6.1. Vector map of LINE-1 ORF2p-RNase H chimeric constructs – The vector contains a human LINE-1, under control of a CMV promoter (black), with the coding sequence of a WT or catalytic dead mutant of *E. coli RNASEHI* (eRHI) or catalytic domain of the human *RNASEH1* (hRH1), fused downstream and in-frame to the C-terminus of L1.3-ORF2p by a linker sequence composed of 5× Glycine-Serine (GS). Annotated features: 5'UTR (light purple), ORF1 (dark purple), IGR (light purple), ORF2 (dark blue) with indicated mutations of the PIPm (Y414A/Y415A) and RTm (D702A) allelic variants, GS-linker (grey), RNase H domain (light blue). The LINE-1 is tagged with an antisense, intron-interrupted neomycin-phosphotransferase gene retrotransposition indicator cassette driven by a SV40 promoter (black) and ending in in a Herpes Simplex Virus Thymidine Kinase (HSV TK) poly(A) signal (grey). The intron is in the antisense orientation to the indicator cassette. Annotated features: Neomycin-phosphotransferase gene split in 2 exons (orange) by the human gamma globin intron 2 in the opposite orientation (grey), HSV TK poly(A) signal (grey). The final SV40 late poly(A) signal facilitates expression of the whole construct. Other vector features include Ampicillin (AmpR; yellow) and Hygromycin (HygroR; green) resistance genes, including regulatory features. ***Table 2.4***: pMK001 - pMK012.

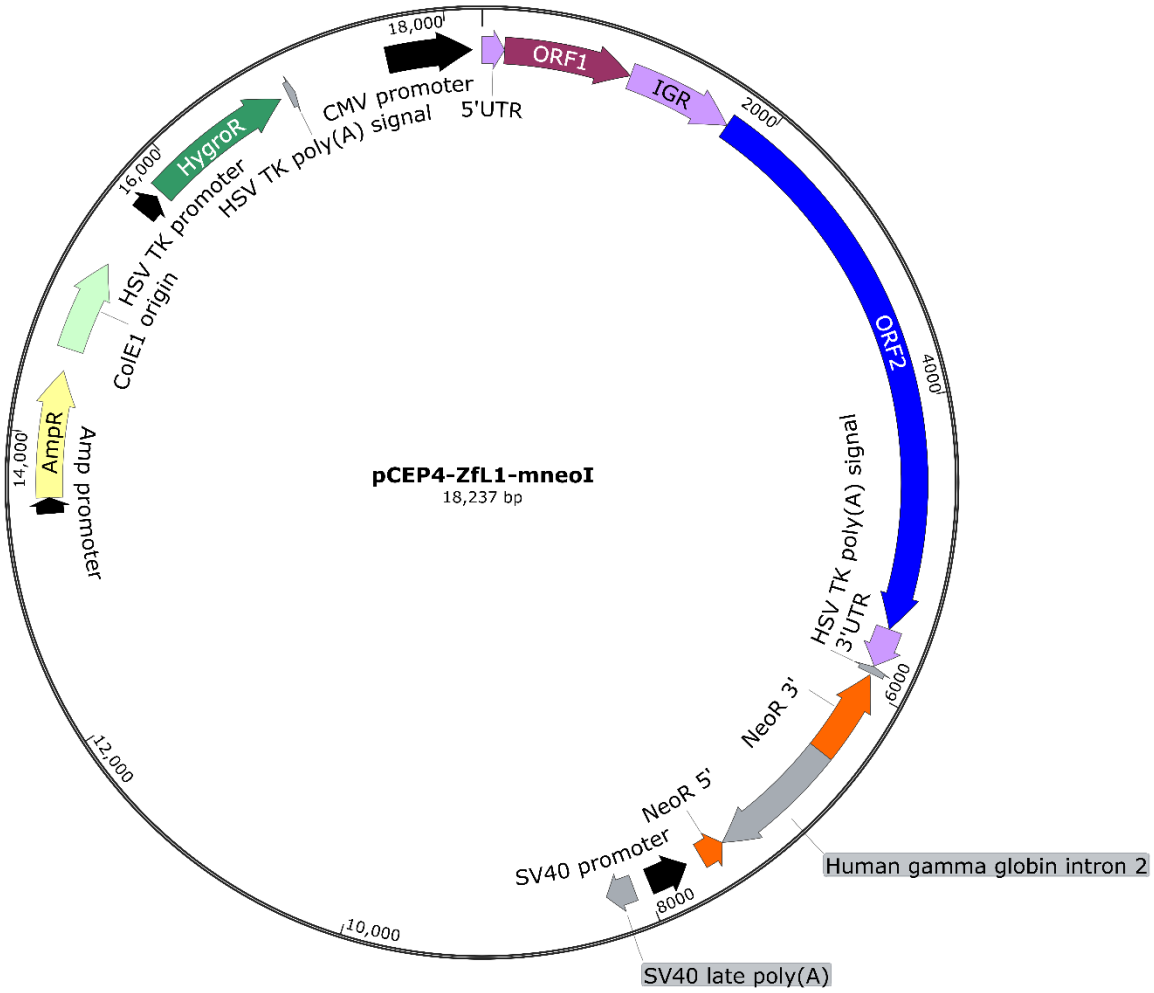


Figure 6.2. Vector map of a ZfL1 element in pCEP4 with 3'UTR before retrotransposition indicator cassette – The vector contains a ZfL1 element, under control of a CMV promoter (black). Annotated features: 5'UTR (light purple), ORF1 (dark purple), IGR (light purple), ORF2 (dark blue), 3'UTR (light purple). The ZfL1 is tagged with an antisense, intron-interrupted neomycin-phosphotransferase gene retrotransposition indicator cassette driven by a SV40 promoter (black) and ending in in a HSV TK poly(A) signal (grey). The intron is in the antisense orientation to the indicator cassette. Annotated features: Neomycin-phosphotransferase gene split in 2 exons (orange) by the human gamma globin intron 2 in the opposite orientation (grey), HSV TK poly(A) signal (grey). The final SV40 late poly(A) signal facilitates expression of the whole construct. Other vector features include Ampicillin (AmpR; yellow) and Hygromycin (HygroR; green) resistance genes, including regulatory features. ***Table 2.4***: pMK036 - pMK046.

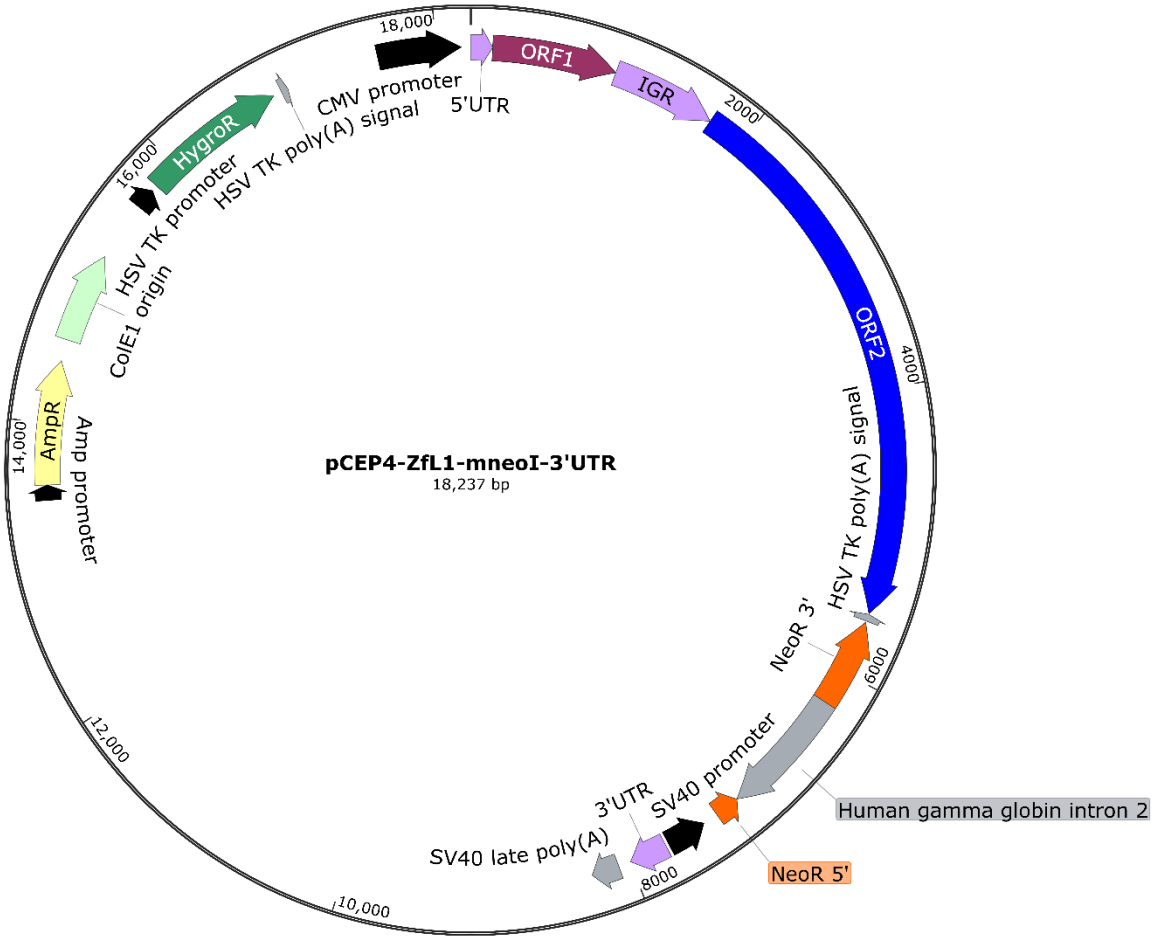


Figure 6.3. Vector map of a ZfL1 element in pCEP4 with 3'UTR after retrotransposition indicator cassette – As Figure 6.2, but containing the sequence of the ZfL1 3'UTR behind the retrotransposition indicator cassette. ***Table 2.4***: pMK047 - pMK057.

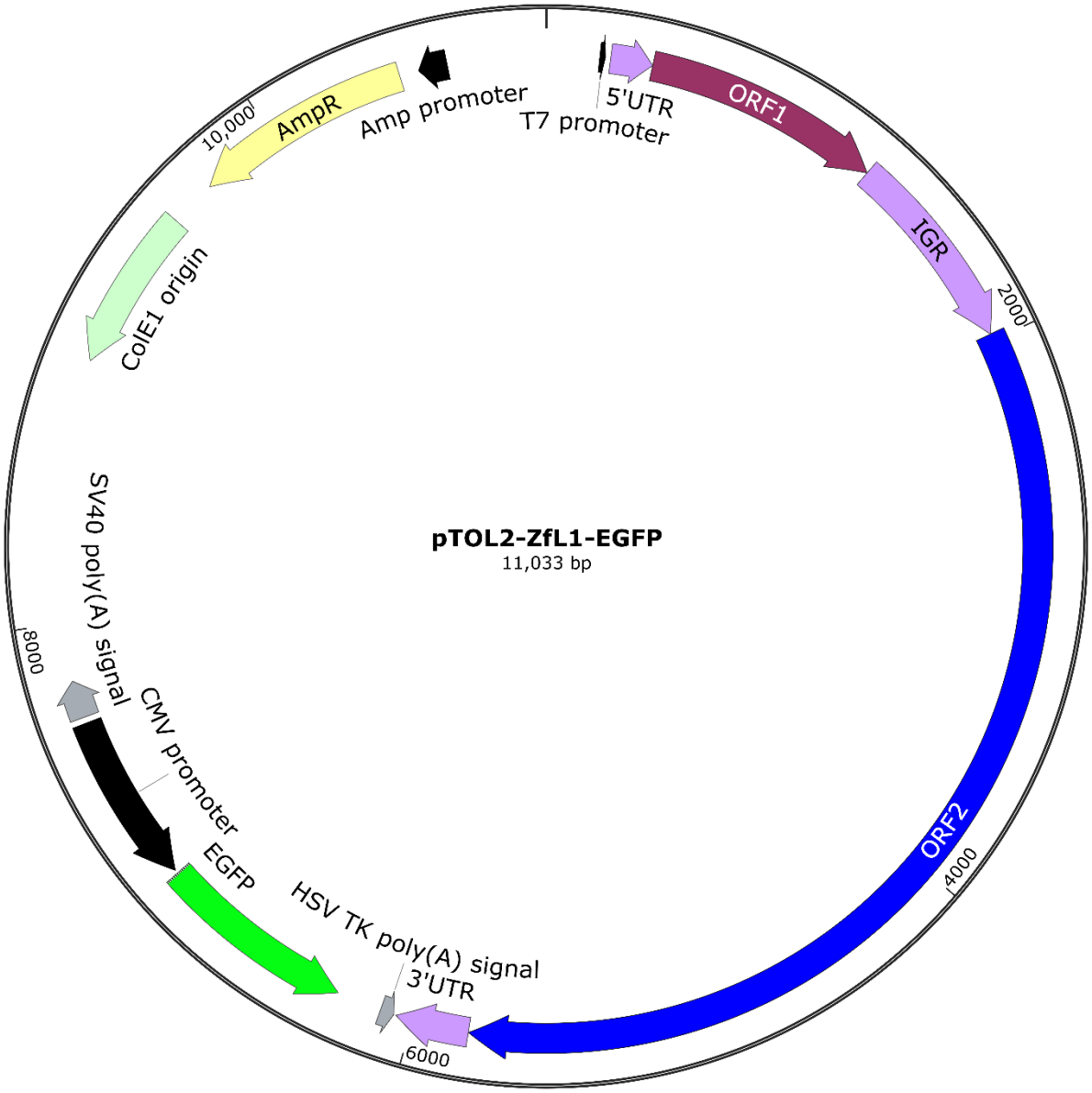


Figure 6.4. Vector map of a ZfL1 element in pTOL2 with 3'UTR before retrotransposition indicator cassette – The vector contains a ZfL1 element, under control of a T7 promoter (black). Annotated features: 5'UTR (light purple), ORF1 (dark purple), IGR (light purple), ORF2 (dark blue), 3'UTR (light purple). The ZfL1 is tagged with an antisense, EGFP gene (EGFP) driven by a CMV promoter (black) ending in in a HSV TK poly(A) signal (grey). The final SV40 late poly(A) signal facilitates expression of the whole construct. Other vector features include Ampicillin (AmpR; yellow) and Hygromycin (HygroR; green) resistance genes, including regulatory features. ***Table 2.4***: pMK014 - pMK024.

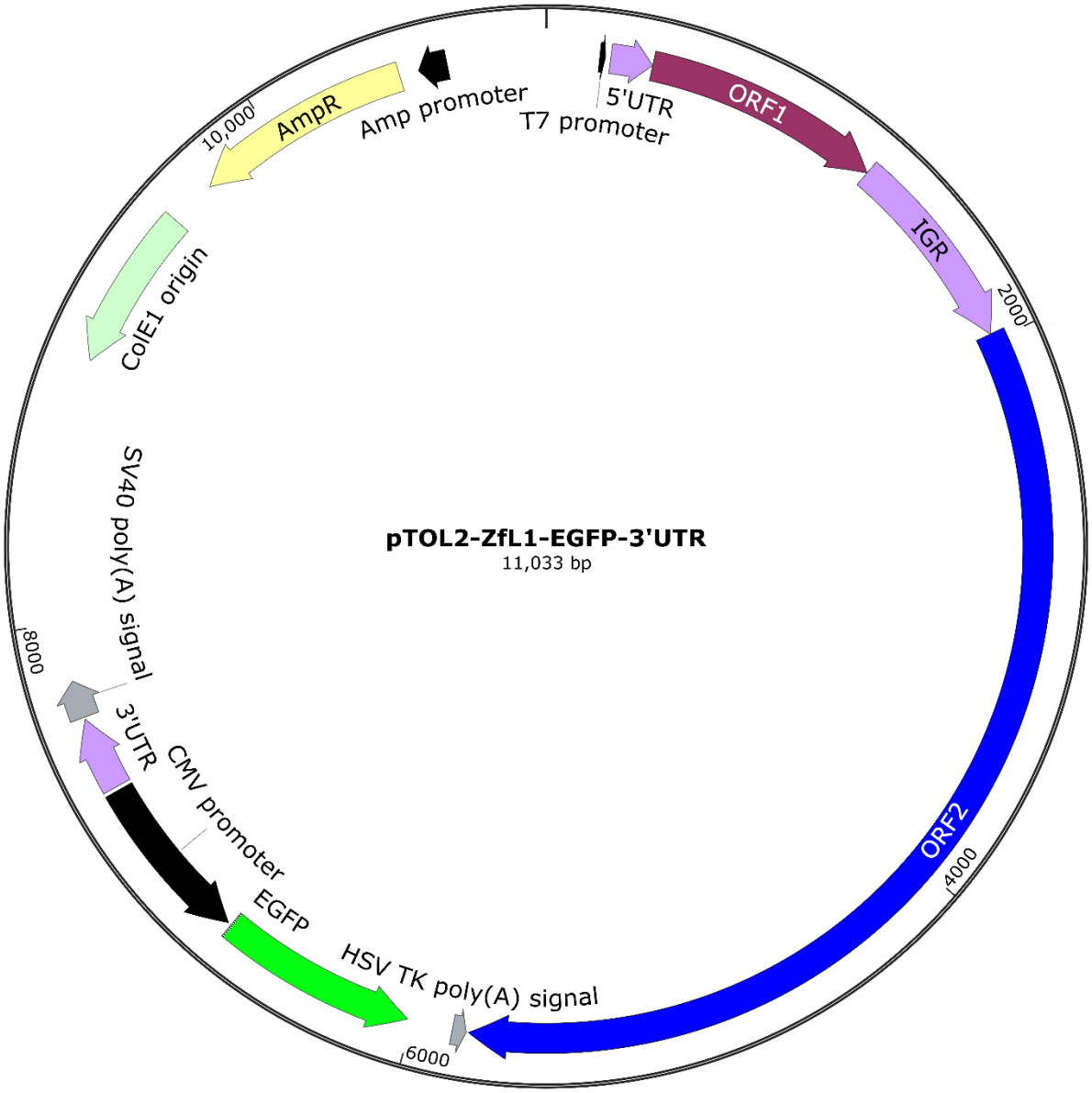


Figure 6.5. Vector map of a ZfL1 element in pTOL2 with 3'UTR after retrotransposition indicator cassette – As Figure 6.4, but containing the sequence of the ZfL1 3'UTR behind the antisense EGFP gene. ***Table 2.4***: pMK025 - pMK035.

