Transposable elements as hidden neuronal gene regulators in health and disease

van Bree, E.J.

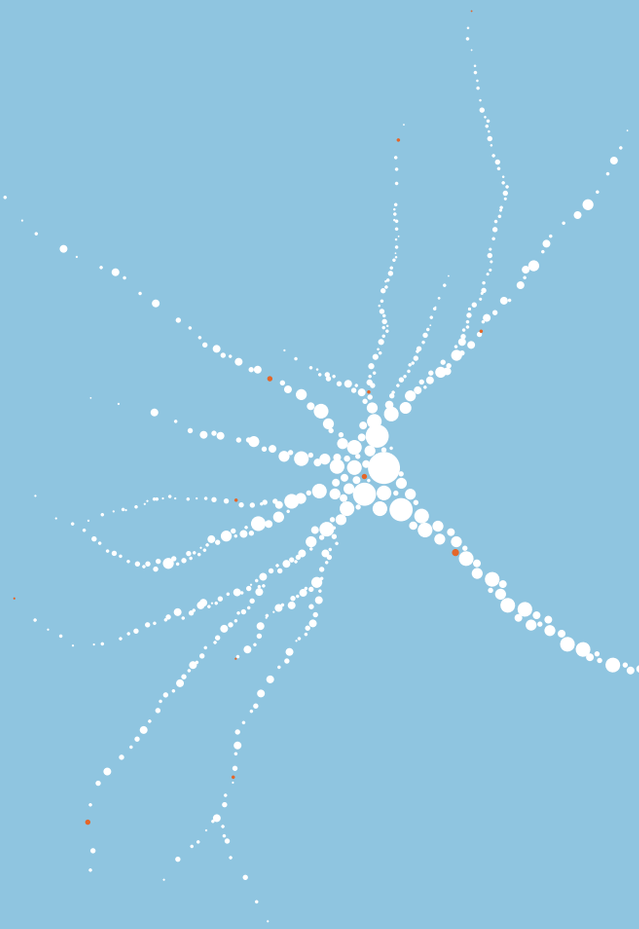
Publication date
2022

TRANSPOSABLE ELEMENTS

as hidden neuronal gene regulators in health and disease

Transposable elements as hidden neuronal gene regulators in health and disease

Elise van Bree



Elise van Bree

Transposable elements as hidden neuronal gene regulators in health and disease

E.J. van Bree

The research in this thesis was carried out at the Swammerdam Institute for Life Sciences, research institute of the Faculty of Science at the University of Amsterdam and at the Translational Research Institute, research institute of Mater Research at the University of Queensland.

The research in this thesis was financially supported by an HFSP Career Development Award (CDA00030/2016C) and ERC starting grant (ERC-2016-stG-716035) to F.M.J. Jacobs.

Printing of this thesis was financially supported by:
Alzheimer Nederland
Stichting Alkemade-Keuls
F.M.J. Jacobs

ISBN: 978-94-6421-888-6
Printed by: Ipskamp Printing, Enschede.

Transposable elements as hidden neuronal gene regulators in health and disease

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op woensdag 2 november 2022, te 13.00 uur

door Elisabeth Johanna van Bree
geboren te Amersfoort

Promotiecommissie

Promotores:	prof. dr. M.P. Smidt	Universiteit van Amsterdam
	dr. F.M.J. Jacobs	Universiteit van Amsterdam
Overige leden:	prof. dr. E.M.A. Aronica	Universiteit van Amsterdam
	prof. dr. R.E. Koes	Universiteit van Amsterdam
	prof. dr. P.J. Verschure	AMC-UvA
	prof. dr. ir. M.J.T. Reinders	Universiteit Leiden
	dr. M.P. Creyghton	Erasmus MC
	dr. V.M. Heine	Vrije Universiteit Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Contents

It’s all very complex but we’re trying to break it down and figure out how simply beautiful it is .

D. Pereira Fernandes

Introduction

1.1 The human genome

The human genome can be seen as an instruction manual for humans, containing all the information needed to assemble a human being from a single cell. Within it, you can find all the details necessary to support growth and development, and to generate all the different cell types that make up the human body. Building blocks, called proteins, are required for all these processes to occur. These assembly pieces are encoded by genes, which are known as the genomic units that have the instructions to produce proteins; the entirety of these units is called the genome. The genome is made of deoxyribonucleic acid (DNA), owing its name to the sugar (deoxyribo) and phosphate with nucleotides (nucleic acid), also known as bases, attached to it. Different combinations of these nucleotides result in a sequence that can be transcribed to produce RNA (ribonucleic acid). The RNA molecule is used to convey the genomic information to produce proteins in a subsequent process called translation. A normal human cell, except for the male and female reproductive cells, contains around 6.1 billion nucleotides (3.06 billion base pairs), divided over 23 pairs of chromosomes (Nurk et al. 2021). Together, these chromosomes carry an estimated 30.000 genes, which can encode even more proteins by for example alternative splicing of the RNA code. Understanding the human genome and its components has a major importance within biomedical research, benefiting both preventive and therapeutic health care.

In 1990, an international consortium initiated “The Human Genome Project” with the aim to generate a complete, high-quality transcript of the human genome. To achieve this, the genome was cut into small fragments and with a laborious approach involving bacterial artificial chromosome (BAC) clones, 500-800 base pairs were read in one stretch in a process called sequencing. The so-called ‘reads’ that were produced were then assembled together with the use of a computer. Thirteen years and 2.7 billion dollars later, a first draft of the human genome was available that included 99% of the total genome (International Human Genome Sequencing Consortium 2001). Still, many gaps remained, which were impossible to fill with the then-used sequencing techniques. In 2015, researchers showed that a new sequencing technique, generating reads which were on average 5.8 thousand base pairs long (5.8 kb), could be used to resolve the remainder of the human genome sequence (Chaisson et al. 2015). Over the years, efforts were made to improve this technique, resulting in longer reads with a higher accuracy. Gaps were filled and finally researchers were able to assemble complete human chromosomes (Wenger et al. 2019; Miga et al. 2020; Logsdon et al. 2021). By the end of 2020, the Telomere-to-Telomere consortium announced the completion of the first, truly complete sequence of a full human genome (Nurk et al. 2021).

The efforts of the last three decades to unravel human genome sequences have pushed advances in DNA sequencing techniques, and laid the basis for researching human diseases. It became clear that even a difference of only one nucleotide in a gene (single nucleotide polymorphism) could result in disease, and furthermore revealed that molecular mechanisms underlying human diseases can be caused by multiple different genetic factors. With 99.9% of the human genome being identical between individuals, important information on human disease likely lies in genomic regions or mechanisms that differ between individuals.

1.2 Species diversity through the noncoding genome

More than 150 years ago, Thomas H. Huxley was the first to discuss human evolution, with the introduction of the phenomenon of a common ancestor of humans and apes (Huxley 1863). Subsequent research has provided evidence for chimpanzees and bonobos to be the closest living relatives to humans, all emerging from a common ancestor around 6-8 million years ago (Diogo et al. 2017; Langergraber et al. 2012; Steiper and Seiffert 2012; Grabowski and Jungers 2017). Studies on the human and chimpanzee genome led to the sensational discovery that the genomes of these species closely resemble each other. Their genomes are most similar in protein-coding regions, with early research from the seventies estimating a 99% similarity (King and Wilson 1975). With extensive studies being performed on the molecular, anatomical, physiological, behavioural and ecological level, it became evident that the genetic differences in protein-coding DNA are too small to account for the organismal diversity (King and Wilson 1975). This gave rise to the idea that there are other mechanisms that account for phenotypic differences between humans and chimpanzees not captured solely by genes, a concept introduced by King and Wilson in 1975 (King and Wilson 1975; Enard et al. 2002). This was a progressive theory, as the noncoding DNA, which makes up around 98.5% of the genome, was long regarded to be 'junk DNA'. In later years, it also became of interest to researchers studying complex human disorders, as large sequencing studies revealed that the protein-coding genome could not account for the complete genetic basis of all disorders. Therefore, studying the noncoding genome, how it is involved in regulating gene expression, and how it differs between individuals might resolve a part of this so-called missing heritability problem.