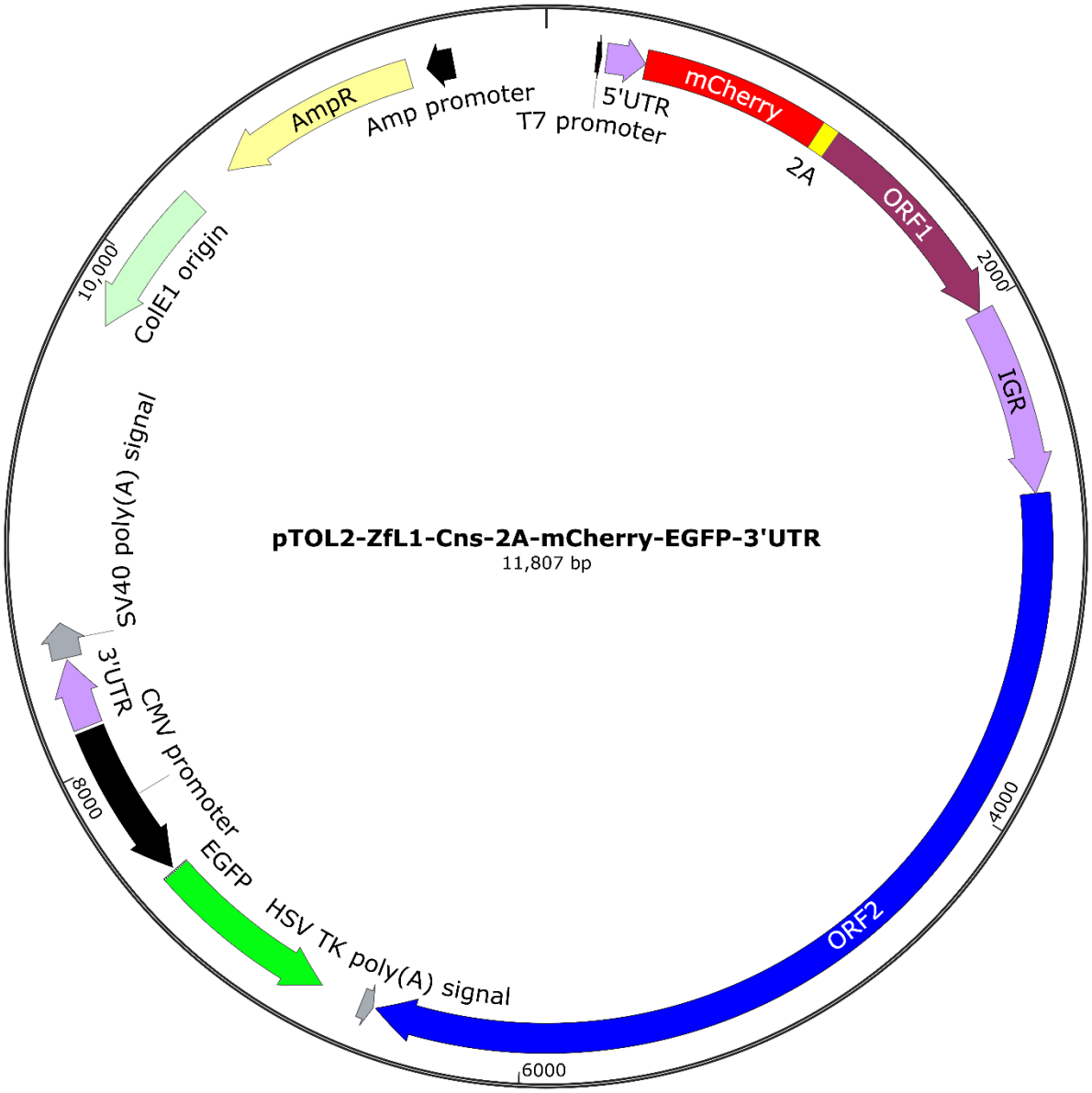


Figure 6.6. Vector map of a ZfL1 element in pTOL2 with mCherry fused to ORF1 with 2A sequence – As Figure 6.5, but containing the sequence of a mCherry fused in-frame to the 5'end of the ORF1 with a self-processing 2A sequence. ***Table 2.4***: pMK058 - pMK060.

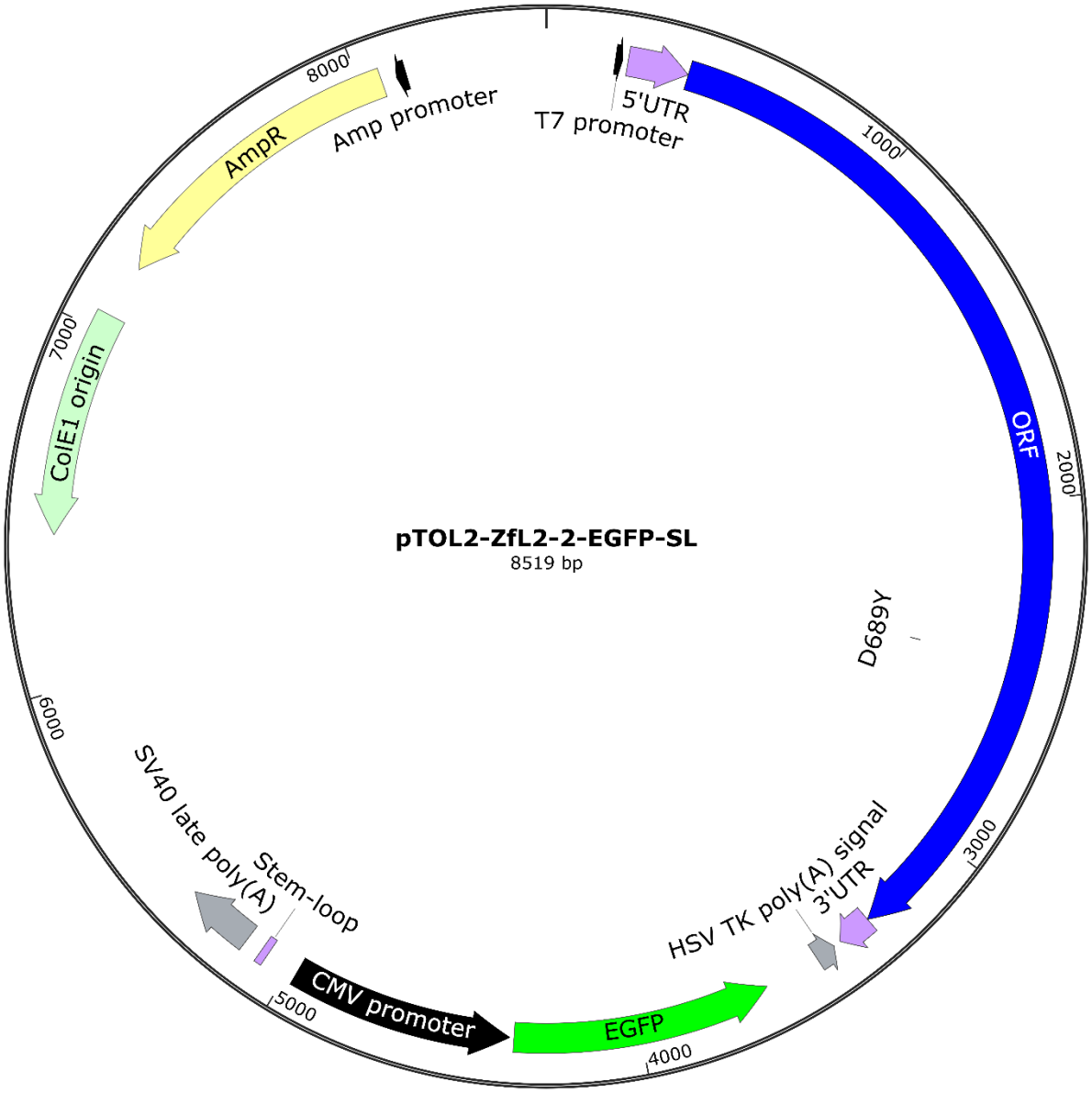


Figure 6.7. Vector map of ZfL2-2 in pTOL2 – The vector contains the ZfL2-2 element, under control of a T7 promoter (black). Annotated features: 5'UTR (light purple), ORF2 (dark blue) with indicated mutation of the RTm (D689Y) allelic variant, 3'UTR (light purple). The ZfL2-2 is tagged with an antisense, EGFP gene (EGFP) driven by a CMV promoter (black) ending in in a HSV TK poly(A) signal (grey). The final stem-loop ensures integration of all the sequence upstream, and the SV40 late poly(A) signal facilitates expression of the whole construct. The vector also include Ampicillin (AmpR; yellow) with regulatory features. ***Table 2.4***: pMK0013, ***Table 2.3***: pTOL2-Zfl2-2-EGFP-SL.

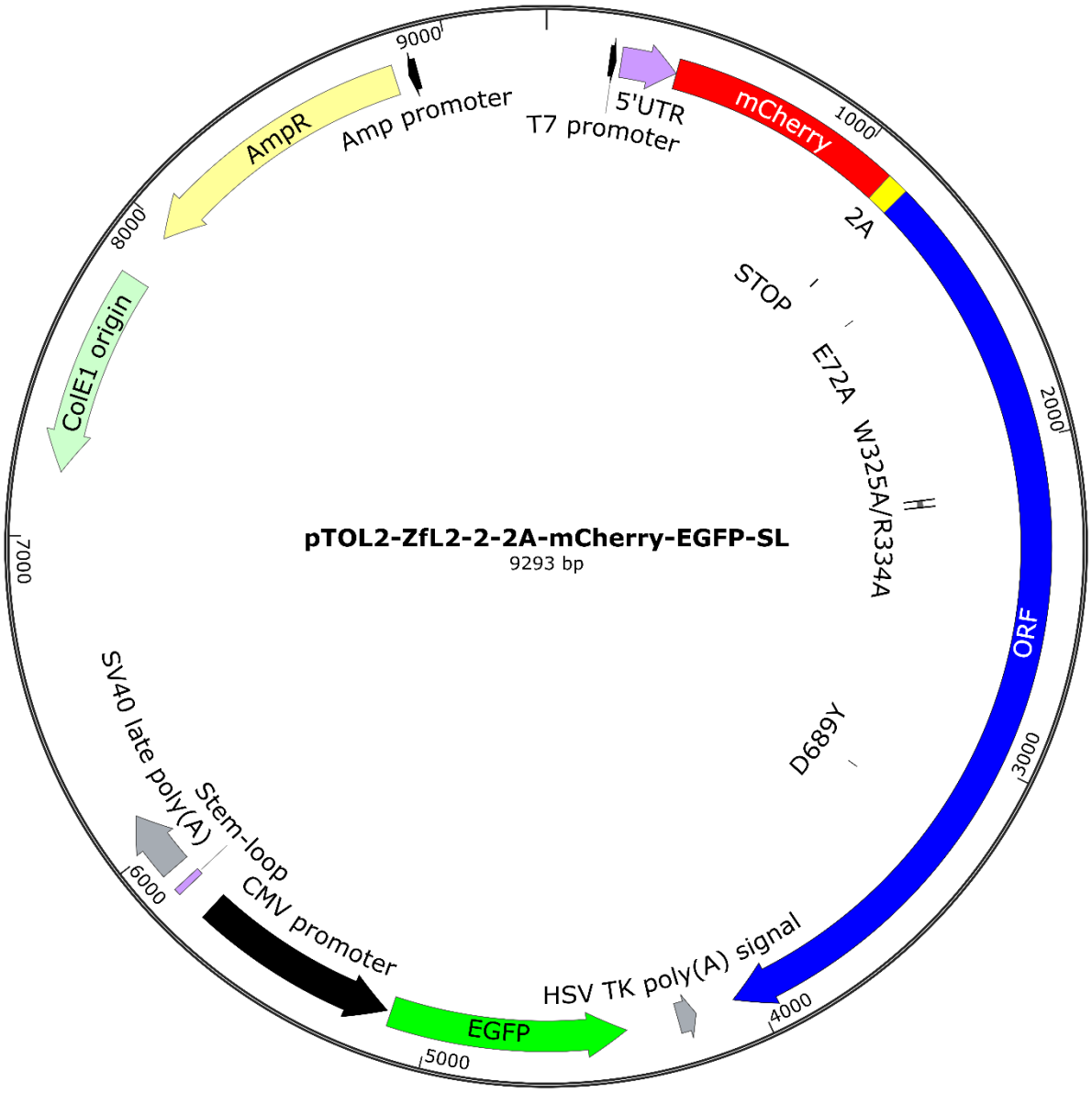


Figure 6.8. Vector map of ZfL2-2 in pTOL2 with mCherry fused to ORF with 2A sequence – As Figure 6.7, but containing the sequence of a mCherry fused in-frame to the 5'end of the ORF with a self-processing 2A sequence, with indicated mutations of the ZfL2-2-2A-mCherry-STOP, ENm (E72A), TBR2m (W325A/R334A) and RTm (D689Y) allelic variants. ***Table 2.4***: pMK061 - pMK066.

6.3. Sequences

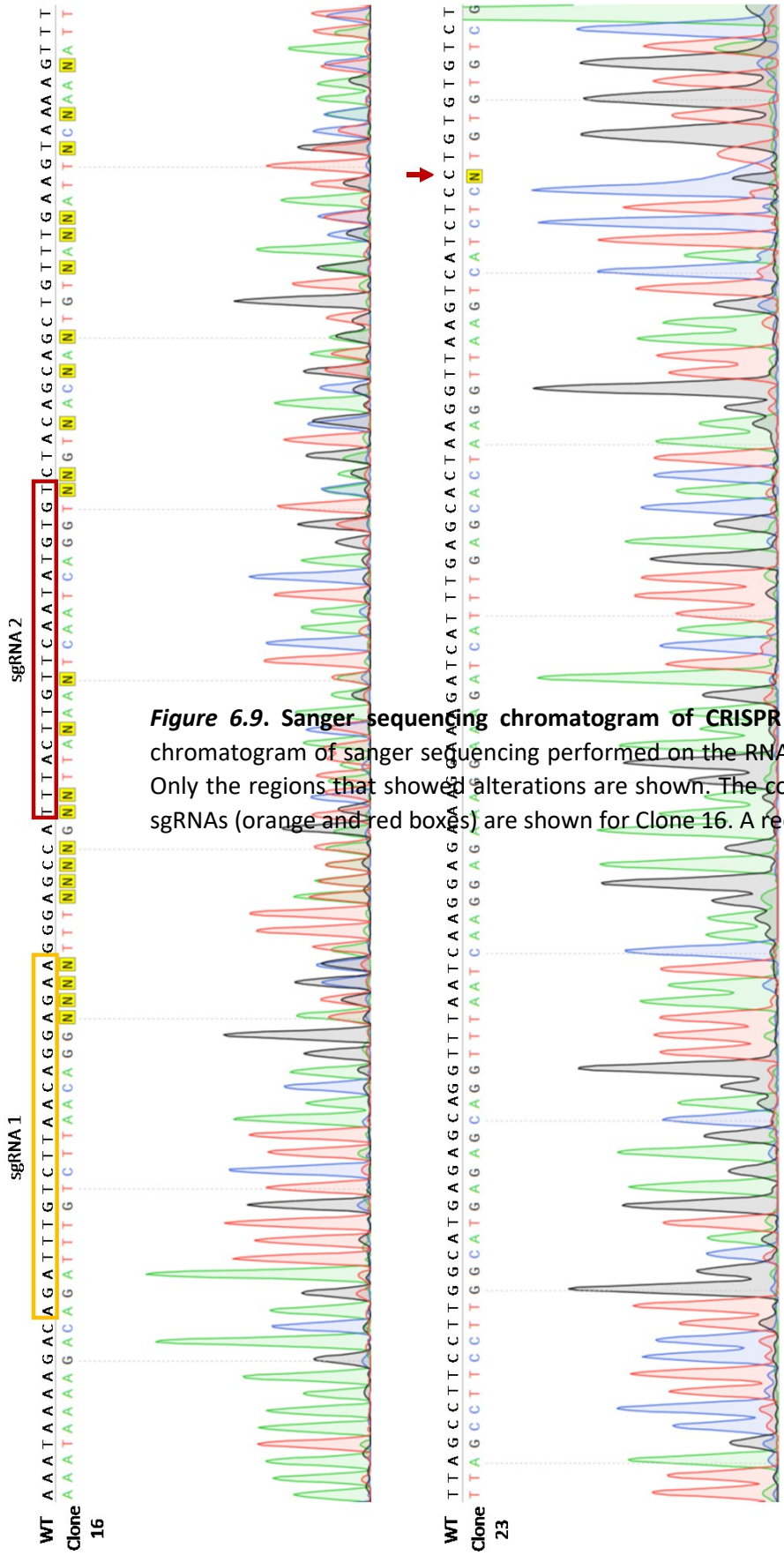


***Figure 6.9*. Sanger sequencing chromatogram of CRISPR targeted *RNASEH2B* gene region of tw**
chromatogram of sanger sequencing performed on the RNASEH2B CRISPR targeted region of the Con
Only the regions that showed alterations are shown. The corresponding WT RNASEH2B sequence in c
sgRNAs (orange and red boxes) are shown for Clone 16. A red arrow above the sequence of Clone 23 in

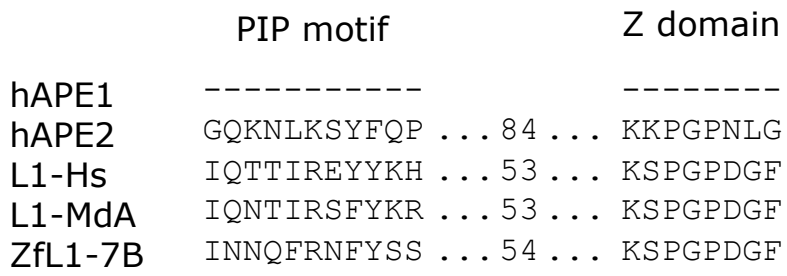


Figure 6.10. Alignment of PIP motif and Z-domain in LINE-1s and human APE1 and 2 – An alignment was done for the human APE1 and 2, to the ORF2p of a human LINE-1 (L1-Hs), mouse LINE-1 (L1-MdA) and zebrafish LINE-1 (ZfL1-7B), using the alignment tools PSI-Coffee and Clustal Omega. Shown is the alignment for the PIP motif and the Z domain found in the different proteins. The PIP motifs aligned perfectly using PSI-Coffee, while the Z domains of the LINE-1s aligned well to the beginning of the Zf-GRF motif of APE2.

>Functional_consensus_ZfL1-7B
AGACTTCCGGTTATGTCGCCTTTGGTATAGACGTGAGTGCGATCAGCTCCCTTTAACTTTTTATTTCTTCTTTAC
AAACCTCGTTGGTTTAGAACTTTTTGGCAACCTACATTTATTTACTTTTCTCTCTTAGAGTACGTAGTATTAAAT
TTTCACTTAACATGTCGAAAGGAAAGGGCAAACAGGTGGAAAAAGACCGTGAAGCGCCGTCCTCGCCTAGCTCTG
GTGTCACGCTAGCCGAAATTAGCGCCCTGCACGAGGAGCTCAGAGCCTCTCTGGCCGCGGATTTTAAAGCCTCTT
TTGAAAGCTTGGACAGTAAGTTGGAGAATATTAACTCCACAGTCGCGGCTCACGACCAGCGCATCAGTTGCCTTG
AATCGGGACTCGCGGAGGTGACCCAGCACCTTGAACACCTGGAGACCGCGTGCGCCGGCCTCACTAAAGAAAATG
AGTGGCTGCGGTCTAAAATGTCAGACCTAGAGGGGCGCAGCCGGCGGCAGAATATTCGCATCGTGGGGCTCCCCG
AGTCAATCGAGGGACCGCGACCAACTAACTTTTTCTCCGAAATGCTTGCTGAGGTATTTGGCGATCAGATTTTGG
CGTCACCGCCTGAACTGGACAGAGCGCACCGAATTCCCGCTCCCAAACCTGCCCCAGGTCAGAGACCGCGCCCTG
TGGTTTTGCGGTTTCACCGTTACCAGGTGAAGGATTTGGTGGTTCGTGAGTCCCGTAAAAAAGGAATTTTGACCT
TTCGTAATCACAAGATTCGGATATTTGAAGATTACAGCCCAGATGTCCTAAAGCTGCGCAGTGAATTCAAAGATG
CCATGGCGGAGCTATACAAGCGCGGACTCAGACCTGCTCTTCTCTTTCCAGCTAAACTACGCATCACTCTACCAA
ACGGTGAGAAACTATGGCTGCTGACAGCTAACGAAGCCAACAACTACATTCGGAACTTAGACGTATAGTGGTTCA
CCTCTATGCTTTGGACGATGACATGACGTGACATAAACCCATCTCTTGTCATAAACAGCTTGGCGCTTTTTTTTC
CCCTTTTTTCCGTAACACTCATGCCACTTCAAGGAGTTACTACTGGGTGACTAACGTTATTCGATGTACTCTCAC
AGAAGTTTGTTCGTCGAGATGCAATGTGAAATTTTTTTTTTCTCTTTTTTTGACAATTCTATTATTGCTTTTTGA
TGAAAATGGTGACTCGTTCTTTTCACCTTTAATATTTTTCATGGACGATGATATATCATCGTGTTCGGGGATAAA
GACTTGTTAAACTGTATACAATGTTTGATACCAACATACGATTACGTCATAGTTCAAAATATGTTTAATTAATCT
ATTTTTAAGTCAGTTATTCATACGGTTGTATGTAGTAGTGGTACAAATACAACTATGAGGTTGTGTGGTAAAATA
TGGGGAGGGATTGTATAAAGGAAGAAGGAAGCTTGCAACAGTGTTGGTTGGAGATATGTTTGATGTAGATGTTCT
TGCCTACATCACTTGTTGGAGTTGTGTGTGTGGGTTAGGAGGGTGGGGGCGGGTTGGTGTTTGTAGTTGTGTGTT
TCTTTTTTTTTTTTTTTTTTTCTTTCCCTTTTCTGTTCTTTTTCTTCTTATGTTTTTTTTTTTTTTTTTTTTT
TTTTTTTTGCGGGGACCAGAGAAAATTTGTGACAACTAGTTTTTAATGTCTCCTGTGTTCAGTGGGACACCTACT
GGTGTACTTACATTCTCCAGCTGGAATGTCAAAGGGCTTAACTCACCTATAAAACGTAACAAAGTTCTCAACCAC
TTGAAGCAGTTAAACACTAAAATAGCTTTTATCCAAGAAACTCATCTTACAAGTGCTGATCACTTAAAAATTAAA
AAGGACTGGGTAGGTCAGCTGTATCATTCCTCGTTTTCCAGTAAGTCACGAGGAGTGGCTATACTTATACATAGA
TTAATTCCATTTTCTGTATCTAAAGTAATTTCTGATCCAAATGGACGTTTTATCATTATTACTGGCCACATTAAT
GGTATTCATCTGGTATTAGCAAATATATATGGACCAAATTGGGACAATGAAGTTTTTTTCAAGAATCTTTTATTT
TCCCTTCCTGATTTAAACACTAGCCAACTTATTTTAGGAGGAGATCTTAACTGCTGCCTGAATCCTACTCTAGAT
TGTTCTTCATCAAAACCCAGAGCTATTTCAAAATCAGCAAAAACTATTGAATCGTTTATGGAACAATATGCAGTT
TCTGATGCCTGGCGTTTTCTAAACCCTAACGCAAAACAGTTTTCTTTTTTTTCACCTGTCCACCAAAGCTTTTCT
CGGATTGATTATTTTCTCCTTGATAATAAATTAATTCCATGTATCAAGAATTGCTCCTATAGTCCTATTGTGATA
TCTGATCATGCTGTGCTTACATTGGATCTATCATTTCCTTGTGGGACAACTCCTAGGTCTTTCTGGCGCCTAAAT
ACTCTATTACTTAATGACCCAGATTTTGTGCAAATTATCAACGATCGGATTGACCTTTTTATTTCTACAAATGTC
ACCCCTAGCATTTCAGCTAGTATTTTATGGGAAGCTTGTAAAGCCTTTCTGAGGGGAGAAATCATATCTTATTCA
GCACATCAGAGAAAAAGGGCATCTGAACAGAAGTCCTCTTTAATTAAATATCTAGCGAAGCTTCAGTCCAAAATG
GAAACCTCACCAACTCCAGATTTGTCTAAAGAAATACTTACAGCAAAAACTGAATTTGACCTGCTGACAACGAAT
GAGGCCACTGAAGCTCTTCATAAAAGCCGTTATAATTACTATGAATTTGGCGATAAACCTAGTAAACTACTAGCA
CACCAGCTACGCCAGTCCTCTGCTCTAAACCATATAACTCAAATTTCAACTTCTTCTGGACCCACAACTGACCCA
CTTCTCATTAACAACCAATTTAGAAATTTCTACAGCTCCTTGTACACCTCAGAAAGCCTGAACGAAGAATCACAC
CTTGATAGTTTTTTTCATAATCTTGATATTCCCCTTATAGATCCTGACTTGGCCTTAAAATTGGATGATCCCATC