1.3 Epigenetics and differential gene expression

To generate phenotypic differences from the same DNA, quantitative and qualitative differences in gene expression need to be established. This control of gene expression not encoded by the genes themselves is referred to as epigenetics. It is a widely used term that was introduced with Conrad Waddington's experiments published in 1956 that showed the inheritance of new phenotypes under the influence of environmental factors, with a presumed unchanged genetic background (Waddington 1956, 1957). This is accomplished by a cell type-specific interplay of different genes expressed at the same time, joined in so-called gene expression networks. While there are different mechanisms involved in regulating gene expression, one of them involves epigenetic modifications on histones that alter the chromatin structure, and thereby contribute to the regulation of gene expression.

When DNA is tightly packed in the nucleus, it is in a state referred to as heterochromatin. This is achieved by wrapping the DNA around histone proteins that make up nucleosomes (H2A, H2B, H3, H4). These histones have 'tails' extending out of the nucleosomes that can contact adjacent nucleosomes and undergo different post-translational modifications. These include for example methylation, acetylation, ubiquitination and phosphorylation (Gibney and Nolan 2010). Many different combinations of histones with specific modifications can be present and affect the accessibility of DNA by changing chromatin structure. With advancements in sequencing techniques, it became evident that these histone modifications mark specific regulatory regions in the DNA. Marks associated with transcriptional activation are for example tri-methylation of histone H3 at lysine

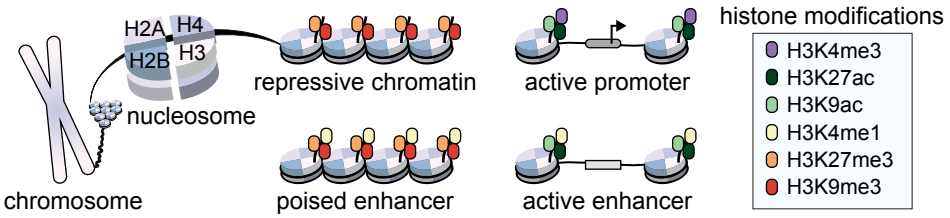


Figure 1: Examples of histone modifications marking genomic regions. Billions of base pairs of DNA containing the genetic code of humans are tightly packaged as chromosomes inside the nucleus of cells. This is established by wrapping negatively charged DNA around positively charged histone protein complexes called nucleosomes. Modifications on tails of the different histones (H2A, H2B, H3, and H4), mark specific DNA regions and can be used to distinguish functional regions. Promoters are located near transcription start sites of genes and form a hub for regulatory elements to bind and help initiate transcription. Enhancers are regions where transcription factors can bind and can promote transcription of genes from much further distances by looping to promoters (Blackwood and Kadonaga 1998; Heintzman et al. 2007).

4 (H3K4me3), and acetylation of histone H3 at lysine 27 (H3K27ac) or lysine 9 (H3K9ac) (Creyghton et al. 2010; Rada-Iglesias et al. 2011; Karmodiya et al. 2012). Heterochromatin, and therefore transcriptional repression, is associated with tri-methylation of histone H3 at lysine 27 (H3K27me3) or lysine 9 (H3K9me3) (Boyer et al. 2006; Mikkelsen et al. 2007a) (**Figure 1**).

Epigenetic studies exploring the distribution of these marks help us gain knowledge about the function of the noncoding genome. With improved DNA sequencing techniques, it became evident in the early 2000s that the noncoding genome is where most genetic variation is seen between humans and our close relative chimpanzees (International Human Genome Sequencing Consortium 2001; Waterson et al. 2005; Suntsova and Buzdin 2020). But also between different human individuals the majority of genetic variation affects the noncoding genome (The 1000 Genomes Project Consortium 2015). Epigenetic studies and studies exploring the noncoding genome are therefore important to contribute to our understanding of evolution, but also for example embryology, ageing, and diseases.

1.4 Transposable elements

Most of the human noncoding genome is derived from transposable elements (TEs): viral DNA with the (now often lost) capability to copy- or cut-and-paste in the genome (Smit 1999; International Human Genome Sequencing Consortium 2001; Deininger et al. 2003). The elements are classified based on their mode of transposition, sequence similarities and structural relationships. Class I TEs use reverse transcription to generate an RNA intermediate to copy through the genome, and include retrotransposons, retroposons and retrointrons. Class II elements on the contrary use a cut-and-paste mechanism facilitated by the enzyme transposase and are therefore DNA-based transposons (**Figure 2**) (McClure 1999; Kidwell and Lisch 2000; Finnegan 1989, 2012).

Class I retrotransposons can be further divided into two subclasses: the LTR retrotransposons, and the non-LTR transposons. The first consists of long terminal repeats (LTRs) flanking DNA sequences essential for reverse transcription. These are similar to the *gag* and *pol* genes of retroviruses, and occasionally include the *env* gene

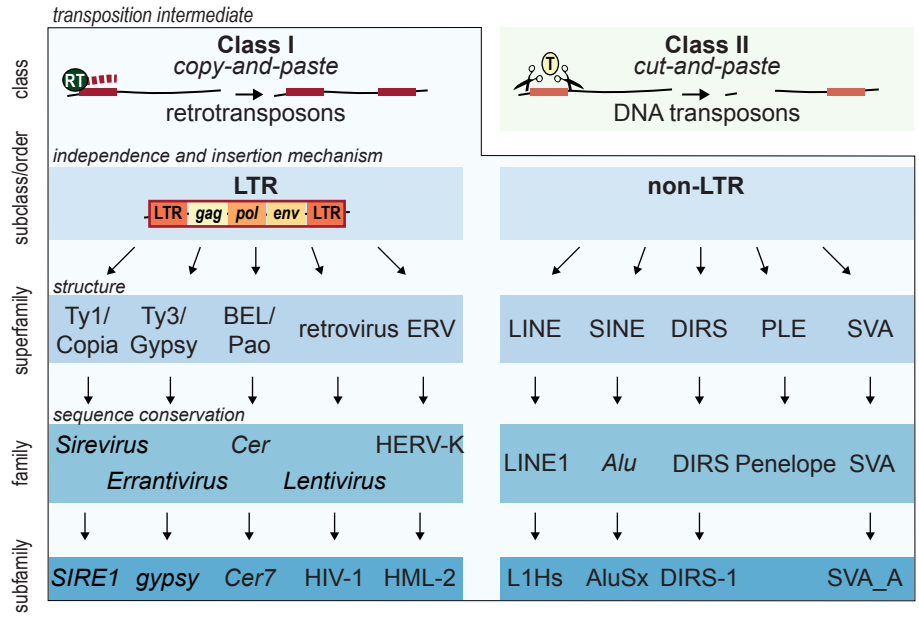


Figure 2: Overview of TE classification system. Grouping based on Wicker et al. (2007), with the emphasis on Class I retrotransposons. Only a few examples of TE (super/sub)families are shown.

(**Figure 2**). LINEs (long interspersed nuclear elements) are non-LTR retrotransposons. As the name reveals, these elements lack long terminal repeats, but they do encode a reverse transcriptase and endonuclease domain for autosomal retrotransposition. SINEs (short interspersed nuclear elements) on the contrary are non-autonomous non-LTR retrotransposons and use the LINE transposition machinery to spread through the genome. This superfamily consists of numerous families classified based on their structure. An example are the *Alu* elements, owing their name to the *Alu*I restriction site they contain (Finnegan 2012). SVA (Sine-VNTR-*Alu*) elements are another family of non-autonomous retrotransposons dependent on the LINE transposon machinery. These elements are composite transposons, containing an *Alu*-like sequence, a variable number of tandem repeat regions (VNTR) and a sequence derived from an HERV-K LTR transposon. They are the youngest family of transposable elements and are exclusively present in humans and other great apes (Wang et al. 2005). Finally, each family of transposons is further divided into subfamilies based on phylogenetics (Wicker et al. 2007). An overview of the TE classification system can be found in **Figure 2**.