ACTCTCGATGAAGTTAAGACAACACTGAGCCTTTTGCAAAGTGGAAAGAGCCCAGGTCCTGATGGTTTTCCAGCT
GAATTTTTTAGAACATTTTCAGAGAAATTATCACCTCTTATGCTGAACATGTTTAATGAATCTTTTAAATCTGGT
ATACTTCCACAAACTTTACGCCAAGCCTCCATCTCCTTAATTCTTAAAAAAGATAAGGACCCCCTTCAGTGTAAT
AATTATCGTCCTATTTCTCTACTGTGTGCAGATGTAAAAGTACTTGCTAAGATACTAGCATGTCGCTTGGAACCA
CTACTACCAGCTATAATCTCCCCAGATCAGACTGGTTTCATTAAAAATAGGCAGTCTTTTCACAACATCAGACGT
CTCTTAAATATTCTTTACACTAGCCCACATAAAAAGACTCCAGAGTTAATTATATCCATGGACGCTGAAAAGGCT
TTCGACAGAGTTGAGTGGGATTTTTTATTTTACACTCTTAAAAAATTTGGAGTGGGTGATTATTTTATATCTTGG
ATAAAGCTTTTGTATACTTCCCCTTTAGCCTGCGTACGCACAAATAATTACTACTCAAAATACTTCCCTCTTGGG
AGAGGAACAAGGCAAGGATGCCCACTCTCCCCTCTTCTGTTCGCTCTGGCTATCGAACCTTTGGCCATTGCAATT
AGAGGCAGTTCAATGTTGGGCATCACTAGGGGGGGAGTTGACCATAAATTATCCTTGTATGCAGACGACCTTTTG
CTGTTCATTGCTGACCCGGCCAGCTCTATCCCAGTGGTGCTTGCATTGCTAACAGAATTCGGCAAAATTTCAGGC
TACAAATTAAATCTGACTAAAAGTGAAGCAATGCCTATTAATAATGCTGCCTCTGAATACCCTTTATCCTCCTTG
CCATTTAAGGCCACTTTACACAGCTTCAAATACTTGGGTGTACAGGTGACAAAAAAATTTGAAGAGCTGTTCGAC
CAAAATTTTTCTCCATTGCTTAGTCGTCTCACACATGATTTTCGCCGCTGGTCTCTTCTCCCATTATCAATGGCA
GGCAGAATCAGCTGCGTAAAGATGAATGTACTCCCTAAATTCTTATACCTTTTTCAGTGTATTCCAATTTTTATC
CCAAAATCTTTTTTTGCTACATTAGATAGTTCTATCTCTCAATTTTTGTGGAATGGTAAGACCCCGAAGATTCGA
AAGGACATTCTACAGAAACCTAAGGGCTTGGGTGGGATGGCCCTCCCTAATTTTTTGTCTTATTATTGGGCTGCG
AACATACGATCCTTATTATATTGGAAAAATAAAAAAAATCTAGACGGTCCCCCAGCATGGTTAAATATTGAGAAC
GCCTCATGTAGATCTTCGGACTTAAAAGCTCTCCTCTGCTTTCCCATCACGCTATCTCCACTTAAATACACTGAT
AATATAATTGTGAAAAACAGTCTTAAAATTTGGATTCAGTTTTTACGTAACTTTGGCTTTCAATCTACCCCTTTA
ACCGCCCCAATACATCTAAACCCATTGTTTTCTCCATCTATATTGGATGGAGCATATGCCTCTTGGAAGAATCAA
GGGATACTCTCAGTAGCAGATCTTTATATCGAAGGTGTTTTTGCCTCATTTGACCAACTGAATGCATTGTATAAC
ATTCCGAGAAATCATTTTTTTAGATATTTACAAGTACGTGACTATGTTCGTAAACATTTCCCTGGTTTTCCCAAT
GTCCCTCAACTATCGAAGATCGACAATATTTTATGTCTTGAAGCAAGCAGTAAAGGCCTAATCTCTAAACTAATC
AGTATTTTTGCAAATTTACAGTCTTTTCCAGGTGACAGTCTCCGCACCAGCTGGTCTGCGGACCTTAATATTGAA
ATAGATCAAGATACATGGGAGGCAGTTCTGGACAGGGTTCATTCTTCATCTATATGCGCACGTCACTGCCTCATT
CAGTGTAAAGTGGTTCACAAAGTACACTGGTCGAAATGTAGGCTCGCACGGATCGACCCAACTATTGATCCAGAA
TGTGATAGGTGTCATTTGGGTTTAGGTACTTTAACCCATATGTTTTGGACATGCCCAGCTCTTTCCTCATTTTGG
GGCTCAGTATTTCAGTCTCTTTCAGCCATTACTTCTATAGATATAACACCATCACCCATTATTGGCCTGTTTGGC
GTGATGCCGGATAATTATGCATTACCCACTTACCTTTTAAACTTCGTGGCCTTTCTTACTTTGCTTGCTAGGCGT
GTCATTTTATTATACTGGAAGAACCCTCAGCCACCCTCGCATGTACGGTGGCTGAAAGATGTTTTGTATTTTGTA
AAACTAGAGAAGATTAGATATTCCCTTCAAAGACCTAACAGAACATTTGCCAAAATTTGGGATGCCCTTTCTTTA
TATGTTAAAACTCTCCAGCTGGATCTAGACTGATCACCTCAACTTTTGATCACAATAGTACATCAGTTTGTAATC
TTTCATCACTCGAAATGTAACATTTGAGATTACTGTATAAATGTAGACATCTGCCTCGCATTAATTGTTTATTTC
ATTTTATTATTATCTTCTTTTTTATTTTAATTTATTTATTCACTTTTTTTATCAATATGTTATTTGTATTTATAT
TTATACATTTTTTTCTATTATTATTTATGTAAGGGTTGTTTGATATGTTTGTAAAATGGAAAAATGTTAAATAAA
TCATAAAAAAAAAAAAAAAAAAAAAAAAAAA

Sequence 6.1. The functional consensus sequence of the ZfL1-7B build from the full-length copies annotated in the zebrafish reference genome – 5'UTR, IGR and 3'UTR (light purple), ORF1 (dark purple) and ORF2 (blue).

>Functional_consensus_ZfL1-10B
TGGACTTCCGGTTATGGCGAGGTGCTGAGAGGTCGAAATTTTTTAGCTCTGCTGCTAACTTGCGATCAATCCTAC
AATACCAACCAGAATTATATCTTAAGTCAAGCAACACAAATATAAAAGCATAAGGAAGAAAAATGTCACGCAAAG
GTACCAGAAAAGACAAGAAAAAGACCAGAAACGAAGATGAATATTCTGAGGGTGAAGAAACTAACGACGCCATGG
CAGCTGGTGAGGGCCCTGACGAAGATGATGACGAAGCAATAACTTCGGATGGTAACAACGAAGATATCATGAAAG
CTATAATGTCCTTGAAAAGTGGACTTTACAAAAAAATTGATGGAGTCCAAGCAACAATTACGACGATAAGTAAGG
AGATAAAGGAATGCACAGGGCGGATAGCACAAGCGGAGCAACGCATTTCGGATGCCGAAGATAATGTCAACGGGC
TGTTATCCAGAGTGAGCACGCTAGAAAATACTGTCAAAACGCTTTCAAACAAAGTGGAAGATTTGGAATGCAGAA
GTCGACGGAACAACATAAGACTTGTGGGACTACCGGAGAAAGCGGAGGGCCAGGATACGGTAACGTTTTTGGAGA
AGTGGATACCGGAGGCTCTAGAATTGGATTCGCGGGAATCTCTTGTAATTGAGCGGGCTCACAGAATCGGCACTT
TGACAAATATCCACTCGCGAACGACACGCCCGAGGACACTAATTATGAAGTTCCTGAACTTCAGAGACAAGGAGC
GAGTGCTAAGGGCAGCCAGAGTTAAAGGGAATGTTCTCTACAATAACGAACAAGTCCGTTTTCACCCAGACCTGT
CCGCCGGCATACATAAGATGCAGCGCGACTATGATGACGTGCGAAAAAGGTTGCGAGATAAGGGAATTCAAAAGC
ACAGAATTATTTTTCCAGCACGGCTCCTACTGACACACGGAGAAAGATCCTACACCTTCCAAACTCCGGCTGAAG
TAGACAACTTTATCCAGTCTCTTTGAAGGATCAAGACAATCCTACCAATATACGAAACTAAAACGTGACAACACT
ACTGACCAATTCGTTTTATGTCTTTTTTTTCCAACGAGAGAGCGCTAATATTCTTTTTGCGTTCTCTATTGGCTC
ACGTTAGTCATATACTTTTTTTTCTGACCACATATGTTTATTCTCGTGATAAAAACTCTTAAAATGTACCTGGGG

TGGGAAAGGAAAACATGGACATAACGCACACTCAAATATTGGACCTTTTGACATTCGCAAGGGACGTTCATCCCA
CAACGAACTTTTCACGGCCAATAGCAGTACTGTTTGATTTGTTCTTTTATTTATATCTATCTTTTTTGCTAATCC
TTATTCAACTCAACCCCCCAGGTATGTCATTATTTGTTAAAAGTTGCAAAATATCAAAATTTTAATGGTTATGTT
TACTAATGTTGCTCAGTTCAAAACAACTAAGTAAATAATGTTGGCAAAGTAGTTAGTACTTGATAATAGCAGAGG
GAGCTAAAGAGAAAGAAATAATCAAAAGTGAGAGACAGACGGCTCTTTGTATTGCCTAGACCCGTATCTGTCATA
AGCGGTTGGTATAAGGTGTGTTTTTTTTTCTTTCCCTTTTTTTTTCTTTTATTTATTTATTTATTTTATTATTTT
TTTTCTTTCCTTTTTTTCTTGTTGAGTGTGGGTATGTTTTCATCAAACAAGCAAATCTGTAAGGTTATATTAGAC
AACTGATGTCTGGTAGCTTAAAGGTGAAATTTACTTCATGGAACTGTAGGGGTTTAAACAAAACTTCAAAAGTTA
AACAAGTGATGAACAGGATAAAATTCTTGCAATCAAAAATAATTTTCTTGCAAGAAACCCATCTAGTAAGTGAGG
ATGTTTCTAAAATAAGAAGAAGATGGCAGGGTCAGGTGCTATCTGCCCCATATACCACCCATGCTAGAGGTGTTA
TAATACTAATTCATAAATCTGTTCCTTTTCAAATTAAAAATATAATTAAGGATTCTGCAGGTAGATACATCATAG
TCCAGGGTATGCTGTTATTTGAAAATATAAATCTGATAAATGTTTATGGACCCAACGAAGATGACCCTAAATTCT
ATAACAATTTGTTTCTGACTATCTCCAATCTCCCGGGGCAAAATATAATTGCTGGTGATTTTAATTGCACGTTAG
ATCCAGCAAGAGATAGGTCATCAGGTTCAGATAATACCCACATTAGATCGAGAAAAACAATAAATAGTTTTATTA
AAGAATTAAATTTGGTGGACATATGGAGACATAGTAAACCAAATGCAACAGAATACTCATGTTATTCCAGCACTT
ATAAAACTTATTCACGCATAGACTATTTTTTGGTTTCAGCACAACTGGTTTCAAAAATTGATGACTGTAATTATG
GTAGTATAGTAATATCGGACCATGCAGCTATGTCACTCACTTATAAAGATGCTAAACTAGTGAAGGATTCTCCAA
GATGGCGCTTTCAACCAGGATGGCTAACAGACCCCACATTTATAGATTTTTTAGATAAACAAATTGACCTGTATT
TTACATGTAATACTTCTCAAACATCTGCTAGCATAAGGTGGGAGGCATTTAAGGCCTTTATTAGGGGTCAAATAA
TTAGTTTTACCAGTTCAAAGAAAAGGAGTGCACAACTGGAAATGAAGACATTAGATGAAAAAATTAAAACACTTG
AATCAGTTCTTTACAATAACAAGTCTAACTCTATAGATACAAATGCACAGTTACTCCTTCTTAGGTCACAGTACA
ATGAATTGTCGGCTAATAAAGCAGCAACTTATTTACTGAAACTTAAACAATCCTATTATGACCAAGGGGAAAAAC
CTGGGAAACTTTTAGCATGGCGCATTAAACAACAACAAACAGAAAGATCTATTAATTACATTGAAGATTCAAGCG
GAAGAACTATAGTAGATCCTAAAGAAATTAATGAGGCCTTTAAATTATTTTATGAGAAATTGTATAGCTCTGATG
GTAACCCCAGACTTGTTGAACAGACACAATTTCTAAATAACATTAATATTCCTAGCATTTCAGAAAATGAAAGTA
AAGCATTAGATGAAGAATTAACAAAAATAGAAATTTTAGAAGCAATCAACTGTATGCAGGCAGGTAAAACGGCAG
GCCCAGATGGCATACCTATAGATATTTATAAAAAATTTCAGTCTAAATTAATTTCACCTCTTTTGGAGATGTTCC
AGGAGTCCTTTAGGAATGGTCTTCTTCCTACATCCATGAGAAGTGCCCTAATAACTTTACTTCCAAAACCAGGAA
AACCAAATATAAAATGCGAAAATATGCGTCCAATCAGCCTTTTAAACTCAGATACAAAAATACTTAGCAAAGTTC
TTGCTAGAAGATTGGAGAGCTTACTCCCTCAGGTGGTGGGAGAAGACCAAAATGGCTTTATTCAAGGGAGACAGG
GCTTCCATAATGTCAGACGAGTGCTCAATATTTTACATAGCCAGAGAGGAGCGAAAGACACAGCACTACTTTCTC
TGGATGCAGAAAAGGCTTTTGACCGTGTCGAATGGCATTATTTATTTGACGTGCTTGCTCGCTTTGGCTTTGGTG
ATACATTTTGTAAATGGGTTAGGCTACTTTGTACGGAGCTCACAGCAGAAGTTTTGACAAATAACATTGTTTCTA
AACCCTTTAAAATATCTAGGAGTTGCCCACAGGGAAGTCCTCTATCACCTTTATTATTTATCCTTGCAATAGAAC
CATTCGCCATAGCAGTGAGAACGCATAATGCCATATATGGAATCCGAGAGAACCACTTGGAACATAAATTAGCAT
TATTTGCAGACGATATTATTTTGTTCCTAAAAAATCTAAGTAGCTCTATTCCTGCACTCCTGGATCTTATTGAAA
CTTTTGGAAAAATATCTGGATATAAAGTTAATCACTCCAAATCATCTATAATGTTACTTAATGAATCAGAGAGAA
AAAATAGCCAAGGTTATGCTTCTATTTTTAATTCAACGGATAATTTTACATATCTGGGTATTAAAATCATGGATG
AAATAAATAAGATTGCTCAAATTAATTATGACCCAATTTTGGATGCTACTATGTCATCAGTAGAACGTTGGACAT
CTTTACCTATCTCAATGATTGGCAGGATAAATATCCTAAAAATGAATATACTCCCTAAATTTCTTTATTTGTTCC
AAAACATCCCCTTGCCCCCTCCCTCATCGTTGTTTACTAGAATTAAGAAACTGTTTACCAACTTTATTTGGCAAA
ATAAACGTCCCAGACTACGTCTAACTTTATTATATTTACCATTCGATCGAGGAGGACTTAAATGTCCAAACTTTC
AGTGGTATTATTGGGCTGCTCAGCTGCGTACAATCATGTATTATTTTTCTTCTGAAAGTCCTCCTTCTTGGATGG
ATTTGGAATCCCGCTCTATCACACCACAAATACCACTGCACTTATACCTATATTCGGCAGACCTCAAATATCTAA
GAAAAAACACAGACAACCCCATTGTATTGAACATGATTAATGTTTGGTTTGATTCTTGTAAATATTTGAATATAA
ATCCTTCCTTGTCACGTTTTAGTCCTATTTGGGGTAATGTTAAGTTTAAACCAGGAAAAAGTGATGCAGGATTTC
GGATATGGGCTGAAAAAGGGTTAAGGAAAGTACAAGATATGTACAATAAGGATGAAATATTTATGTCCTTTGCAG
TAATTTCAAAAAAATATGATATTCCAAAGACGCATTTTTTTAAATATCTTCAACTTAGAAGCTTTATATCCTCAT
CTCAAAATCATTCACTAGATATTCCTAACATTTCCTCATTAGAGGCAGCAATCCTAGGACACTGTTATGATAAAG
GTTCAATTTCCTCTTTGTACAGATTGTTTGTGTCTGGCTCTAACGAATCCTCAGAATCAAAACTGAAGTTATGGA
AAGAAGATATGGAAGATGAAATATCTCTAGAAGACTGGAACGAGGCATGTAGGAGAGCTCAGATACAGACAATCA
ATAGTAGCCTGAAACTCGTACAGTACAAATGGTTAATGCGAACATACATAACTCCTGCTATATTGCACAAGTTTA
ATGATAATATTCCAGATACCTGTATAAAGTGCAATGAGGTAAGAGGAACACTTTATCACTGTATTTGGGAATGTG
TGAAAGTGAAACCCTTTTGGCAAGATCTTATTGAAATAATTGACCAAATTTTATTAAAGAAATTACCGTTGGATC
CTAAGATTTTTATCCTGGGTATATTTCCTGTCAGCCCTCTCTTACAAAGCAAAGAGGTCAGATTCATAAACATGT
GTATTTTACAAGCTAAACGTATTATTGCTCTTAATTGGAAAAATATAAATGAACCTAGAATAGGAAAGTGGATTA
AAGAAATGGCATCTCATATGTCAATGGAAAAGATCACGTATATTATCAAAAAGAAACAAAATGTGTTTGAAGATA
CCTGGAGACCTTTTAATAACTTTCTAAAGTACAATGTAAATGTGGACAGTCTACTACAAATAGAACAGGCTTAGG
AATATGGGAAGTATCATGGTCTAAACAGACTACCTGATTTAAACATATCCCCTCTTTTTTTTTTTTTTTTTTTT
CTCTTCCCCCTTCTTATTATATTTTTTTTAATTGTTTATTTATTTTTATTTTACTTATTTACTTATTATTGTTTA
TTTATTTATTTTTAAGGGAATGGGTTGTTGGGAGGGAGAATGGGGGGTAAAAAGTTTGCTGTGCCTTCATGATTA

TTGTTCTGTAATTGTGTAATCTGTTGTGATATGCAAATTGAAAAAAGTAATAAATATATTGTACAAAAAAAAAAA
AAAAAAAA

Sequence 6.2. The functional consensus sequence of the ZfL1-10B build from the full-length copies annotated in the zebrafish reference genome – 5'UTR, IGR and 3'UTR (light purple), ORF1 (dark purple) and ORF2 (blue).

>Functional_consensus_ZfL1-12B
GCTTCCGGTTCCGGCGAGTGAGGGAATGGCAGCATAGTGCATCGTCTCCTCCGACGAAGATATTAATAACCCCAC
TTAAACTTGAGAAAAACCCCTAACTGTTTCAAAATAACCGATAATAGTGGGGCAACAATGGAGACGAAAAAAAAG
AGCAGGAAAACCAACACGGAACTTGTGGATAGTGGGGCAACAATGGAGACGAAAAAAAAGAGCAGGAAAACCAAC
ACGGAACTTGTGGATGCTAATCGTCAACAGCCTAGTAACCCTGACAACACGTATATCAATCTGGAACGAATCCTG
GAGGAAATTCAAGACTTTCGGAAGGACAACAACCGCCAGTTGGATGAAATCAAAACCGAATTAAATAAGACTAAT
CAAAGAATCAGCCACGCAGAGGATCGTATTGAAGAAACTGAAACCCGATTGCAAAACGTGGAGGAAATCATGCAA
AAAATCATCAAAATACAATCGCAATTGGAAAACAAACAAATTGACCAAGAAGGCAGATCAAGAAGAGACAATATA
AGGATATACAACGTGCCTGAAGATGCAGAGAAAAACTCTATGGTTGATTTCGTAGAACAACTGTTGCGGGACACT
TTGGATTTTCCCCCAGACAGGGAGCTGTATGTAGAAAGAGCCCATCGGGCGCTCGTGCCGAAACCCGGAACGAAT
GCAAAACCTAGATCGATCATAGTAAAATTCCTCAGATACAGAACGAAAGAAGAAGTGATCCGGAAGGCATGGGCG
AAAAAAGAAATCCTCATGGGAGAGCAAAAAATATACTTCGATCATGACTATCCTCCCGCGGTTCTCCAGAAAAGG
AAGGATTACTCACAGGTGAAAAAAGTACTTAAAGAAAAGAAGATCCGATTCCAAACTCCATATCCAGCAAGACTA
CGGGTGTACTATGAGGACGGGGGTTCACACGTACCATACAGCTGAAGAGGCATACACAGACATGGTGAAAAAAGGG
TTCCGTGTGGAAAAAACGCCGCAGACAGAATCGTTGATGGAACAACTCAAACACACGTCGTGGGAAACTGTCTCG
CATAAAAAGACAAAAGAGAAAACAAATCAACAAGCGACTGACATCCGAGAGCGTCTACAGGCATATAGAAGATAA
TTTAAGTCTGTCTACGCCATCTGCAACAAATACTTGAATTGGTACATTTTTATGAACCGCAAAACTTGAGATAAT
TTCTGCTCTTTTTTTTCTCCGTATCGTTCAATAATGTCATATTACGTCGGACACGTGAGGTATGTTTACTTTCAT
CTCATGTAAATACGGTGAACTGGGGGGCCCTGAATACCACCACAGAAAGAAAGGTCTCCCTCCCACGCTAGATAT
CATCAGACTGTCCCAGGGTTGTTTGTTTGAGACCCCTAACTTGGAAGCTCAATATGTTCATCAGGTTCTTGTAGT
GTTCGTTATTGTTTCGTATATTGTTCTCCAGGGTTATTGTCGTCCTTACTCAAGAGATATTGAAATACTAAACCA
ACATAAGTGACGAGACAAAAATAATTTCATTTAATGTTAATGGCCTACTTAATCCAATCAAAAGGAAAAAAATAT
TGACTAAAATGAAGAAAGAAAATGCCCATATAGTATACTTACAAGAGACACACCTGAATAATTCAGAACATGAAA
AACTTAAAAGAATGGGATTTACCCAGATCTATTACTCATCATACCCCACAGGGCACAGGAGAGGGGTTATAATAT
TGATTTCTCATAAAGTTATTTTTGAAAAAAAATATGAATATGCTGACAAAGAAGGTCGTTTTATTCTGGTTAAGG
GGATACTAAATGGAACATATGTTACTTTGTTTAATATTTATGCTCCTCCTAATAGTAAAATTAATCTCTTTCAAA
ATATATTTCAGACCATAGCCTCTGAAACGCAGGGTATTCTGATCTGTGGGGGTGATCTAAATATACACCTACAAC
CTGATTTAGACTCCACAAATAAGAAAAGAATTTTTCCCAAACCTATTCTTCGGAAAATTAATGCAATGATGAATG
ATATTGGCATAATTGATATATGGAGAGACTTGTACCCAAATACAAAAGATTATACACATTTTTCTTCTCCGCATA
ATACTTACTCAAGGATTGATTATTACCTTATGTTTAAAACAGATAGAGCCAGAGTTACGAGTTGTGATATTGGAA
CTATAGACATTAGTGATCATGCTCCAGTCTATTTAAAACTACAATTAAATAACAAATACAAAGACACACTATGGA
AATTTAACCTAAACCTTCTTAATAATCCATCTTTTAAATTACATCTTAAAAATGAAATTCATACCTATTTAGAAA
ATAATGATAATGAAGAGGTCACTCCTCCTATACTGTGGGATGCCGCAAAAGCAGTATTAAGAGGGAAAACTATAT
CATTTGCTTCACGAAACAAGAAACTCCGGCAACAGGAATTAAACAAATTACAGCATCAACTAAAATCATTAGAAA
AGGAACACAAGATTAATCAAAACCCAAAAATATTAAAACAAATTAAAAAAGTACAGAATGAAATTAATATGCTAT
ATACTCAAGAGATTGAGAAAAAAATATCATTCTCCAGGCAAAAATATTATGAAAGTGGCCCCAAATTTATGAAAA
TATTGGCATGGAAATTAAGAAAACAACAAGCAGACCAAACAATATATAGAATCAGAGATCCTGTGACCAATACAA
TACAAAACAAACAAGAGAATATTCAAACTGTTTTTGAAATGTTTTACAAAAAACTATACTCTAAAATGCCTGAAG
ATAAAAACAAAGAAATTGACATTTTTCTTGATAAATTGGAACTACCAACACTAACTGATGAACAAAACAAATTAT
TGACAGCAGAAATAACAGAAACCGAAGTAAAAAAGGCCATTACTAAGCTTAAATGTAACAAATCTCCAGGACCAG
ATGGATTTACTGGGGAGTGGTATAGGGCATTCCAAAAAGAGATAATACCCATCTTGGCCCGGTCCTTTAATTGGA
CCCTAAAAAATGCGACTGTTCCACCATCATGGAAAGAAGCAATAATATCTATCATTCCAAAAGAGGACAAAGACA
AGTTACAGTGCGGATCATACAGACCTCTATCTGTGTTAAATGTTGATTATAGATTATATACATCAATTATGGCCA
GAAGAATGGAAGACCTCATGCCCAATCTGATAAACCAAGACCAATCTGGATTCATCCGACAGCGTCAAACTAATG
ATAATATTAGACGCACTTTACATGTTATGAACCATATCAGGAGAAATAAGATAACAGCAATGATTTTAAGTTTAG
ACGCTGAAAAAGCTTTCGATTCTGTCAGCTGGACATATCTTTACAAAGTCTTACACAAATTTGGCTTCCACGAAA
CTATAATCAAAAGTATACAAGCACTTTATGATACACCAACTGCAAAAATAAAGATCAACGGCCACCTCTCAAACT
CATTCACATTGGAAAGAGGCACCAGACAGGGATGCCCGTGGTCCCCTCAATTGTTTGCACTTTATTTGGAACCCC

TAGTACAAAGCATAAGACAGAATAAAACAATTCAAGGTATCAATATTAAAGGAGTGGAACACAAGATTGCCTGCT
ACGCAGATGATGTATTGATTTATTTAAGAAATCCAACCACTTCACTACCCTATCTCATGGAACAGCTCCAAAACT
CTGGCCCTATCTCTGGATATAAATTAAACATAGAAAAAACTGAAATAATTACTTATAATTATAATCCCCCTACAA
ATATAAAAAATGTATACTCATTAAAATGGCATACAAAATCTTTCAAATATTTAGGTATCCATCTAACAAAAAACA
CAGAAAAAATCCAAAAAATGAACTTCGACCCTATAACTGCTAAAACTAGAGAAGATTTGACCAGATGGAACCTAA
TCCCCTTTTTTAGCTTTAGTTCGAGGATCGAAACAATTAAGATGAACATTGTACCAAAATTTTTATATTTATTTC
AGTGTCTACCTGTAGAAATACCGGAGAAACAATTTATAGAATGGGATAAAATGCTTTCGCGCTTCATTTGGCAGG
GAAAAAAGCCGAGAGTACGCTTTAAAACACTGCAGCTGCCGAAAGATAAAGGAGGATGGGCTCTCCCCTCTCTAA
AAGACTATTATATAGCAGCACAAATCAGGACTATAGTAAACTGGTGTGACCCACACTATCATGCCCCATGGAAAG
ATATAGAAAACCTAACATTTAGAAATACGCCTGTTCAAGCCATATTAATAGATGAAAAATTACAAAAATATACAG
AACAAATAGAAGACCCATGGACAAAACTTACTTTAAATGTCTGGACATCAGTGATGAAAGAATTTAACCTGCTAA
AATACTGTACAGTTCTAAAATGGATCGCTTATGATTCTGATTTCACACCAAACAGATTGGATAACAGATTCAAAT
TATGGGCAAACAAGGGAATCACTTCCTTCAGTACCATCATTAAAAAGGGAGAAATTTTAAGCTTTCAACAATTAA
AGGACAATTTTCTACTTGAAAATCAAGATTTTTATAGGTACCTCCAAGTGCGAAACTATTATGATCAGAAAATAA
AAGGAAAACTTGAAGACAATCAGAATCCACTGATAGATATAATTAAAAAAGCTCACACATCAGGTATAAAGTACA
AAATAGTCTCTCTGATATACAAGAGTCTGAGAAGTATGAAAACACAATCCACAAATTATATAAAAGAAAAATGGG
AGACAGAAAGTGGAACAACATTCACTGATGAAGAGTGGACAAATATATGGGAATTTCAATGGAAATCAACAAGCT
CATTAAACTGGAAAGAAAACTGTTGGAAAAACATTATTAGATATTTCAAAACACCAGCACTAACAGCTAAATATA
CAAATAAGTCTCCAAAATGCTGGCGTAACTGCGGATGCAATAAAGCAAACCATTATCATATCTTTTGGGAATGTA
TGTATATTCAAGAATACTGGAAAGAAATACAAGAGGCTCTTGAATATATATTTGAAAAAGATATACCTCTAGAAA
GTAAATTGTTATATTTTGGCTGTGTGGATCCAGATATGGTAACGGCAGACAATAAATACCTTATGCGTGCTCTCA
TGGCTGCCAGCAAAAAAGCAATCACACGCAAATGGCTACAACAAGAGCGACCAACCCTTAGTAATTGGATAGACG
CAACTATTGAGATTTACGCAATGGAAAGAATCACCTTCGTTGTCAACTTAAAGAAAGATATATTCCTGAGAAAAT
GGAGGAAATGGGTTGAATATATTTTAGTTAGAAGACCTGACTTTATTGCTATAAATGAATAGAAGATGATGCTCT
GATAGATAAATATATGCTAATGCATGCACATGTATATGACTACACTTATTCATATTATAATGTGTCCCCATATGT
ATAATGAAAATGGCCACTTGAGATTACTCTTGAAGGAAATATATTATGACATGTGTTCTAACCCTGGATGTAAAT
AGTGTTTTAATTTTCATTTATTTTTGATTATTGTATTTTTATTTTATCTTTTTCCTATTATTATTATTATTTATT
TATTTATCTATTTTTTATTTATTTTATTTTTCCTCTCTTTATGCTGTTTTTCACTTGATAATTTACAAAATCTGA
GTACTACTGTGTAGTAAATACACCTGACAATAATGCACATAATTTTTATACCACCTAAGTTCACATACTGAGATA
CCAATGTACCTTGAAATGCCGTATGATTGTACTCAATTGAATGTTCAAAAAAAAAAAAAAAAAAAAAA

Sequence 6.3. The functional consensus sequence of the ZfL1-12B build from the full-length copies annotated in the zebrafish reference genome – 5'UTR, IGR and 3'UTR (light purple), ORF1 (dark purple) and ORF2 (blue).

Bibliography

Adams, J. W., Kaufman, R. E., Kretschmer, P. J., Harrison, M., & Nienhuis, A. W. (1980). A family of long reiterated DNA sequences, one copy of which is next to the human beta globin gene. *Nucleic Acids Research*, *8*(24), 6113–6128. https://doi.org/10.1093/nar

Ahl, V., Keller, H., Schmidt, S., & Weichenrieder, O. (2015). Retrotransposition and Crystal Structure of an Alu RNP in the Ribosome-Stalling Conformation. *Molecular Cell*, *60*(5), 715–727. https://doi.org/10.1016/j.molcel.2015.10.003

Alisch, R. S., Garcia-Perez, J. L., Muotri, A. R., Gage, F. H., & Moran, J. V. (2006). Unconventional translation of mammalian LINE-1 retrotransposons. *Genes and Development*, *20*(2), 210–224. https://doi.org/10.1101/gad.1380406

Alves, G., Tatro, A., & Fanning, T. (1996). Differential methylation of human LINE-1 retrotransposons in malignant. *Gene*, *176*, 39–44. https://doi.org/10.1016/b978-1-4831-6698-8.50009-6

Amir, R. E., Van den Veyver, I. B., Wan, M., Tran, C. Q., Francke, U., & Zoghbi, H. Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nature Genetics*, *23*(2), 185–188. https://doi.org/10.1038/13810

An, W., Han, J. S., Schrum, C. M., Maitra, A., Koentgen, F., & Boeke, J. D. (2008). Conditional activation of a single-copy L1 transgene in mice by Cre. *Genesis*, *46*(7), 373–383. https://doi.org/10.1002/dvg.20407

An, W., Han, J. S., Wheelan, S. J., Davis, E. S., Coombes, C. E., Ye, P., … Boeke, J. D. (2006). Active retrotransposition by a synthetic L1 element in mice. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(49), 18662–18667. https://doi.org/10.1073/pnas.0605300103

Antinucci, P., & Hindges, R. (2016). A crystal-clear zebrafish for in vivo imaging. *Scientific Reports*, *6*(1), 29490. https://doi.org/10.1038/srep29490

Ardeljan, D., Taylor, M. S., Ting, D. T., & Burns, K. H. (2017). The Human Long Interspersed Element-1 Retrotransposon: An Emerging Biomarker of Neoplasia. *Clinical Chemistry*, *63*(4), 816–822. https://doi.org/10.1373/clinchem.2016.257444

Ardeljan, D., Wang, X., Oghbaie, M., Taylor, M. S., Husband, D., Deshpande, V., … Lacava, J. (2020). LINE-1 ORF2p expression is nearly imperceptible in human cancers. *Mobile DNA*, *11*(1), 1–19. https://doi.org/10.1186/s13100-019-0191-2

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., … Stockinger, H. (2012). ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Research*, *40*(Web Server issue), W597-603. https://doi.org/10.1093/nar/gks400

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

Attig, J., & Ule, J. (2019). Genomic Accumulation of Retrotransposons Was Facilitated by Repressive RNA-Binding Proteins: A Hypothesis. *BioEssays*, *41*(2), 1–12. https://doi.org/10.1002/bies.201800132

Babushok, D. V., Ostertag, E. M., Courtney, C. E., Choi, J. M., & Kazazian, H. H. (2006). L1 integration in a transgenic mouse model. *Genome Research*, *16*(2), 240–250. https://doi.org/10.1101/gr.4571606

Badge, R. M., Alisch, R. S., & Moran, J. V. (2003). ATLAS: A system to selectively identify human-specific L1 insertions. *American Journal of Human Genetics*, *72*(4), 823–838. https://doi.org/10.1086/373939

Baillie, J. K., Barnett, M. W., Upton, K. R., Gerhardt, D. J., Richmond, T. A., De Sapio, F., … 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

Barchitta, M., Quattrocchi, A., Maugeri, A., Vinciguerra, M., & Agodi, A. (2014). LINE-1 hypomethylation in blood and tissue samples as an epigenetic marker for cancer risk: A

systematic review and meta-analysis. *PLoS ONE*, *9*(10). https://doi.org/10.1371/journal.pone.0109478

Bartsch, K., Knittler, K., Borowski, C., Rudnik, S., Damme, M., Aden, K., … Rabe, B. (2017). Absence of RNase H2 triggers generation of immunogenic micronuclei removed by autophagy. *Human Molecular Genetics*, *26*(20), 3960–3972. https://doi.org/10.1093/hmg/ddx283

Beck, C. R., Collier, P., Macfarlane, C., Malig, M., Kidd, J. M., Eichler, E. E., … Moran, J. V. (2010). LINE-1 Retrotransposition Activity in Human Genomes. *Cell*, *141*(7), 1159–1170. https://doi.org/10.1016/j.cell.2010.05.021

Beck, C. R., Garcia-Perez, J. L., Badge, R. M., & Moran, J. V. (2011). LINE-1 elements in structural variation and disease. *Annual Review of Genomics and Human Genetics*, *12*, 187–215. https://doi.org/10.1146/annurev-genom-082509-141802

Becker, K. G., Swergold, G. D., Ozato, K., & Thayer, R. E. (1993). Binding of the ubiquitous nuclear transcription factor YY1 to a cis regulatory sequence in the human LINE-1 transposable element. *Human Molecular Genetics*, *2*(10), 1697–1702. https://doi.org/10.1093/hmg/2.10.1697

Beilhartz, G. L., & Götte, M. (2010). HIV-1 Ribonuclease H: Structure, Catalytic Mechanism and Inhibitors. *Viruses*, *2*, 900–926. https://doi.org/10.3390/v2040900

Belancio, V. P., Roy-Engel, A. M., & Deininger, P. (2008). The impact of multiple splice sites in human L1 elements. *Gene*, *411*(1), 38–45. https://doi.org/10.1016/j.trsl.2012.08.005

Belancio, Victoria P, Hedges, D. J., & Deininger, P. (2006). LINE-1 RNA splicing and influences on mammalian gene expression. *Nucleic Acids Research*, *34*(5), 1512–1521. https://doi.org/10.1093/nar/gkl027

Belancio, Victoria P, Roy-Engel, A. M., Pochampally, R. R., & Deininger, P. (2010). Somatic expression of LINE-1 elements in human tissues. *Nucleic Acids Research*, *38*(12), 3909–3922. https://doi.org/10.1093/nar/gkq132

Belancio, Victoria P, Whelton, M., & Deininger, P. (2007). Requirements for polyadenylation at the 3' end of LINE-1 elements. *Gene*, *390*(1–2), 98–107. https://doi.org/10.1016/j.gene.2006.07.029

Belgnaoui, S. M., Gosden, R. G., Semmes, O. J., & Haoudi, A. (2006). Human LINE-1 retrotransposon induces DNA damage and apoptosis in cancer cells. *Cancer Cell International*, *6*, 1–10. https://doi.org/10.1186/1475-2867-6-13