1.5 Transposable elements as gene-regulatory elements

Over seventy years ago, Barbara McClintock was the first to discover mobile DNA elements that are able to influence gene expression dependent on the location of their insertion, for which she was later awarded the Nobel prize (McClintock 1950, 1956). These findings were made in maize, and it took decades before the importance of TEs in human evolution and gene-regulatory networks was widely acknowledged (Deininger et al. 2003; Feschotte 2008; Cordaux and Batzer 2009). Advances in DNA sequencing techniques provided an overview of the genome-wide location of TE-derived sequences and showed that many

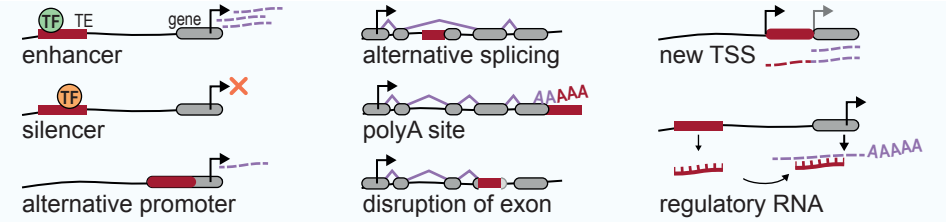


Figure 3: Transposable elements can be incorporated in gene-regulatory networks in various ways. They can for example influence the level of gene expression by functioning as enhancers, silencers or alternative promoters. They can change transcription by influencing splicing, adding an alternative polyA site, disrupting exons, or acting as alternative transcription start sites (TSSs). A last example is shown where TEs produce regulatory RNAs binding to mRNA transcripts. TE in red, TSS indicated with arrow and transcript(ion) in purple (based on Ecco et al. 2017).

reside in regulatory regions (International Human Genome Sequencing Consortium 2001; Jordan et al. 2003; Bourque et al. 2008; Trizzino et al. 2017, 2018). There they can function as enhancers or silencers through the transcription factor binding sites they carry (Thornburg et al. 2006; Bourque et al. 2008; Sundaram et al. 2014). Other mechanisms through which they can be incorporated into gene-regulatory networks are by functioning as alternative promoters, changing transcripts by alternative splicing of mRNA or adding a new polyA site, disrupting exons, and producing regulatory RNAs (Conley et al. 2008; Ecco et al. 2017) (**Figure 3**). Early research showed that over a 1,000 highly conserved eutherian-specific noncoding elements (conserved between human and mouse/rat/dog, but absent in opossum and chicken) are derived from TEs (Mikkelsen et al. 2007b). Later research showed more than 10,000 TE fragments residing close to developmental genes and genes involved in the regulation of transcription have been under strong selection since the origin of eutherians (Lowe et al. 2007; Feschotte 2008). This makes TEs a source for regulatory innovation and suggests a crucial role for the elements in gene-regulatory networks during mammalian evolution (Britten 1997; Lowe et al. 2007).