Benitez-Guijarro, M., Lopez-Ruiz, C., Tarnauskaitė, Ž., Murina, O., Mohammad, M. M., Williams, T. C., … Garcia-Perez, J. L. (2018). RNase H2, mutated in Aicardi-Goutières syndrome, promotes LINE-1 retrotransposition. *EMBO Journal, 37*, e98506. https://doi.org/10.15252/embj.201798506

Bessa, J., Luengo, M., Rivero-Gil, S., Ariza-Cosano, A., Maia, A. H. F., Ruiz-Ruano, F. J., … Gómez-Skarmeta, J. L. (2014). A mobile insulator system to detect and disrupt cis-regulatory landscapes in vertebrates. *Genome Research*, *24*(3), 487–495. https://doi.org/10.1101/gr.165654.113

Bestor, T. H., & Bourc'his, D. (2004). Transposon silencing and imprint establishment in mammalian germ cells. *Cold Spring Harbor Symposia on Quantitative Biology*, *69*, 381–387. https://doi.org/10.1101/sqb.2004.69.381

Bodak, M., Yu, J., & Ciaudo, C. (2014). Regulation of LINE-1 in mammals. *Biomolecular Concepts*, *5*(5), 409–428. https://doi.org/10.1515/bmc-2014-0018

Boeke, J. D., Garfinkel, D. J., Styles, C. A., & Fink, G. R. (1985). Ty elements transpose through an RNA intermediate. *Cell*, *40*(3), 491–500

Bogearts, J., & Prenen, H. (2014). Molecular genetics of colorectal cancer. *Annals of Gastroenterology*, *27*, 9–14. https://doi.org/10.1053/j.scrs.2016.04.013

Böhne, A., Brunet, F., Galiana-Arnoux, D., Schultheis, C., & Volff, J.-N. (2008). Transposable

elements as drivers of genomic and biological diversity in vertebrates. *Chromosome Research*, *16*(1), 203–215. https://doi.org/10.1007/s10577-007-1202-6

Boissinot, S., & Sookdeo, A. (2016). The Evolution of LINE-1 in Vertebrates. *Genome Biology Evolution*, *8*(12), 3485–3507. https://doi.org/10.1093/gbe/evw247

Boo, S. H., & Kim, Y. K. (2020). The emerging role of RNA modifications in the regulation of mRNA stability. *Experimental & Molecular Medicine*, *52*(3), 400–408. https://doi.org/10.1038/s12276-020-0407-z

Bourque, G. (2009). Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Current Opinion in Genetics & Development*, *19*(6), 607–612. https://doi.org/10.1016/j.gde.2009.10.013

Bratthauer, G. L., Cardiff, R. D., & Fanning, T. G. (1994). Expression of LINE-1 retrotransposons in human breast cancer. *Cancer*, *73*(9), 2333–2336

Brégnard, C., Guerra, J., Déjardin, S., Passalacqua, F., Benkirane, M., & Laguette, N. (2016). Upregulated LINE-1 Activity in the Fanconi Anemia Cancer Susceptibility Syndrome Leads to Spontaneous Pro-inflammatory Cytokine Production. *EBioMedicine*, *8*, 184–194. https://doi.org/10.1016/j.ebiom.2016.05.005

Bringaud, F., García-Pérez, J. L., Heras, S. R., Ghedin, E., El-Sayed, N. M., Andersson, B., … Lopez, M. C. (2002). Identification of non-autonomous non-LTR retrotransposons in the genome of Trypanosoma cruzi. *Molecular and Biochemical Parasitology*, *124*(1–2), 73–78. https://doi.org/10.1016/S0166-6851(02)00167-6

Brouha, B., Schustak, J., Badge, R. M., Lutz-Prigge, S., Farley, A. H., Moran, J. V, & Kazazian, H. H. (2003). Hot L1s account for the bulk of retrotransposition in the human population. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(9), 5280–5285. https://doi.org/10.1073/pnas.0831042100

Bruning, J. B., & Shamoo, Y. (2004). Structural and thermodynamic analysis of human PCNA with peptides derived from DNA polymerase-δ p66 subunit and flap endonuclease-1. *Structure*, *12*(12), 2209–2219. https://doi.org/10.1016/j.str.2004.09.018

Bubeck, D., Reijns, M. A. M., Graham, S. C., Astell, K. R., Jones, E. Y., & Jackson, A. P. (2011). PCNA directs type 2 RNase H activity on DNA replication and repair substrates. *Nucleic Acids Research*, *39*(9), 3652–3666. https://doi.org/10.1093/nar/gkq980

Bundo, M., Toyoshima, M., Okada, Y., Akamatsu, W., Ueda, J., Nemoto-Miyauchi, T., … Iwamoto, K. (2014). Increased L1 Retrotransposition in the Neuronal Genome in Schizophrenia. *Neuron*, *81*(2), 306–313. https://doi.org/10.1016/j.neuron.2013.10.053

Burkovics, P., Szukacsov, V., Unk, I., & Haracska, L. (2006). Human Ape2 protein has a 3′-5′ exonuclease activity that acts preferentially on mismatched base pairs. *Nucleic Acids Research*, *34*(9), 2508–2515. https://doi.org/10.1093/nar/gkl259

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

Castro-Diaz, N., Ecco, G., Coluccio, A., Kapopoulou, A., Yazdanpanah, B., Friedli, M., … Trono, D. (2014). Evolutionally dynamic L1 regulation in embryonic stem cells. *Genes and Development*, *28*(13), 1397–1409. https://doi.org/10.1101/gad.241661.114

Cecco, M. De, Ito, T., Petrashen, A. P., Elias, A. E., Skvir, N. J., Criscione, S. W., … Sedivy, J. M. (2019). LINE-1 derepression in senescent cells triggers interferon and inflammaging. *Nature*, *566*(7742), 73-78. https://doi.org/10.1038/s41586-018-0784-9

Cerritelli, S. M., & Crouch, R. J. (2009). Ribonuclease H: the enzymes in Eukaryotes. *FEBS Journal*, *276*(6), 1494–1505. https://doi.org/10.1111/j.1742-4658.2009.06908.x

Chon, H., Vassilev, A., DePamphilis, M. L., Zhao, Y., Zhang, J., Burgers, P. M., … Cerritelli, S. M. (2009). Contributions of the two accessory subunits, RNASEH2B and RNASEH2C, to the activity and properties of the human RNase H2 complex. *Nucleic Acids Research*, *37*(1), 96–110. https://doi.org/10.1093/nar/gkn913

Chow, J. C., Ciaudo, C., Fazzari, M. J., Mise, N., Servant, N., Glass, J. L., … Heard, E. (2010). LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. *Cell*, *141*(6), 956–969. https://doi.org/10.1016/j.cell.2010.04.042

Christian, C. M., Deharo, D., Kines, K. J., Sokolowski, M., & Belancio, V. P. (2016). Identification of L1 ORF2p sequence important to retrotransposition using Bipartile Alu retrotransposition (BAR). *Nucleic Acids Research*, *44*(10), 4818–4834. https://doi.org/10.1093/nar/gkw277

Clements, A. P., & Singer, M. F. (1998). The human LINE-1 reverse transcriptase: effect of deletions outside the common reverse transcriptase domain. *Nucleic Acids Research*, *26*(15): 3528–3535. https://doi.org/10.1093/nar/26.15.3528

Cortelazzo, A., De Felice, C., Guerranti, R., Signorini, C., Leoncini, S., Pecorelli, A., … Hayek, J. (2014). Subclinical inflammatory status in Rett syndrome. *Mediators of Inflammation*, *2014*: 480980 . https://doi.org/10.1155/2014/480980

Cost, G. J., Feng, Q., Jacquier, A., & Boeke, J. D. (2002). Human L1 element target-primed reverse transcription in vitro. *The EMBO Journal*, *21*(21), 5899–5910. https://doi.org/10.1093/emboj/cdf592

Coufal, N. G., Garcia-Perez, J. L., Peng, G. E., Marchetto, M. C. N., Muotri, A. R., Mu, Y., … Gage, F. H. (2011). Ataxia telangiectasia mutated (ATM) modulates long interspersed element-1 (L1) retrotransposition in human neural stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(51), 20382–20387. https://doi.org/10.1073/pnas.1100273108

Coufal, N. G., Garcia-Perez, J. L., Peng, G. E., Yeo, G. W., Mu, Y., Lovci, M. T., … Gage, F. H. (2009). L1 retrotransposition in human neural progenitor cells. *Nature*, *460*(7259), 1127–1131. https://doi.org/10.1038/nature08248

Crichton, J. H., Dunican, D. S., Maclennan, M., Meehan, R. R., & Adams, I. R. (2014). Defending the genome from the enemy within: mechanisms of retrotransposon suppression in the mouse germline. *Cellular and Molecular Life Sciences : CMLS*, *71*(9), 1581–1605. https://doi.org/10.1007/s00018-013-1468-0

Crick, F., & Watson, J. (1953). Molecular Structure of Nucleic Acids. *Nature*, *171*, 737–738.

Crow, Y. J., Vanderver, A., Orcesi, S., Kuijpers, T. W., & Rice, G. I. (2014). Therapies in Aicardi-Goutières syndrome. *Clinical and Experimental Immunology*, *175*(1), 1–8. https://doi.org/10.1111/cei.12115

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

Daskalos, A., Nikolaidis, G., Xinarianos, G., Savvari, P., Cassidy, A., Zakopoulou, R., … Liloglou, T. (2009). Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer. *International Journal of Cancer*, *124*(1), 81–87. https://doi.org/10.1002/ijc.23849

de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A., & Pollock, D. D. (2011). Repetitive elements may comprise over Two-Thirds of the human genome. *PLoS Genetics*, *7*(12). https://doi.org/10.1371/journal.pgen.1002384

de la Peña, M., & Cervera, A. (2017). Circular RNAs with hammerhead ribozymes encoded in eukaryotic genomes: The enemy at home. *RNA Biology*, *14*(8), 985–991. https://doi.org/10.1080/15476286.2017.1321730

De Luca, C., Guadagni, F., Sinibaldi-Vallebona, P., Sentinelli, S., Gallucci, M., Hoffmann, A., … Sciamanna, I. (2016). Enhanced expression of LINE-1-encoded ORF2 protein in early stages of colon and prostate transformation. *Oncotarget*, *7*(4), 4048–4061. https://doi.org/10.18632/oncotarget.6767

De March, M., Merino, N., Barrera-Vilarmau, S., Crehuet, R., Onesti, S., Blanco, F. J., & De

Biasio, A. (2017). Structural basis of human PCNA sliding on DNA. *Nature Communications*, *8*(13935). https://doi.org/10.1038/ncomms13935

de Souza, F. S. J., Franchini, L. F., & Rubinstein, M. (2013). Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? *Molecular Biology and Evolution*, *30*(6), 1239–1251. https://doi.org/10.1093/molbev/mst045

Deberardinis, R. J., & Kazazian, H. H. (1999). Analysis of the Promoter from an Expanding Mouse Retrotransposon Subfamily aligned the sequences of 53 monomers to generate a consensus T F monomer and determined that most T F elements are truncated near a potential binding site. *Genomics*, *323*(56), 317–323. https://doi.org/10.1006/geno.1998.5729

Deininger, P. L., & Batzer, M. A. (1999). Alu repeats and human disease. *Molecular Genetics and Metabolism*, *67*(3), 183–193. https://doi.org/10.1006/mgme.1999.2864

Denli, A. M., Narvaiza, I., Kerman, B. E., Pena, M., Benner, C., Marchetto, M. C. N., … 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

Dmitriev, S. E., Andreev, D. E., Terenin, I. M., Olovnikov, I. A., Prassolov, V. S., Merrick, W. C., & Shatsky, I. N. (2007). Efficient Translation Initiation Directed by the 900-Nucleotide-Long and GC-Rich 5′ Untranslated Region of the Human Retrotransposon LINE-1 mRNA Is Strictly Cap Dependent Rather than Internal Ribosome Entry Site Mediated. *Molecular and Cellular Biology*, *27*(13), 4685–4697. https://doi.org/10.1128/mcb.02138-06

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

Doolittle, W. F., & Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature*, *284*(5757), 601–603. https://doi.org/10.1038/284601a0

Doucet-O'Hare, T. T., Rodić, N., Sharma, R., Darbari, I., Abril, G., Choi, J. A., … Kazazian, H. H. (2015). LINE-1 expression and retrotransposition in Barrett's esophagus and esophageal carcinoma. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(35), E4894–E4900. https://doi.org/10.1073/pnas.1502474112

Doucet, A. J., Hulme, A. E., Sahinovic, E., Kulpa, D. A., Moldovan, J. B., Kopera, H. C., … Gilbert, N. (2010). Characterization of LINE-1 Ribonucleoprotein Particles. *PLoS Genetics*, *6*(10), e1001150. https://doi.org/10.1371/journal.pgen.1001150

Dovrat, D., Stodola, J. L., Burgers, P. M. J., & Aharoni, A. (2014). Sequential switching of binding partners on PCNA during in vitro Okazaki fragment maturation. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(39), 14118–14123. https://doi.org/10.1073/pnas.1321349111

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., … Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Duvernell, D. D., Pryor, S. R., & Adams, S. M. (2004). Teleost fish genomes contain a diverse array of L1 retrotransposon lineages that exhibit a low copy number and high rate of turnover. *Journal of Molecular Evolution*, *59*(3), 298–308. https://doi.org/10.1007/s00239-004-2625-8

Eddy, S. R. (2004). What is a hidden Markov model? *Nature Biotechnology*, *22*(10), 1315–1316. https://doi.org/10.1038/nbt1004-1315

Eickbush, D. G., & Eickbush, T. H. (1995). Vertical Transmission of the Retrotransposable

Elements Rl and R2 During the Evolution of the Drosophila melanogaster Species Subgroup. *Genetics Society of America*, *139*, 671–684. https://doi.org/10.1134/S1062359019030038

Eickbush, Danna G., & Eickbush, T. H. (2010). R2 Retrotransposons Encode a Self-Cleaving Ribozyme for Processing from an rRNA Cotranscript. *Molecular and Cellular Biology*, *30*(13), 3142–3150. https://doi.org/10.1128/mcb.00300-10

Eickbush, T. H., & Eickbush, D. G. (2015). Integration, Regulation, and Long-Term Stability of R2 Retrotransposons. *Microbiology Spectrum*, *3*(2). https://doi.org/10.1128/microbiolspec.MDNA3-0011-2014.Integration

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., … Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432. https://doi.org/10.1093/nar/gky995

Ellison, V., & Stillman, B. (2003). Biochemical characterization of DNA damage checkpoint complexes: Clamp loader and clamp complexes with specificity for 5′ recessed DNA. *PLoS Biology*, *1*(2). https://doi.org/10.1371/journal.pbio.0000033

Ergün, S., Buschmann, C., Heukeshoven, J., Dammann, K., Schnieders, F., Lauke, H., … Schumann, G. G. (2004). Cell type-specific expression of LINE-1 open reading frames 1 and 2 in fetal and adult human tissues. *Journal of Biological Chemistry*, *279*(26), 27753–27763. https://doi.org/10.1074/jbc.M312985200

Erwin, J. A., Paquola, A. C. M., Singer, T., Gallina, I., Novotny, M., Quayle, C., … 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

Essers, J., Theil, A. F., Baldeyron, C., van Cappellen, W. A., Houtsmuller, A. B., Kanaar, R., & Vermeulen, W. (2005). Nuclear dynamics of PCNA in DNA replication and repair. *Molecular and Cellular Biology*, *25*(21), 9350–9359. https://doi.org/10.1128/MCB.25.21.9350-9359.2005

Etchegaray, E., Naville, M., Volff, J. N., & Haftek-Terreau, Z. (2021). Transposable element-derived sequences in vertebrate development. *Mobile DNA*, *12*(1), 1–24. https://doi.org/10.1186/s13100-020-00229-5

Evrony, G. D., Cai, X., Lee, E., Hills, L. B., Elhosary, P. C., Lehmann, H. S., … 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., Mehta, B. K., Benjamini, Y., Johnson, R. M., Cai, X., … Walsh, C. A. (2015). Cell lineage analysis in human brain using endogenous retroelements. *Neuron*, *85*(1), 49–59. https://doi.org/10.1016/j.neuron.2014.12.028

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*. https://doi.org/10.7554/eLife.12966

Ewing, A. D., Gacita, A., Wood, L. D., Ma, F., Xing, D., Kim, M. S., … Solyom, S. (2015). Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution. *Genome Research*, *25*(10), 1536–1545. https://doi.org/10.1101/gr.196238.115

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

Fanning, T., & Singer, M. (1987). The LINE-1 DNA sequences in four mammalian orders predict proteins that conserve homologies to retrovirus proteins. *Nucleic Acids Research*, *15*(5), 2251–2260. https://doi.org/10.1093/nar/15.5.2251

Farabaugh, P. J., & Fink, G. R. (1980). Insertion of the eukaryotic transposable element Ty1 creates a 5-base pair duplication. *Nature*, *286*(5771), 352–356. https://doi.org/10.1038/286352a0

Faulkner, G. J., Kimura, Y., Daub, C. O., Wani, S., Plessy, C., Irvine, K. M., … Carninci, P. (2009). The regulated retrotransposon transcriptome of mammalian cells. *Nature Genetics*, *41*(5), 563–571. https://doi.org/10.1038/ng.368

Fawcett, D. H., Lister, C. K., Kellett, E., & Finnegan, D. J. (1986). Transposable Elements Controlling I-R Hybrid Dysgenesis in D. melanogaster Are Similar to Mammalian LINES. *Cell*, *47*(6), 1007-1015. https://doi.org/10.1016/0092-8674(86)90815-9

Fedorov, A. V., Lukyanov, D. V., & Podgornaya, O. I. (2006). Identification of the proteins specifically binding to the rat LINE1 promoter. *Biochemical and Biophysical Research Communications*, *340*(2), 553–559. https://doi.org/10.1016/j.bbrc.2005.12.040

Feng, Q., Moran, J. V, Kazazian, H. H., & Boeke, J. D. (1996). Human L1 Retrotransposon Encodes a Conserved Endonuclease Required for Retrotransposition. *Cell*, *87*(5), 905–916. https://doi.org/10.1016/S0092-8674(00)81997-2

Feng, Q., Schumann, G., & Boeke, J. D. (1998). Retrotransposon R1Bm endonuclease cleaves the target sequence. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(5), 2083–2088. https://doi.org/10.1073/pnas.95.5.2083

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

Feschotte, C., & Gilbert, C. (2012). Endogenous viruses: Insights into viral evolution and impact on host biology. *Nature Reviews Genetics*, *13*(4), 283–296. https://doi.org/10.1038/nrg3199

Feusier, J., Watkins, W. S., Thomas, J., Farrell, A., Witherspoon, D. J., Baird, L., … Jorde, L. B. (2019). Pedigree-based estimation of human mobile element retrotransposition rates. *Genome Research*, *29*(10), 1567–1577. https://doi.org/10.1101/gr.247965.118

Finnegan, D. J. (1989). Eukaryotic transposable elements and genome evolution. *Trends in Genetics*, *5*(C), 103–107. https://doi.org/10.1016/0168-9525(89)90039-5

Flasch, D. A., Macia, Á., Sánchez, L., Ljungman, M., Heras, S. R., García-Pérez, J. L., … Moran, J. V. (2019). Genome-wide de novo L1 Retrotransposition Connects Endonuclease Activity with Replication. *Cell*, *177*(4), 837-851.e28. https://doi.org/10.1016/j.cell.2019.02.050

Furano, A. V., Duvernell, D. D., & Boissinot, S. (2004). L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. *Trends in Genetics*, *20*(1), 9–14. https://doi.org/10.1016/J.TIG.2003.11.006

Garcia-Perez, J. L., Marchetto, M. C. N., Muotri, A. R., Coufal, N. G., Gage, F. H., O'Shea, K. S., & Moran, J. V. (2007). LINE-1 retrotransposition in human embryonic stem cells. *Human Molecular Genetics*, *16*(13), 1569–1577. https://doi.org/10.1093/hmg/ddm105

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

Gary, R., Ludwig, D. L., Cornelius, H. L., MacInnes, M. A., & Park, M. S. (1997). The DNA repair endonuclease XPG binds to proliferating cell nuclear antigen (PCNA) and shares sequence elements with the PCNA-binding regions of FEN-1 and cyclin-dependent kinase inhibitor p21. *Journal of Biological Chemistry*, *272*(39), 24522–24529. https://doi.org/10.1074/jbc.272.39.24522

Gasior, S. L., Wakeman, T. P., Xu, B., & Deininger, P. L. (2006). The human LINE-1 retrotransposon creates DNA double-strand breaks. *Journal of Molecular Biology*, *357*(5), 1383–1393. https://doi.org/10.1016/j.jmb.2006.01.089

George, J. A., Traverse, K. L., DeBaryshe, P. G., Kelley, K. J., & Pardue, M. Lou. (2010). Evolution of diverse mechanisms for protecting chromosome ends by Drosophila TART telomere retrotransposons. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(49), 21052–21057. https://doi.org/10.1073/pnas.1015926107

Gilbert, C., & Feschotte, C. (2018). Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. *Current Opinion in Genetics & Development*, *49*, 15–24. https://doi.org/10.1016/j.gde.2018.02.007.