1.6 Transposable elements in disease and ageing

While most TEs are not transpositioning anymore, those belonging to the LINE-1 (L1), *Alu* and SVA families are still spreading through the genome and thereby continue to be an innovative force for human gene regulation. When this occurs in germ cells, it results in heritable TE polymorphic variation in the population. For *Alu* elements, this occurs around once in every 40 births. For L1 and SVA elements, the germline retrotransposition rate is one in 63 births (Feusier et al. 2019). New insertions can also occur in somatic cells, as mosaicism is for example seen in the brain and cancer cells (Coufal et al. 2009; Muotri et al. 2005; van den Hurk et al. 2007; Garcia-Perez et al. 2007; Baillie et al. 2011; Erwin et al. 2014; Richardson et al. 2014; Tubio et al. 2014; Steely et al. 2021). The initial discovery of a polymorphic L1 insertion in haemophilia A patients sparked the interest of researchers for the involvement of TEs in disease. Now, over 130 TE insertions have been reported to be associated with disease (Hancks and Kazazian 2016; Kazazian and Moran 2017). These include a wide variety of diagnoses, including haemophilia A and B (Kazazian et al. 1988; Nakamura et al. 2015), Alström syndrome (Taşkesen et al. 2012), X-linked dystonia-Parkinsonism (XDP) (Aneichyk et al. 2018), cystic fibrosis (Chen et al. 2008), lynch syndrome (van der Klift et al. 2012), and Duchenne muscular dystrophy (Narita et al. 1993). TEs can contribute to disease via different mechanisms, which are thought to depend

on the type, length, orientation and exact location of the TE insertion (Chen et al. 2006). L1 elements are for example shown to be associated with disease through disruption of exons and altering transcripts (Kondo-Iida et al. 1999; Ostertag and Kazazian 2001; Meischl et al. 2000; Clayton et al. 2020), or reducing transcript levels (Schwahn et al. 1998). Complete abolition of gene expression and abolishment of functional transcripts is also seen for *Alu* insertions associated with disease (Apoil et al. 2007; Mustajoki et al. 1999). In a case of XDP, an SVA inserted in an intron of the *TAF1* gene leads to intron retention and decreases *TAF1* expression levels (Aneichyk et al. 2018). Besides altering gene expression, transposition activity of TEs causes genome instability, which can have detrimental or even lethal effects. L1-mediated transduction, where imperfect cleavage of L1-derived transcripts results in the inclusion of the 3' flanking regions during retrotransposition is one mechanism through which TEs affect the host genome integrity (Moran et al. 1999; Holmes et al. 1994). Even larger alterations occur via unequal homologous recombination, facilitated by two nearby repetitive sequences (Burwinkel and Kilimann 1998; Gilbert et al. 2002).

The activity of TEs is thought to increase with ageing and believed to influence or promote the process and age-related diseases (Gorbunova et al. 2021; Simon et al. 2019; Blaudin de Thé et al. 2018; Tam et al. 2019). An age-dependent increase of TE expression is for example shown in the *Drosophila* (fruit fly) brain (Li et al. 2013), head and fat body (Chen et al. 2016; Wood et al. 2016). In mice, activation of transposable elements has also been reported upon ageing in liver and muscle tissues and age-associated cancer tissues (De Cecco et al. 2013b). The derepression of TEs upon ageing can work via multiple mechanisms. A major process generally considered to be involved is the age-related remodelling of the epigenome, resulting in a loss of heterochromatin. This is associated with transcriptional derepression throughout the genome (López-Otín et al. 2013). Interestingly, stabilisation of heterochromatin reduced the age-related increases in TE expression seen in *Drosophila*, and increased lifespan (Wood et al. 2016). Hypomethylation of DNA is a common feature of ageing in somatic cells, and is associated with increased gene expression, as DNA methylation ensures transcriptional repression of DNA regions (Bird and Wolffe 1999; Jones et al. 1998; Maegawa et al. 2010). In blood samples of elderly people, ageing was negatively correlated with DNA methylation levels of *Alu* and L1 elements. Comparison of overall methylation of *Alu* elements at two different time points also showed a decrease in methylation at the later stage compared to around 4 years earlier (Bollati et al. 2009). This decrease in methylation could promote the transposition activity of TEs. Other mechanisms that can lead to increased activity of TEs are for example the redistribution of chromatin modifiers involved in TE repression. One of those is SIRT6, a longevity regulating protein and strong repressor of L1 elements (Van Meter et al. 2014; Simon et al. 2019). During ageing, L1 activity is seen to increase, and thought to be a consequence of the reported cellular redistribution of SIRT6 proteins away from L1 loci (Van Meter et al. 2014). For SIRT1 it is believed a similar phenomenon can take place, based on observations in mice and *Drosophila* (Oberdoerffer et al. 2008; Wood et al. 2016; Gorbunova et al. 2021). In Alzheimer's disease, of which the greatest risk factor is age, widespread activation of TEs is seen (Guo et al. 2018). Tau protein, which forms the neurofibrillary tangles characteristic for this neurodegenerative disease, is shown to be sufficient for increasing TE-derived transcript levels. With TEs being the incorporation in gene-regulatory networks, knowledge

about which elements become active during ageing might provide clues about affected processes and provide therapeutic targets to increase lifespan.

1.7 KRAB-ZNFs: the guardians of the genome

The fact that new TE insertions in the genome can benefit species fitness is widely acknowledged, but their involvement in diseases shows that this also comes at a cost. Higher vertebrates have developed mechanisms to protect the genome from instability caused by TE activity. One of the key mechanisms that evolved in vertebrate genomes to repress TEs involves Krüppel-associated box (KRAB) domain zinc finger proteins (KZNFs) to transcriptionally repress TEs (Wolf and Goff 2009; Jacobs et al. 2014; Schmitges et al. 2016; Imbeault et al. 2017; Turelli et al. 2020). KZNFs are the largest transcription factor (TF) family in mammalian genomes, with over 700 structurally distinct proteins encoded by approximately 400 different loci in our genome (reviewed by Urrutia 2003; Huntley et al. 2006).

The repressive potential of KZNFs is mediated by their KRAB domain, which is made up of either one or two subdomains (Margolin et al. 1994). The A box of the KRAB domain is important for repression through recruitment of KAP1, which in turn recruits the cofactors HP1, SETDB1, and HDAC. Together this protein complex silences DNA by condensing it into heterochromatin (**Figure 4a**, Schultz et al. 2002). The B box is thought to enhance the repressive effect of the A box (Vissing et al. 1995). The A- and B-box are always encoded by separate exons, thereby providing the possibility for generating different repressive capacities by alternative splicing (Urrutia 2003; Vissing et al. 1995; Shao et al. 2006). Recognition and binding to target sites is mediated through the C_2H_2 zinc-finger domain of KZNFs. This domain consists usually of 10 or more finger-like protrusions that each recognize three to four nucleotides (**Figure 4b**). The combination of multiple zinc fingers generates a unique sequence that KZNFs can bind (Gebelein and Urrutia 2001). It is thought that the proteins do not necessarily always bind DNA with all of their zinc-fingers, thereby creating additional variations in the sequences they can recognize (Urrutia 2003).

New KZNFs arise through duplication events and subsequent functional divergence, and are often found in clusters on the genome (Shannon et al. 2003; Emerson and Thomas 2009; Nowick et al. 2010). Significant differences in the binding domains of parent-daughter KZNF pairs that have recently emerged suggests a drive for diversification of KZNF function (Nowick et al. 2010). This is thought to be facilitated by the modular design of KZNF, with separate exons coding for the different domains. A mutation in a specific zinc finger domain can thereby result in subtle changes ensuring new DNA binding capabilities, without affecting the overall protein structure important for recruitment of repressive co-factors (Pabo et al. 2001; Shannon et al. 2003). Interestingly, KZNFs likely did not start like this, as coelacanths, a 'living fossil' species of fish, mainly have mono-exonic KZNF genes (Imbeault et al. 2017).

Early work linked individual KZNFs to the repression of TEs (Wolf and Goff 2009; Tan et al. 2013; Jacobs et al. 2014; Wolf et al. 2015), and a correlation between the number of TEs and KZNFs in vertebrates shows co-evolution of the viral DNA elements and their repressors (Thomas and Schneider 2011). This contributed to the belief that an evolutionary arms