Gilbert, N., Lutz, S., Morrish, T. A., & Moran, J. V. (2005). Multiple Fates of L1 Retrotransposition Intermediates in Cultured Human Cells. *Molecular and Cellular Biology*, *25*(17), 7780–7795. https://doi.org/10.1128/mcb.25.17.7780-7795.2005

Gilbert, N., & Moran, J. V. (2002). Mammalian LINE-1 Retrotransposons and Related Elements. *American Society of Microbiology*. https://doi.org/10.1128/9781555817954.ch35

Goodier, J. L., Ostertag, E. M., Du, K., & Kazazian, J. (2001). A novel active L1 retrotransposon subfamily in the mouse. *Genome Research*, *11*(10), 1677–1685. https://doi.org/10.1101/gr.198301.

Goodier, John L. (2016). Restricting retrotransposons: A review. *Mobile DNA*, *7*(1). https://doi.org/10.1186/s13100-016-0070-z

Goodier, John L., Cheung, L. E., & Kazazian, H. H. (2012). MOV10 RNA Helicase Is a Potent Inhibitor of Retrotransposition in Cells. *PLoS Genetics*, *8*(10), e1002941. https://doi.org/10.1371/journal.pgen.1002941

Grimaldi, G., Skowronski, J., & Singer, M. (1984). Defining the beginning and end of KpnI family segments. *The EMBO Journal*, *3*(8), 1753–1759. https://doi.org/10.1002/j.1460-2075.1984.tb02042.x

Gulbis, J. M., Kelman, Z., Hurwitz, J., O'Donnell, M., & Kuriyan, J. (1996). Structure of the C-terminal region of p21(WAF1/CIP1) complexed with human PCNA. *Cell*, *87*(2), 297–306. https://doi.org/10.1016/S0092-8674(00)81347-1

Hadi, M. Z., Ginalski, K., Nguyen, L. H., & Wilson, D. M. (2002). Determinants in nuclease specificity of Ape1 and Ape2, human homologues of Escherichia coli exonuclease III. *Journal of Molecular Biology*, *316*(3), 853–866. https://doi.org/10.1006/jmbi.2001.5382

Hall, L. L., Carone, D. M., Gomez, A. V., Kolpa, H. J., Byron, M., Mehta, N., … Lawrence, J. B. (2014). Stable C0T-1 repeat RNA is abundant and is associated with euchromatic interphase chromosomes. *Cell*, *156*(5), 907–919. https://doi.org/10.1016/j.cell.2014.01.042

Han, J. S., & Boeke, J. D. (2004). A highly active synthetic mammalian retrotransposon. *Nature*, *429*(6989), 314–318. https://doi.org/10.1038/nature02535

Han, J. S., Szak, S. T., & Boeke, J. D. (2004). Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature*, *429*(6989), 268–274. https://doi.org/10.1038/nature02536

Hancks, D. C., Goodier, J. L., Mandal, P. K., Cheung, L. E., & Kazazian, H. H. (2011). Retrotransposition of marked SVA elements by human L1s in cultured cells. *Human Molecular Genetics*, *20*(17), 3386–3400. https://doi.org/10.1093/hmg/ddr245

Hancks, D. C., & Kazazian, H. H. (2012). Active Human Retrotransposons: Variation and Disease. *Current Opinion in Genetics & Development*, *22*(3), 191–203. https://doi.org/10.1016/j.gde.2012.02.006.Active

Hancks, D. C., & Kazazian, H. H. (2016). Roles for retrotransposon insertions in human disease. *Mobile DNA*, *7*(1). https://doi.org/10.1186/s13100-016-0065-9

Haracska, L., Johnson, R. E., Unk, I., Phillips, B. B., Hurwitz, J., Prakash, L., & Prakash, S. (2001a). Targeting of human DNA polymerase ι to the replication machinery via interaction with PCNA. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(25), 14256–14261. https://doi.org/10.1073/pnas.261560798

Haracska, L., Johnson, R. E., Unk, I., Phillips, B., Hurwitz, J., Prakash, L., & Prakash, S. (2001b). Physical and Functional Interactions of Human DNA Polymerase η with PCNA. *Molecular and Cellular Biology*, *21*(21), 7199–7206.

https://doi.org/10.1128/mcb.21.21.7199-7206.2001
Haracska, L., Unk, I., Johnson, R. E., Phillips, B. B., Hurwitz, J., Prakash, L., & Prakash, S. (2002). Stimulation of DNA Synthesis Activity of Human DNA Polymerase κ by PCNA. *Molecular and Cellular Biology*, *22*(3), 784–791. https://doi.org/10.1128/mcb.22.3.784-791.2002
Harada, K., Baba, Y., Ishimoto, T., Chikamoto, A., Kosumi, K., Hayashi, H., … Baba, H. (2015). LINE-1 methylation level and patient prognosis in a database of 208 hepatocellular carcinomas. *Annals of Surgical Oncology*, *22*(4), 1280–1287. https://doi.org/10.1245/s10434-014-4134-3
Hardies, S. C., Wang, L., Zhou, L., Zhao, Y., Casavant, N. C., & Huang, S. (2000). Line-1 (L1) lineages in the mouse. *Molecular Biology and Evolution*, *17*(4), 616–628. https://doi.org/10.1093/oxfordjournals.molbev.a026340
Harris, C. R., Normart, R., Yang, Q., Stevenson, E., Haffty, B. G., Ganesan, S., … Tang, L. H. (2010). Association of nuclear localization of a long interspersed nuclear element-1 protein in breast tumors with poor prognostic outcomes. *Genes and Cancer*, *1*(2), 115–124. https://doi.org/10.1177/1947601909360812
Härtlova, A., Erttmann, S. F., Raffi, F. A. M., Schmalz, A. M., Resch, U., Anugula, S., … Gekara, N. O. (2015). DNA Damage Primes the Type I Interferon System via the Cytosolic DNA Sensor STING to Promote Anti-Microbial Innate Immunity. *Immunity*, *42*(2), 332–343. https://doi.org/10.1016/j.immuni.2015.01.012
Hattori, M., Kuhara, S., Takenaka, O., & Sakaki, Y. (1986). L1 family of repetitive DNA sequences in primates may be derived from a sequence encoding a reverse transcriptase-related protein. *Nature*, *321*(6070), 625–628. https://doi.org/10.1038/321625a0
Hayashi, Y., Kajikawa, M., Matsumoto, T., & Okada, N. (2014). Mechanism by which a LINE protein recognizes its 3′ tail RNA. *Nucleic Acids Research*, *42*(16), 10605–10617. https://doi.org/10.1093/nar/gku753
Helman, E., Lawrence, M. S., Stewart, C., Sougnez, C., Getz, G., & Meyerson, M. (2014). Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Research*, *24*(7), 1053–1063. https://doi.org/10.1101/gr.163659.113
Heras, S. R., Macias, S., Plass, M., Fernandez, N., Cano, D., Eyras, E., … Cáceres, J. F. (2013). The Microprocessor controls the activity of mammalian retrotransposons. *Nature Structural & Molecular Biology*, *20*(10), 1173–1181. https://doi.org/10.1038/nsmb.2658
Herrmann, A., Wittmann, S., Thomas, D., Shepard, C. N., Kim, B., Ferreirós, N., & Gramberg, T. (2018). The SAMHD1-mediated block of LINE-1 retroelements is regulated by phosphorylation. *Mobile DNA*, *9*(1), 11. https://doi.org/10.1186/s13100-018-0116-5
Hiller, B., Achleitner, M., Glage, S., Naumann, R., Behrendt, R., & Roers, A. (2012). Mammalian RNase H2 removes ribonucleotides from DNA to maintain genome integrity. *Journal of Experimental Medicine*, *209*(8), 1419–1426. https://doi.org/10.1084/jem.20120876
Hishiki, A., Hashimoto, H., Hanafusa, T., Kamei, K., Ohashi, E., Shimizu, T., … Sato, M. (2009). Structural basis for novel interactions between human translesion synthesis polymerases and proliferating cell nuclear antigen. *Journal of Biological Chemistry*, *284*(16), 10552–10560. https://doi.org/10.1074/jbc.M809745200
Hohjoh, H., & Singer, M. F. (1996). Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. *The EMBO Journal*, *15*(3), 630–639.
Honda, H., Ichiyanagi, K., Suzuki, J., Ono, T., Koyama, H., Kajikawa, M., & Okada, N. (2007). A new system for analyzing LINE retrotransposition in the chicken DT40 cell line widely used for reverse genetics. *Gene*, *395*(1–2), 116–124. https://doi.org/10.1016/j.gene.2007.02.017
Hotter, D., Bosso, M., Jønsson, K. L., Krapp, C., Christina, M., Das, A., … Kirchhoff, F.

(2019). IFI16 targets the transcription factor Sp1 to suppress HIV-1 transcription and latency reactivation, *25*(6), 858–872. https://doi.org/10.1016/j.chom.2019.05.002.IFI16

Houwing, S., Kamminga, L. M., Berezikov, E., Cronembold, D., Girard, A., van den Elst, H., … Ketting, R. F. (2007). A Role for Piwi and piRNAs in Germ Cell Maintenance and Transposon Silencing in Zebrafish. *Cell*, *129*(1), 69–82. https://doi.org/10.1016/j.cell.2007.03.026

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., … Stemple, D. L. (2013a). The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, *496*(7446), 498–503. https://doi.org/10.1038/nature12111

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., … Stemple, D. L. (2013b). The zebrafish reference genome sequence and its relationship to the human genome. https://doi.org/10.1038/nature12111

Hu, S., Li, J., Xu, F., Mei, S., Le Duff, Y., Yin, L., … Guo, F. (2015). SAMHD1 Inhibits LINE-1 Retrotransposition by Promoting Stress Granule Formation. *PLOS Genetics*, *11*(7), e1005367. https://doi.org/10.1371/journal.pgen.1005367

Idica, A., Sevrioukov, E. A., Zisoulis, D. G., Hamdorf, M., Daugaard, I., Kadandale, P., & Pedersen, I. M. (2017). MicroRNA miR-128 represses LINE-1 (L1) retrotransposition by down-regulating the nuclear import factor TNPO1. *Journal of Biological Chemistry*, *292*(50), 20494–20508. https://doi.org/10.1074/jbc.M117.807677

Iskow, R. C., McCabe, M. T., Mills, R. E., Torene, S., Pittard, W. S., Neuwald, A. F., … Devine, S. E. (2010). Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell*, *141*(7), 1253–1261. https://doi.org/10.1016/j.cell.2010.05.020

Ivancevic, A. M., Kortschak, R. D., Bertozzi, T., & Adelson, D. L. (2018). Horizontal transfer of BovB and L1 retrotransposons in eukaryotes. *Genome Biology*, *19*(1), 85. https://doi.org/10.1186/s13059-018-1456-7

Iwagami, S., Baba, Y., Watanabe, M., Shigaki, H., Miyake, K., Ishimoto, T., … Baba, H. (2013). LINE-1 hypomethylation is associated with a poor prognosis among patients with curatively resected esophageal squamous cell carcinoma. *Annals of Surgery*, *257*(3), 449–455. https://doi.org/10.1097/SLA.0b013e31826d8602

Jacobs, F. M. J., Greenberg, D., Nguyen, N., Haeussler, M., Ewing, A. D., Katzman, S., … Haussler, D. (2014). An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature*, *516*(7530), 242–245. https://doi.org/10.1038/nature13760

Jamburuthugoda, V. K., & Eickbush, T. H. (2014). Identification of RNA binding motifs in the R2 retrotransposon-encoded reverse transcriptase. *Nucleic Acids Research*, *42*(13), 8405–8415. https://doi.org/10.1093/nar/gku514

Jeong, H.-S., Backlund, P. S., Chen, H.-C., Karavanov, A. A., & Crouch, R. J. (2004). RNase H2 of Saccharomyces cerevisiae is a complex of three proteins. *Nucleic Acids Research*, *32*(2), 407–414. https://doi.org/10.1093/nar/gkh209

Jönsson, M. E., Ludvik Brattås, P., Gustafsson, C., Petri, R., Yudovich, D., Pircs, K., … Jakobsson, J. (2019). Activation of neuronal genes via LINE-1 elements upon global DNA demethylation in human neural progenitors. *Nature Communications*, *10*(1), 1–11. https://doi.org/10.1038/s41467-019-11150-8

Jónsson, Z. O., Hindges, R., & Hübscher, U. (1998). Regulation of DNA replication and repair proteins through interaction with the front side of proliferating cell nuclear antigen. *EMBO Journal*, *17*(8), 2412–2425. https://doi.org/10.1093/emboj/17.8.2412

Jurka, J. (1997). Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(5), 1872–1877. https://doi.org/10.1073/PNAS.94.5.1872

Jurka, Jerzy, Zietkiewicz, E., & Labuda, D. (1995). Ubiquitous mammalian-wide interspersed

repeats (MIRs) are molecular fossils from the mesozoic era. *Nucleic Acids Research*, *23*(1), 170–175. https://doi.org/10.1093/nar/23.1.170

Kajikawa, M., & Okada, N. (2002). LINEs Mobilize SINEs in the Eel through a Shared 3′ Sequence. *Cell*, *111*(3), 433–444. https://doi.org/10.1016/S0092-8674(02)01041-3

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E., & Schulman, A. H. (2000). Genome evolution of wild barley (Hordeum spontaneum) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(12), 6603–6607. https://doi.org/10.1073/pnas.110587497

Kalueff, A. V., Stewart, A. M., & Gerlai, R. (2014). Zebrafish as an emerging model for studying complex brain disorders. *Trends in Pharmacological Sciences*, *35*(2), 63–75. https://doi.org/10.1016/j.tips.2013.12.002

Kano, H., Godoy, I., Courtney, C., Vetter, M. R., Gerton, G. L., Ostertag, E. M., & Kazazian, H. H. (2009). L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. *Genes & Development*, *23*(11), 1303–1312. https://doi.org/10.1101/gad.1803909

Kazazian, H. H., Wong, C., Youssoufian, H., Scott, A. F., Phillips, D. G., & Antonarakis, S. E. (1988). Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature*, *332*(6160), 164–166. https://doi.org/10.1038/332164a0

Kearns, N. a, Genga, R. M. J., Enuameh, M. S., Garber, M., Wolfe, S. a, Maehr, R., … Fineran, P. C. (2014). High-throughput gene targeting and phenotyping in zebrafish using CRISPR/Cas9. *International Journal of Molecular Sciences*, *5*(1), 1–10. https://doi.org/10.1101/gr.186379.114.Freely

Kendall, J., Liu, Q., Bakleh, A., Krasnitz, A., Nguyen, K. C. Q., Lakshmi, B., … Mu, D. (2007). Oncogenic cooperation and coamplification of developmental transcription factor genes in lung cancer. *Proceedings of the National Academy of Sciences of the United States of America, 104*(42):16663-16668. https://doi.org/10.1073/pnas.0708286104

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Research*, *12*(4), 656–664. https://doi.org/10.1101/gr.229202

Khadgi, B. B., Govindaraju, A., & Christensen, S. M. (2019). Completion of LINE integration involves an open “4-way” branched DNA intermediate. *Nucleic Acids Research*, *47*(16), 8708–8719. https://doi.org/10.1093/nar/gkz673

Khan, H., Smit, A., & Boissinot, S. (2006). Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Research*, *16*(1), 78–87. https://doi.org/10.1101/gr.4001406

Khazina, E., Truffault, V., Büttner, R., Schmidt, S., Coles, M., & Weichenrieder, O. (2011). Trimeric structure and flexibility of the L1ORF1 protein in human L1 retrotransposition. *Nature Structural and Molecular Biology*, *18*(9), 1006–1014. https://doi.org/10.1038/nsmb.2097

Khazina, E., & Weichenrieder, O. (2009). Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(3), 731–736. https://doi.org/10.1073/pnas.0809964106

Kim, J. H., Lee, S. R., Li, L. H., Park, H. J., Park, J. H., Lee, K. Y., … Choi, S. Y. (2011). High cleavage efficiency of a 2A peptide derived from porcine teschovirus-1 in human cell lines, zebrafish and mice. *PLoS ONE*, *6*(4), 1–8. https://doi.org/10.1371/journal.pone.0018556

Kimberland, M. L., Divoky, V., Prchal, J., Schwahn, U., Berger, W., & Kazazian, H. H. (1999). Full-length human L1 insertions retain the capacity for high frequency retrotransposition

in cultured cells. *Human Molecular Genetics*, *8*(8), 1557-1560. https://doi.org/10.1093/hmg/8.8.1557

Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B., & Schilling, T. F. (1995). *Stages of Embryonic Development of the Zebrafish*.

Kinzler, K. W., & Vogelstein, B. (1996). Lessons from hereditary colorectal cancer. *Cell*, *87*(2), 159–170. https://doi.org/10.1016/S0092-8674(00)81333-1

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

Kopera, H. C., Flasch, D. A., Nakamura, M., Miyoshi, T., Doucet, A. J., & Moran, J. V. (2016). LEAP: L1 Element Amplification Protocol. *Methods in Molecular Biology*, *1400*(60), 339–355. https://doi.org/10.1007/978-1-4939-3372-3

Kubo, S., Del Carmen Seleme, M., Soifer, H. S., Perez, J. L. G., Moran, J. V., Kazazian, H. H., & Kasahara, N. (2006). L1 retrotransposition in nondividing and primary human somatic cells. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(21), 8036–8041. https://doi.org/10.1073/pnas.0601954103

Kulpa, D. a., & Moran, J. V. (2005). Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. *Human Molecular Genetics*, *14*(21), 3237–3248. https://doi.org/10.1093/hmg/ddi354

Kuwabara, T., Hsieh, J., Muotri, A., Yeo, G., Warashina, M., Lie, D. C., … Gage, F. H. (2009). Wnt-mediated activation of NeuroD1 and retro-elements during adult neurogenesis. *Nature Neuroscience*, *12*(9), 1097–1105. https://doi.org/10.1038/nn.2360

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Iii, L. J. L., … Peter, V. (2012). Landscape of Somatic Retrotransposition in Human Cancers. *Science*, *337*(6097), 967–971. https://doi.org/10.1126/science.1222077

Levin, D. S., Bai, W., Yao, N., O'Donnell, M., & Tomkinson, A. E. (1997). An interaction between DNA ligase I and proliferating cell nuclear antigen: Implications for Okazaki fragment synthesis and joining. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(24), 12863–12868. https://doi.org/10.1073/pnas.94.24.12863

Li, P., Du, J., Goodier, J. L., Hou, J., Kang, J., Kazazian, H. H., … Yu, X.-F. (2017). Aicardi–Goutières syndrome protein TREX1 suppresses L1 and maintains genome integrity through exonuclease-independent ORF1p depletion. *Nucleic Acids Research*, *45*(8), 4619–4631. https://doi.org/10.1093/nar/gkx178

Li, P. W.-L., Li, J., Timmerman, S. L., Krushel, L. A., & Martin, S. L. (2006). The dicistronic RNA from the mouse LINE-1 retrotransposon contains an internal ribosome entry site upstream of each ORF: implications for retrotransposition. *Nucleic Acids Research*, *34*(3), 853–864. https://doi.org/10.1093/nar/gkj490

Li, X., Zhang, J., Jia, R., Cheng, V., Xu, X., Qiao, W., … Cen, S. (2013). The MOV10 helicase inhibits LINE-1 mobility. *The Journal of Biological Chemistry*, *288*(29), 21148–21160. https://doi.org/10.1074/jbc.M113.465856

Lieschke, G. J., & Currie, P. D. (2007). Animal models of human disease: zebrafish swim into view. *Nature Reviews Genetics*, *8*(5), 353–367. https://doi.org/10.1038/nrg2091

Lim, Y. W., Sanz, L. A., Xu, X., Hartono, S. R., & Chédin, F. (2015). Genome-wide DNA hypomethylation and RNA:DNA hybrid accumulation in Aicardi–Goutières syndrome. *ELife*, *4*. https://doi.org/10.7554/eLife.08007

Liu, N., Lee, C. H., Swigut, T., Grow, E., Gu, B., Bassik, M. C., & Wysocka, J. (2018). Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators. *Nature*, *553*(7687), 228–232. https://doi.org/10.1038/nature25179

Liu, Y., Mi, Y., Mueller, T., Kreibich, S., Williams, E. G., Van Drogen, A., … Aebersold, R. (2019). Multi-omic measurements of heterogeneity in HeLa cells across laboratories. *Nature Biotechnology*, *37*(3), 314–322. https://doi.org/10.1038/s41587-019-0037-y

Luan, D. D., & Eickbush, T. H. (1995). RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element. *Molecular and Cellular Biology*, *15*(7), 3882–3891. https://doi.org/10.1128/mcb.15.7.3882

Luan, D. D., Korman, M. H., Jakubczak, J. L., & Eickbush, T. H. (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: A mechanism for non-LTR retrotransposition. *Cell*, *72*(4), 595–605. https://doi.org/10.1016/0092-8674(93)90078-5

Lutz, S. M., Vincent, B. J., Kazazian, H. H., Batzer, M. A., & Moran, J. V. (2003). Allelic Heterogeneity in LINE-1 Retrotransposition Activity. *American Journal of Human Genetics*, *73*(6), 1431–1437. https://doi.org/10.1086/379744

Macia, A., Blanco-Jimenez, E., & García-Pérez, J. L. (2015). Retrotransposons in pluripotent cells: Impact and new roles in cellular plasticity. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, *1849*(4), 417–426. https://doi.org/10.1016/j.bbagrm.2014.07.007

Macia, A., Muñoz-Lopez, M., Cortes, J. L., Hastings, R. K., Morell, S., Lucena-Aguilar, G., … Garcia-Perez, J. L. (2011). Epigenetic Control of Retrotransposon Expression in Human Embryonic Stem Cells. *Molecular and Cellular Biology*, *31*(2), 300–316. https://doi.org/10.1128/mcb.00561-10

Macia, A., Widmann, T. J., Heras, S. R., Ayllon, V., Sanchez, L., Benkaddour-Boumzaouad, M., … Garcia-Perez, J. L. (2017). Engineered LINE-1 retrotransposition in nondividing human neurons. *Genome Research*, *27*(3), 335–348. https://doi.org/10.1101/gr.206805.116

MacLennan, M., García-Cañadas, M., Reichmann, J., Khazina, E., Wagner, G., Playfoot, C. J., … Adams, I. R. (2017). Mobilization of LINE-1 retrotransposons is restricted by Tex19.1 in mouse embryonic stem cells. *ELife*, *6*. https://doi.org/10.7554/eLife.26152

Malik, H. S. (2005). Ribonuclease H evolution in retrotransposable elements. *Cytogenetic and Genome Research*, *110*, 392–401. https://doi.org/10.1159/000084971

Malik, H. S., Burke, W. D., & Eickbush, T. H. (1999). The age and evolution of non-LTR retrotransposable elements. *Molecular Biology and Evolution*, *16*(6), 793–805. https://doi.org/10.1093/oxfordjournals.molbev.a026164

Mankan, A. K., Schmidt, T., Chauhan, D., Goldeck, M., Höning, K., Gaidt, M., … Hornung, V. (2014). Cytosolic RNA:DNA hybrids activate the cGAS –STING axis . *The EMBO Journal*, *33*(24), 2937–2946. https://doi.org/10.15252/embj.201488726

Marchetto, M. C. N., Narvaiza, I., Denli, A. M., Benner, C., Lazzarini, T. A., Nathanson, J. L., … Gage, F. H. (2013). Differential L1 regulation in pluripotent stem cells of humans and apes. *Nature*, *503*(7477), 525–529. https://doi.org/10.1038/nature12686

Martin, S. L., & Bushman, F. D. (2001). Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Molecular and Cellular Biology*, *21*(2), 467–475. https://doi.org/10.1128/MCB.21.2.467-475.2001

Mátés, L., Chuah, M. K. L., Belay, E., Jerchow, B., Manoj, N., Acosta-Sanchez, A., … Izsvák, Z. (2009). Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. https://doi.org/10.1038/ng.343

Mathias, S. L., Scott, A. F., Kazazian, H. H., Boeke, J. D., & Gabriel, A. (1991). Reverse transcriptase encoded by a human transposable element. *Science*, *254*(5039), 1808–1810.

https://doi.org/10.3389/fchem.2016.00006
Mavragani, C. P., Sagalovskiy, I., Guo, Q., Nezos, A., Kapsogeorgou, E. K., Lu, P., … Crow, M. K. (2016). Expression of Long Interspersed Nuclear Element 1 Retroelements and Induction of Type I Interferon in Patients With Systemic Autoimmune Disease. *Arthritis and Rheumatology*, *68*(11), 2686–2696. https://doi.org/10.1002/art.39795
Mayanagi, K., Ishino, S., Shirai, T., Oyama, T., Kiyonari, S., Kohda, D., … Ishino, Y. (2018). Direct visualization of DNA baton pass between replication factors bound to PCNA. *Scientific Reports*, *8*(1), 1–10. https://doi.org/10.1038/s41598-018-34176-2
McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America*, *36*(6), 344–355. https://doi.org/10.1073/pnas.36.6.344
Mcclure, M. A. (1991). Evolution of Retroposons by Acquisition or Deletion of Retrovirus-like Genes. *Molecular Biology and Evolution*, *8*(6), 835–856. https://doi.org/10.1093/oxfordjournals.molbev.a040686
Miki, Y., Nishisho, I., Horii, A., Miyoshi, Y., Utsunomiya, J., Kinzler, K. W., … Nakamura, Y. (1992). Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Research*, *52*(3), 643–645.
Minakami, R., Kurose, K., Etoh, K., Furuhata, Y., Hattori, M., & Sakaki, Y. (1992). Identification of an internal cis-element essential for the human L1 transcription and a nuclear factor(s) binding to the element. *Nucleic Acids Research*, *20*(12), 3139–3145. https://doi.org/10.1093/nar/20.12.3139
Mita, P., Wudzinska, A., Sun, X., Andrade, J., Nayak, S., Kahler, D. J., … Boeke, J. D. (2018). LINE-1 protein localization and functional dynamics during the cell cycle. *ELife*, *7*, 1–35. https://doi.org/10.7554/eLife.30058
Mizrokhi, L. J., Georgieva, S. G., & Ilyin, Y. V. (1988). Jockey, a mobile drosophila element similar to mammalian LINEs, is transcribed from the internal promoter by RNA polymerase II. *Cell*, *54*(5), 685–691. https://doi.org/10.1016/S0092-8674(88)80013-8
Moldovan, G. L., Pfander, B., & Jentsch, S. (2007). PCNA, the Maestro of the Replication Fork. *Cell*, *129*(4), 665–679. https://doi.org/10.1016/j.cell.2007.05.003
Monot, C., Kuciak, M., Viollet, S., Mir, A. A., Gabus, C., Darlix, J. L., & Cristofari, G. (2013). The Specificity and Flexibility of L1 Reverse Transcription Priming at Imperfect T-Tracts. *PLoS Genetics*, *9*(5). https://doi.org/10.1371/journal.pgen.1003499
Moran, J. V., Holmes, S. E., Naas, T. P., DeBerardinis, R. J., Boeke, J. D., & Kazazian, H. H. (1996). High frequency retrotransposition in cultured mammalian cells. *Cell*, *87*(5), 917–927. https://doi.org/10.1016/S0092-8674(00)81998-4
Moran, J. V., DeBerardinis, R. J., & Kazazian, H. H. (1999). Exon shuffling by L1 retrotransposition. *Science*, *283*(5407), 1530–1534. https://doi.org/10.1126/science.283.5407.1530
Moran, J. V., Holmes, S. E., Naas, T. P., DeBerardinis, R. J., Boeke, J. D., & Kazazian, H. H. (1996). High Frequency Retrotransposition in Cultured Mammalian Cells. *Cell*, *87*(5), 917–927. https://doi.org/10.1016/S0092-8674(00)81998-4
Morrish, T. A., Gilbert, N., Myers, J. S., Vincent, B. J., Stamato, T. D., Taccioli, G. E., … Moran, J. V. (2002). DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nature Genetics*, *31*(2), 159–165. https://doi.org/10.1038/ng898
Morse, B., Rotherg, P. G., South, V. J., Spandorfer, J. M., & Astrin, S. M. (1988). Insertional mutagenesis of the myc locus by a LINE-1 sequence in a human breast carcinoma. *Nature*, *333*(6168), 87–90. https://doi.org/10.1038/333087a0
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., Marchetto, M. C. N., Coufal, N. G., Oefner, R., Yeo, G., Nakashima, K., & Gage, F. H. (2010). L1 retrotransposition in neurons is modulated by MeCP2. *Nature*, *468*(7322), 443–446. https://doi.org/10.1038/nature09544

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

Nevers, P., & Saedler, H. (1977). Transposable genetic elements as agents of gene instability and chromosomal rearrangements. *Nature*. *268*(5616), 109-115. https://doi.org/10.1038/268109a0

Nguyen, H. D., Yadav, T., Giri, S., Saez, B., Graubert, T. A., & Zou, L. (2017). Functions of Replication Protein A as a Sensor of R Loops and a Regulator of RNaseH1. *Molecular Cell*, *65*(5), 832-847. https://doi.org/10.1016/j.molcel.2017.01.029

Nguyen, T. H. M., Carreira, P. E., Sanchez-Luque, F. J., Schauer, S. N., Fagg, A. C., Richardson, S. R., … Faulkner, G. J. (2018). L1 Retrotransposon Heterogeneity in Ovarian Tumor Cell Evolution. *Cell Reports*, *23*(13), 3730–3740. https://doi.org/10.1016/j.celrep.2018.05.090

Nick Mcelhinny, S., Kumar, D., Clark, A., Watt, D., Watts, B., Lundström, E., … Kunkel, T. (2010). Genome instability due to ribonucleotide incorporation into DNA. *Nature Chemical Biology*, *6*(10), 774-81 . https://doi.org/10.1038/nchembio.424

Norton, W., & Bally-Cuif, L. (2010). Adult zebrafish as a model organism for behavioural genetics. *BMC Neuroscience*, *11*(1), 90. https://doi.org/10.1186/1471-2202-11-90

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment. https://doi.org/10.1006/jmbi.2000.4042

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., … Phillippy, A. M. (2021). The complete sequence of a human genome. *BioRxiv*.

O'Donnell, K. A., An, W., Schrum, C. T., Wheelan, S. J., & Boeke, J. D. (2013). Controlled insertional mutagenesis using a LINE-1 (ORFeus) gene-trap mouse model. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(29), 2706-2713. https://doi.org/10.1073/pnas.1302504110

Ogino, S., Nosho, K., Kirkner, G. J., Kawasaki, T., Chan, A. T., Schernhammer, E. S., … Fuchs, C. S. (2008). A cohort study of tumoral LINE-1 hypomethylation and prognosis in colon cancer. *Journal of the National Cancer Institute*, *100*(23), 1734–1738. https://doi.org/10.1093/jnci/djn359

Okada, N., Hamada, M., Ogiwara, I., & Ohshima, K. (1997). SINEs and LINEs share common 3' sequences: a review. *Gene*, *205*(1-2), 229-43. https://doi.org/10.1016/S0378-1119(97)00409-5

Olivares, M., García-Pérez, J. L., Thomas, M. C., Heras, S. R., & López, M. C. (2002). The non-LTR (long terminal repeat) retrotransposon L1Tc from Trypanosoma cruzi codes for a protein with RNase H activity. *The Journal of Biological Chemistry*, *277*(31), 28025–28030. https://doi.org/10.1074/jbc.M202896200

Olivares, M., Thomas, M. C., Alonso, C., & López, M. C. (1999). The L1Tc, long interspersed nucleotide element from Trypanosoma cruzi, encodes a protein with 3'-phosphatase and 3'-phosphodiesterase enzymatic activities. *Journal of Biological Chemistry*, *274*(34), 23883–23886. https://doi.org/10.1074/jbc.274.34.23883

Orecchini, E., Doria, M., Antonioni, A., Galardi, S., Ciafré, S. A., Frassinelli, L., … Michienzi, A. (2017). ADAR1 restricts LINE-1 retrotransposition. *Nucleic Acids Research*, *45*(1), 155–168. https://doi.org/10.1093/nar/gkw834

Ostertag, E. M. (2000). Determination of L1 retrotransposition kinetics in cultured cells. *Nucleic Acids Research*, *28*(6), 1418–1423. https://doi.org/10.1093/nar/28.6.1418

Ostertag, Eric M., DeBerardinis, R. J., Goodier, J. L., Zhang, Y., Yang, N., Gerton, G. L., &

Kazazian, H. H. (2002). A mouse model of human L1 retrotransposition. *Nature Genetics*, *32*(4), 655–660. https://doi.org/10.1038/ng1022

Ostertag, Eric M., Goodier, J. L., Zhang, Y., & Kazazian, H. H. (2003). SVA Elements Are Nonautonomous Retrotransposons that Cause Disease in Humans. *American Journal of Human Genetics*, *73*(6), 1444–1451. https://doi.org/10.1086/380207

Otsu, M., Kajikawa, M., Okada, N., & Kawai, G. (2017). Solution structure of a reverse transcriptase recognition site of a LINE RNA from zebrafish. *Journal of Biochemistry*, *162*(4), 279–285. https://doi.org/10.1093/jb/mvx026

Parajuli, S., Teasley, D. C., Murali, B., Jackson, J., Vindigni, A., & Stewart, S. A. (2017). Human ribonuclease H1 resolves R-loops and thereby enables progression of the DNA replication fork. *Journal of Biological Chemistry*, *292*(37), 15216–15224. https://doi.org/10.1074/jbc.M117.787473

Paterson, A. L., Weaver, J. M. J., Eldridge, M. D., Tavaré, S., Fitzgerald, R. C., & Edwards, P. A. W. (2015). Mobile element insertions are frequent in oesophageal adenocarcinomas and can mislead paired-end sequencing analysis. *BMC Genomics*, *16*(1), 1–14. https://doi.org/10.1186/s12864-015-1685-z

Pattamadilok, J., Huapai, N., Rattanatanyong, P., Vasurattana, A., Triratanachat, S., Tresukosol, D., & Mutirangura, A. (2008). LINE-1 hypomethylation level as a potential prognostic factor for epithelial ovarian cancer. *International Journal of Gynecological Cancer*, *18*(4), 711–717. https://doi.org/10.1111/j.1525-1438.2007.01117.x

Percharde, M., Lin, C. J., Yin, Y., Guan, J., Peixoto, G. A., Bulut-Karslioglu, A., … 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

Perepelitsa-Belancio, V., & Deininger, P. (2003). RNA truncation by premature polyadenylation attenuates human mobile element activity. *Nature Genetics*, *35*(4), 363–366. https://doi.org/10.1038/ng1269

Piskareva, O., Ernst, C., Higgins, N., & Schmatchenko, V. (2013). The carboxy-terminal segment of the human LINE-1 ORF2 protein is involved in RNA binding. *FEBS Open Bio*, *3*, 433–437. https://doi.org/10.1016/j.fob.2013.09.005

Pitkänen, E., Cajuso, T., Katainen, R., Kaasinen, E., Välimäki, N., Palin, K., … Kilpivaara, O. (2014). Frequent L1 retrotranspositions originating from TTC28 in colorectal cancer. *Oncotarget*, *5*(3), 853–859. https://doi.org/10.18632/oncotarget.1781

Potter, S. S., Brorein, W. J., Dunsmuir, P., & Rubin, G. M. (1979). Transposition of elements of the 412, copia and 297 dispersed repeated gene families in drosophila. *Cell*, *17*(2), 415–427. https://doi.org/10.1016/0092-8674(79)90168-5

Prak, E. T. L., Dodson, A. W., Farkash, E. A., & Kazazian, H. H. (2003). Tracking an embryonic L1 retrotransposition event. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(4), 1832–1837. https://doi.org/10.1073/pnas.0337627100

Prestel, A., Wichmann, N., Martins, J. M., Marabini, R., Kassem, N., Broendum, S. S., … Kragelund, B. B. (2019). The PCNA interaction motifs revisited: thinking outside the PIP-box. *Cellular and Molecular Life Sciences*, *76*(24), 4923–4943. https://doi.org/10.1007/s00018-019-03150-0

Raiz, J., Damert, A., Chira, S., Held, U., Klawitter, S., Hamdorf, M., … Schumann, G. G. (2012). The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Research*, *40*(4), 1666–1683. https://doi.org/10.1093/nar/gkr863

Rebollo, R., Karimi, M. M., Bilenky, M., Gagnier, L., Miceli-Royer, K., Zhang, Y., … Mager, D. L. (2011). Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. *PLoS Genetics*, *7*(9), e1002301.

https://doi.org/10.1371/journal.pgen.1002301
Rebollo, R., Romanish, M. T., & Mager, D. L. (2012). Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annual Review of Genetics*, *46*, 21–42. https://doi.org/10.1146/annurev-genet-110711-155621
Reijns, M. A., & Jackson, A. P. (2014). Ribonuclease H2 in health and disease. *Biochemical Society Transactions*, *42*(4), 717 – 725.
Reijns, M., Rabe, B., Rigby, R. E., Mill, P., Astell, K. R., Lettice, L. A., … Jackson, A. P. (2012). Enzymatic Removal of Ribonucleotides from DNA Is Essential for Mammalian Genome Integrity and Development. *Cell*, *149*(5), 1008–1022. https://doi.org/10.1016/J.CELL.2012.04.011
Richardson, S. R., Doucet, A. J., Kopera, H. C., Moldovan, J. B., Garcia-Perez, J. L., & Moran, J. V. (2015). The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiology Spectrum*, *3*(2), MDNA3-0061–2014. https://doi.org/10.1128/microbiolspec.MDNA3-0061-2014
Richardson, S. R., Gerdes, P., Gerhardt, D. J., Sanchez-Luque, F. J., Bodea, G. O., Muñoz-Lopez, M., … Faulkner, G. J. (2017). Heritable L1 retrotransposition in the mouse primordial germline and early embryo. *Genome Research*, *27*(8), 1395–1405. https://doi.org/10.1101/gr.219022.116
Richardson, S. R., Narvaiza, I., Planegger, R. A., Weitzman, M. D., & Moran, J. V. (2014). APOBEC3A deaminates transiently exposed single-strand DNA during LINE-1 retrotransposition. *ELife*, *3*, e02008. https://doi.org/10.7554/eLife.02008
Rodić, N., Sharma, R., Sharma, R., Zampella, J., Dai, L., Taylor, M. S., … Burns, K. H. (2014). Long interspersed element-1 protein expression is a hallmark of many human cancers. *American Journal of Pathology*, *184*(5), 1280–1286. https://doi.org/10.1016/j.ajpath.2014.01.007
Rodić, N., Steranka, J. P., Makohon-moore, A., Moyer, A., Sharma, R., Kohutek, Z. A., … Iacobuzio-, C. A. (2015). Retrotransposon insertions in the clonal evolution of pancreatic ductal adenocarcinoma. *Nature Medicine*, *21*(9), 1060–1064. https://doi.org/10.1038/nm.3919.Retrotransposon
Rodriguez-Martin, B., Alvarez, E. G., Baez-Ortega, A., Zamora, J., Supek, F., Demeulemeester, J., … Rodriguez-Martin, B. (2020). Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nature Genetics*, *52*(3), 306–319. https://doi.org/10.1038/s41588-019-0562-0
Roy-Engel, A. M., El-Sawy, M., Farooq, L., Odom, G. L., Perepelitsa-Belancio, V., Bruch, H., … Deininger, P. L. (2005). Human retroelements may introduce intragenic polyadenylation signals. *Cytogenetic and Genome Research*, *110*(1–4), 365–371. https://doi.org/10.1159/000084968
Sahakyan, A. B., Murat, P., Mayer, C., & Balasubramanian, S. (2017). G-quadruplex structures within the 3′ UTR of LINE-1 elements stimulate retrotransposition. *Nature Structural and Molecular Biology*, *24*(3), 243–247. https://doi.org/10.1038/nsmb.3367
Sanchez-Luque, F. J., Kempen, M. J. H. C., Gerdes, P., Vargas-Landin, D. B., Richardson, S. R., Troskie, R. L., … Faulkner, G. J. (2019). LINE-1 Evasion of Epigenetic Repression in Humans. *Molecular Cell*, *75*(3), 590-604.e12. https://doi.org/10.1016/j.molcel.2019.05.024
Sánchez-Luque, F. J., López, M. C., MacIas, F., Alonso, C., & Thomas, M. C. (2011). Identification of an hepatitis delta virus-like ribozyme at the mRNA 5′-end of the L1Tc retrotransposon from Trypanosoma cruzi. *Nucleic Acids Research*, *39*(18), 8065–8077. https://doi.org/10.1093/nar/gkr478
Sassaman, D. M., Dombroski, B. A., Moran, J. V, Kimberland, M. L., Naas, T. P., DeBerardinis, R. J., … Kazazian, H. H. (1997). Many human L1 elements are capable of

retrotransposition. *Nature Genetics*, *16*(1), 37–43. https://doi.org/10.1038/ng0597-37
Schauer, S. N., Carreira, P. E., Shukla, R., Gerhardt, D. J., Gerdes, P., Sanchez-Luque, F. J., … Faulkner, G. J. (2018). L1 retrotransposition is a common feature of mammalian hepatocarcinogenesis. *Genome Research*, *28*(5), 639–653. https://doi.org/10.1101/gr.226993.117
Schrodinger, E. (1944). *What is Life?*
Scott, A. F., Schmeckpeper, B. J., Abdelrazik, M., Comey, C. T., O'Hara, B., Rossiter, J. P., … Margolet, L. (1987). Origin of the human L1 elements: Proposed progenitor genes deduced from a consensus DNA sequence. *Genomics*, *1*(2), 113–125. https://doi.org/10.1016/0888-7543(87)90003-6
Scott, E. C., & Devine, S. E. (2017). The role of somatic L1 retrotransposition in human cancers. *Viruses*, *9*(6), 1–19. https://doi.org/10.3390/v9060131
Scott, E. C., Gardner, E. J., Masood, A., Chuang, N. T., Vertino, P. M., & Devine, S. E. (2016). A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Research*, *26*(6), 745–755. https://doi.org/10.1101/gr.201814.115
Seleme, M. del C., Vetter, M. R., Cordaux, R., Bastone, L., Batzer, M. A., & Kazazian, H. H. (2006). Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(17), 6611–6616. https://doi.org/10.1073/pnas.0601324103
Shen, W., Sun, H., De Hoyos, C. L., Bailey, J. K., Liang, X. H., & Crooke, S. T. (2017). Dynamic nucleoplasmic and nucleolar localization of mammalian RNase H1 in response to RNAP I transcriptional R-loops. *Nucleic Acids Research*, *45*(18), 10672–10692. https://doi.org/10.1093/nar/gkx710
Shiloh, Y. (2001). ATM (ataxia telangiectasia mutated): expanding roles in the DNA damage response and cellular homeostasis. *Biochemical Society Transactions*, *29*(6), 661–666. https://doi.org/10.1042/bst0290661
Shukla, R., Upton, K. R., Muñoz-Lopez, M., Gerhardt, D. J., Fisher, M. E., Nguyen, T., … Faulkner, G. J. (2013). Endogenous Retrotransposition Activates Oncogenic Pathways in Hepatocellular Carcinoma. *Cell*, *153*(1), 101–111. https://doi.org/10.1016/j.cell.2013.02.032
Singer, M. F. (1982, March 1). SINEs and LINEs: Highly repeated short and long interspersed sequences in mammalian genomes. *Cell*. Cell Press. https://doi.org/10.1016/0092-8674(82)90194-5
Singer, M. F., Krek, V., McMillan, J. P., Swergold, G. D., & Thayer, R. E. (1993). LINE-1: a human transposable element. *Gene*, *135*(1–2), 183–188. https://doi.org/10.1016/0378-1119(93)90064-A
Skowronski, J, Fanning, T. G., & Singer, M. F. (1988). Unit-length line-1 transcripts in human teratocarcinoma cells. *Molecular and Cellular Biology*, *8*(4), 1385–1397. https://doi.org/10.1128/mcb.8.4.1385-1397.1988
Skowronski, Jacek, & Singer, M. F. (1985). Expression of a cytoplasmic LINE-1 transcript is regulated in a human teratocarcinoma cell line (repeated DNA/transcription/differentiation). *Proceedings of the National Academy of Sciences USA, 82*(18):6050-6054. https://doi.org/10.1073/pnas.82.18.6050
Smyshlyaev, G., Voigt, F., Blinov, A., Barabas, O., & Novikova, O. (2013). Acquisition of an Archaea-like ribonuclease H domain by plant L1 retrotransposons supports modular evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(50), 20140–20145. https://doi.org/10.1073/pnas.1310958110
Solyom, S., Ewing, A. D., Rahrmann, E. P., Doucet, T., Nelson, H. H., Burns, M. B., … Kazazian, H. H. (2012). Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Research*, *22*(12), 2328–2338. https://doi.org/10.1101/gr.145235.112

Sookdeo, A., Hepp, C. M., McClure, M. A., & Boissinot, S. (2013). Revisiting the evolution of mouse LINE-1 in the genomic era. *Mobile DNA*, *4*(1), 1–15. https://doi.org/10.1186/1759-8753-4-3

Speek, M. (2001). Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Molecular and Cellular Biology*, *21*(6), 1973–1985. https://doi.org/10.1128/MCB.21.6.1973-1985.2001

Steinhoff, C., & Schulz, W. A. (2003). Transcriptional regulation of the human LINE-1 retrotransposon L1.2B. *Molecular Genetics and Genomics : MGG*, *270*(5), 394–402. https://doi.org/10.1007/s00438-003-0931-2

Stetson, D. B., Ko, J. S., Heidmann, T., & Medzhitov, R. (2008). Trex1 prevents cell-intrinsic initiation of autoimmunity. *Cell*, *134*(4), 587–598. https://doi.org/10.1016/j.cell.2008.06.032.Trex1

Sudhindar, P. D., Wainwright, D., Saha, S., Howarth, R., McCain, M., Bury, Y., … Shukla, R. (2021). HCV Activates Somatic L1 Retrotransposition—A Potential Hepatocarcinogenesis Pathway. *Cancers,* *13*(20), 5079. https://doi.org/10.3390/CANCERS13205079

Sugano, T., Kajikawa, M., & Okada, N. (2006). Isolation and characterization of retrotransposition-competent LINEs from zebrafish. *Gene*, *365*(1-2), 74–82. https://doi.org/10.1016/j.gene.2005.09.037

Suzuki, J., Yamaguchi, K., Kajikawa, M., Ichiyanagi, K., Adachi, N., Koyama, H., … Okada, N. (2009). Genetic Evidence That the Non-Homologous End-Joining Repair Pathway Is Involved in LINE Retrotransposition. *PLoS Genet*, *5*(4). https://doi.org/10.1371/journal.pgen.1000461

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

Tang, Z., Steranka, J. P., Ma, S., Grivainis, M., Rodić, N., Huang, C. R. L., … Burns, K. H. (2017). Human transposon insertion profiling: Analysis, visualization and identification of somatic LINE-1 insertions in ovarian cancer. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(5), 733–740. https://doi.org/10.1073/pnas.1619797114

Taylor, M. S., LaCava, J., Mita, P., Molloy, K. R., Huang, C. R. L., Li, D., … Dai, L. (2013). Affinity proteomics reveals human host factors implicated in discrete stages of LINE-1 retrotransposition. *Cell*, *155*(5), 1034–1048. https://doi.org/10.1016/j.cell.2013.10.021

Tchénio, T., Casella, J. F., & Heidmann, T. (2000). Members of the SRY family regulate the human LINE retrotransposons. *Nucleic Acids Research*, *28*(2), 411–415. https://doi.org/10.1093/nar/28.2.411

Thomas, C. A., Tejwani, L., Trujillo, C. A., Negraes, P. D., Herai, R. H., Mesci, P., … Muotri, A. R. (2017). Modeling of TREX1-Dependent Autoimmune Disease using Human Stem Cells Highlights L1 Accumulation as a Source of Neuroinflammation. *Cell Stem Cell*, *21*(3), 319-331. https://doi.org/10.1016/j.stem.2017.07.009

Tollis, M., & Boissinot, S. (2012). The Evolutionary Dynamics of Transposable Elements in Eukaryote Genomes. *Genome Dynamics, 7,* 68-91. https://doi.org/10.1159/000337126

Treangen, T. J., & Salzberg, S. L. (2013). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, *13*(1), 36–46. https://doi.org/10.1038/nrg3117.Repetitive

Trelogan, S. A., & Martin, S. L. (1995). Tightly regulated, developmentally specific expression of the first open reading frame from LINE-1 during mouse embryogenesis. *Proceedings of the National Academy of Sciences*, *92*(5), 1520–1524. https://doi.org/10.1073/pnas.92.5.1520

Tristán-Ramos, P., Rubio-Roldan, A., Peris, G., Sánchez, L., Amador-Cubero, S., Viollet, S., … Heras, S. R. (2020). The tumor suppressor microRNA let-7 inhibits human LINE-1 retrotransposition. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-19430-4

Tsuchimoto, D., Sakai, Y., Sakumi, K., Nishioka, K., Sasaki, M., Fujiwara, T., & Nakabeppu, Y. (2001). Human APE2 protein is mostly localized in the nuclei and to some extent in the mitochondria, while nuclear APE2 is partly associated with proliferating cell nuclear antigen. *Nucleic Acids Research*, *29*(11), 2349–2360. https://doi.org/10.1093/nar/29.11.2349

Tsurimoto, T., & Stillman, B. (1990). Functions of replication factor C and proliferating-cell nuclear antigen: Functional similarity of DNA polymerase accessory proteins from human cells and bacteriophage T4. *Proceedings of the National Academy of Sciences of the United States of America*, *87*(3), 1023–1027. https://doi.org/10.1073/pnas.87.3.1023

Tubio, J. M. C., Li, Y., Ju, Y. S., Martincorena, I., Cooke, S. L., Tojo, M., … Campbell, P. . J. (2014). Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science*, *345*(6196), 1251343. https://doi.org/10.1126/science.1251343.Extensive

Tunbak, H., Enriquez-Gasca, R., Tie, C. H. C., Gould, P. A., Mlcochova, P., Gupta, R. K., … Rowe, H. M. (2020). The HUSH complex is a gatekeeper of type I interferon through epigenetic regulation of LINE-1s. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-19170-5

Umar, A., Buermeyer, A. B., Simon, J. A., Thomas, D. C., Clark, A. B., Liskay, R. M., & Kunkel, T. A. (1996). Requirement for PCNA in DNA mismatch repair at a step preceding DNA resynthesis. *Cell*, *87*(1), 65–73. https://doi.org/10.1016/S0092-8674(00)81323-9

Unk, I., Haracska, L., Gomes, X. V., Burgers, P. M. J., Prakash, L., & Prakash, S. (2002). Stimulation of 3′→5′ Exonuclease and 3′-Phosphodiesterase Activities of Yeast Apn2 by Proliferating Cell Nuclear Antigen. *Molecular and Cellular Biology*, *22*(18), 6480–6486. https://doi.org/10.1128/mcb.22.18.6480-6486.2002

Upton, K. R., Gerhardt, D. J., Jesuadian, J. S., Richardson, S. R., Sánchez-Luque, F. J., Bodea, G. O., … 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

van den Hurk, J. a J. M., Meij, I. C., del Carmen Seleme, M., Kano, H., Nikopoulos, K., Hoefsloot, L. H., … Cremers, F. P. M. (2007). L1 retrotransposition can occur early in human embryonic development. *Human Molecular Genetics*, *16*(13), 1587–1592. https://doi.org/10.1093/hmg/ddm108

van Hoesel, Q. A., van de Velde, C. J. H., Kuppen, P. J. K., Liefers, G. J., Putter, H., Sato, Y., … Hoon, D. S. B. (2012). Hypomethylation of LINE-1 in primary tumor has poor prognosis in young breast cancer patients: a retrospective cohort study. *Breast Cancer Research and Treatment*, *134*(3), 1103–1114.

Van Meter, M., Kashyap, M., Rezazadeh, S., Geneva, A. J., Morello, T. D., Seluanov, A., & Gorbunova, V. (2014). SIRT6 represses LINE1 retrotransposons by ribosylating KAP1 but this repression fails with stress and age. *Nature Communications*, *5*, 5011. https://doi.org/10.1038/ncomms6011

Van Valen, L. (1973). A new evolutionary law, *1*, 1–30.

Viollet, S., Monot, C., & Cristofari, G. (2014). L1 retrotransposition: The snap-velcro model and its consequences. *Mobile Genetic Elements*, *4*(2), e28907. https://doi.org/10.4161/mge.28907

Volkman, H. E., & Stetson, D. B. (2014). The enemy within: Endogenous retroelements and autoimmune disease. *Nature Immunology, 15,* 415–422. https://doi.org/10.1038/ni.2872

Wallace, B. D., Berman, Z., Mueller, G. A., Lin, Y., Chang, T., Andres, S. N., … Scott

Williams, R. (2017). APE2 Zf-GRF facilitates 3'-5' resection of DNA damage following oxidative stress. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(2), 304–309. https://doi.org/10.1073/pnas.1610011114

Wallace, N. A., Belancio, V. P., & Deininger, P. L. (2008). L1 mobile element expression causes multiple types of toxicity. *Gene*, *419*(1–2), 75–81. https://doi.org/10.1016/j.gene.2008.04.013

Warbrick, E., Lane, D. P., Glover, D. M., & Cox, L. S. (1997). Homologous regions of Fen1 and p21(Cip1) compete for binding to the same site on PCNA: A potential mechanism to co-ordinate DNA replication and repair. *Oncogene*, *14*(19), 2313–2321. https://doi.org/10.1038/sj.onc.1201072

Ward, J. R., Vasu, K., Deutschman, E., Halawani, D., Larson, P. A., Zhang, D., … Longworth, M. S. (2017). Condensin II and GAIT complexes cooperate to restrict LINE-1 retrotransposition in epithelial cells. *PLoS Genetics*, *13*. https://doi.org/10.1371/journal.pgen.1007051

Warkocki, Z., Krawczyk, P. S., Adamska, D., Bijata, K., Garcia-Perez, J. L., & Dziembowski, A. (2018). Uridylation by TUT4/7 Restricts Retrotransposition of Human LINE-1s. *Cell*, *174*(6), 1537-1548. https://doi.org/10.1016/j.cell.2018.07.022

Wei, W., Gilbert, N., Ooi, S. L., Lawler, J. F., Ostertag, E. M., Kazazian, H. H., … Moran, J. V. (2001). Human L1 retrotransposition: cis preference versus trans complementation. *Molecular and Cellular Biology*, *21*(4), 1429–1439. https://doi.org/10.1128/MCB.21.4.1429-1439.2001

Weichenrieder, O., Repanas, K., & Perrakis, A. (2004). Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. *Structure*, *12*(6), 975–986. https://doi.org/10.1016/j.str.2004.04.011

Weinreb, J. T., Ghazale, N., Pradhan, K., Gupta, V., Potts, K. S., Tricomi, B., … Bowman, T. V. (2021). Excessive R-loops trigger an inflammatory cascade leading to increased HSPC production. *Developmental Cell*, *56*(5), 627-640.e5. https://doi.org/10.1016/j.devcel.2021.02.006

Wheelan, S. J., Aizawa, Y., Han, J. S., & Boeke, J. D. (2005). Gene-breaking: A new paradigm for human retrotransposon-mediated gene evolution. *Genome Research*, *15*(8), 1073–1078. https://doi.org/10.1101/gr.3688905

Wheeler, D. L., Church, D. M., Federhen, S., Lash, A. E., Madden, T. L., Pontius, J. U., … Wagner, L. (2003). Database resources of the National Center for Biotechnology. *Nucleic Acids Research*, *31*(1), 28–33. https://doi.org/10.1093/nar/gkg033

White, R. M., Sessa, A., Burke, C., Bowman, T., LeBlanc, J., Ceol, C., … Zon, L. I. (2008). Transparent adult zebrafish as a tool for in vivo transplantation analysis. *Cell Stem Cell*, *2*(2), 183–189. https://doi.org/10.1016/j.stem.2007.11.002

Wylie, A., Jones, A. E., D'Brot, A., Lu, W.-J., Kurtz, P., Moran, J. V, … Abrams, J. M. (2016). p53 genes function to restrain mobile elements. *Genes & Development*, *30*(1), 64–77. https://doi.org/10.1101/gad.266098.115

Xie, Y., Rosser, J. M., Thompson, T. L., Boeke, J. D., & An, W. (2011). Characterization of L1 retrotransposition with high-throughput dual-luciferase assays. *Nucleic Acids Research*, *39*(3), e16–e16. https://doi.org/10.1093/nar/gkq1076

Xiong, Y., & Eickbush, T. H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO Journal*, *9*(10), 3353–3362. https://doi.org/10.1002/j.1460-2075.1990.tb07536.x

Xu, M., You, Y., Hunsicker, P., Hori, T., Small, C., Griswold, M. D., & Hecht, N. B. (2008). Mice deficient for a small cluster of Piwi-lnteracting RNAs Implicate Piwi-interacting RNAs in transposon control. *Biology of Reproduction*, *79*(1), 51–57. https://doi.org/10.1095/biolreprod.108.068072

Yang, N., Zhang, L., Zhang, Y., & Kazazian, H. H. (2003). An important role for RUNX3 in human L1 transcription and retrotransposition. *Nucleic Acids Research*, *31*(16), 4929–4940. https://doi.org/10.1093/nar/gkg663

Yoder, J. A., Walsh, C. P., & Bestor, T. H. (1997). Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics*, *13*(8), 335–340. https://doi.org/10.1016/s0168-9525(97)01181-5

Yooyongsatit, S., Ruchusatsawat, K., Noppakun, N., Hirankarn, N., Mutirangura, A., & Wongpiyabovorn, J. (2015). Patterns and functional roles of LINE-1 and Alu methylation in the keratinocyte from patients with psoriasis vulgaris. *Journal of Human Genetics*, *60*(7), 349–355. https://doi.org/10.1038/jhg.2015.33

Young, R., Way, J., Way, S., Yin, J., & Syvanen, M. (1979). Transposition mutagenesis of bacteriophage lambda. A new gene affecting cell lysis. *Journal of Molecular Biology*, *132*(3), 307–322. https://doi.org/10.1016/0022-2836(79)90262-6

Yu, F. (2001). Methyl-CpG-binding protein 2 represses LINE-1 expression and retrotransposition but not Alu transcription. *Nucleic Acids Research*, *29*(21), 4493–4501. https://doi.org/10.1093/nar/29.21.4493

Yu, F., Zingler, N., Schumann, G., & Strätling, W. H. (2001). Methyl-CpG-binding protein 2 represses LINE-1 expression and retrotransposition but not Alu transcription. *Nucleic Acids Research*, *29*(21), 4493–4501. https://doi.org/10.1093/nar/29.21.4493

Zaki-Dizaji, M., Akrami, S. M., Azizi, G., Abolhassani, H., & Aghamohammadi, A. (2018). Inflammation, a significant player of Ataxia–Telangiectasia pathogenesis? *Inflammation Research*, *67*(7), 559–570. https://doi.org/10.1007/s00011-018-1142-y

Zhao, B., Wu, Q., Ye, A. Y., Guo, J., Zheng, X., Yang, X., … Huang, A. Y. (2019). Somatic LINE-1 retrotransposition in cortical neurons and non-brain tissues of Rett patients and healthy individuals. *PLoS Genetics*, *15*. https://doi.org/10.1371/journal.pgen.1008043

Zhao, K., Du, J., Han, X., Goodier, J. L., Li, P., Zhou, X., … Yu, X.-F. (2013). Modulation of LINE-1 and Alu/SVA retrotransposition by Aicardi-Goutières syndrome-related SAMHD1. *Cell Reports*, *4*(6), 1108–1115. https://doi.org/10.1016/j.celrep.2013.08.019

Zhong, J., & Lambowitz, A. M. (2003). Group II intron mobility using nascent strands at DNA replication forks to prime reverse transcription. *EMBO Journal*, *22*(17), 4555–4565. https://doi.org/10.1093/emboj/cdg433