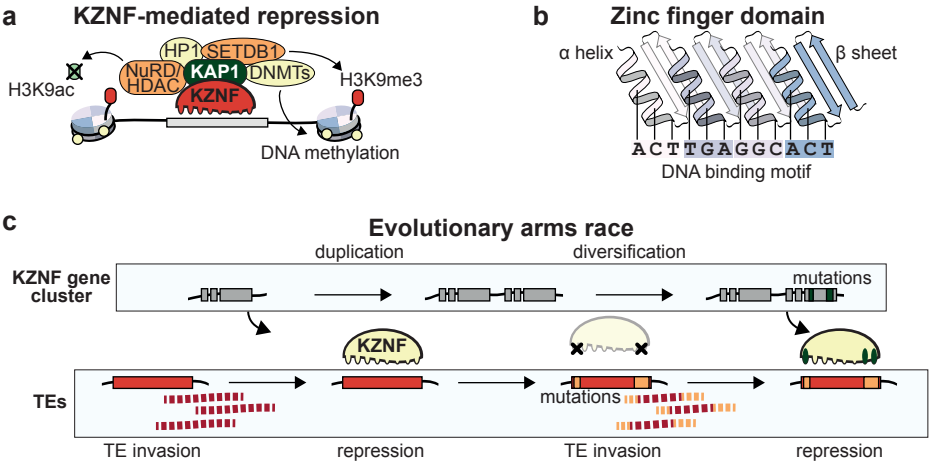


Figure 4: KZNF repression of TEs. a, The KRAB domain of KZNFs recruits KAP1, which interacts with nucleosome remodelling deacetylase complex/histone deacetylase (NuRD/HDAC), heterochromatin protein 1 (HP1), SET domain bifurcated 1 (SETDB1), and DNA methyltransferase (DNMTs) to induce repression (based on Ecco et al. 2017). **b**, Each zinc finger of KZNFs is composed of a ββα-structure. Amino acids located at the -1, 3 and 6 position of the α-helix bind specific DNA nucleotides, and thereby produce a DNA binding motif (based on Heil and Noor 2012). **c**, Depiction of the evolutionary arms between KZNFs and TEs. KZNFs repress TEs, until binding is deteriorated by mutations accumulating in TE sequences. TEs continue to invade the genome until a new, optimised KZNF is born through duplication and diversification.

race is taking place, where the birth of new TEs drives the evolution and selection of the KZNF protein family to maintain genome integrity (**Figure 4c**, Thomas and Schneider 2011; Jacobs et al. 2014; Castro-Diaz et al. 2014). Mutations that accumulate in TEs may allow the elements to escape KZNF repression, resulting in new TE invasions until another KZNF is optimised to repress the elements. Interestingly, the transcriptional activity of TEs seems to be dynamically regulated by KZNFs, especially facilitating their activity in early embryonic stages, where they are thought to be important for pluripotency (Göke et al. 2015; Fort et al. 2014; Lu et al. 2014; Gifford et al. 2013; Macfarlan et al. 2012; Pontis et al. 2019; Santoni et al. 2012). This deeper understanding of the complex relationship between TEs and their host resulted in a domestication model. This describes a model in which host species are dependent on a fine balance between the potentially adaptive advantages and deleterious effects that TEs bring (Friedli and Trono 2015; Ecco et al. 2017). Recent elaborate studies examined the genome wide targets of KZNFs and showed that next to TEs, KZNFs also target promoter regions (Schmitges et al. 2016; Imbeault et al. 2017; Barazandeh et al. 2018; Farmiloe et al. 2020; Helleboid et al. 2019). In humans, these are often the oldest KZNFs, for which no TE-derived sequence can be identified at their binding sites in promoter regions. This adds another layer to the complex interplay of TEs and KZNFs in species-specific gene regulation. Moreover, these additional functions could withhold KZNFs from becoming redundant in a genome when their TE targets become less harmful over time.

1.8 Transposable elements and KZNFs in the brain

While the relationship between TEs and KZNFs enables dynamic activity of the elements and can thereby promote speciation, most tissues are under physiological and

environmental constraints that limit the viability of TE-induced genomic and regulatory alterations. The brain might be an organ where more TE activity is tolerated, because many independent labs have reported somatic (non-germline) TE insertions in the brain and neuronal cell types (Muotri et al. 2005; Coufal et al. 2009; Baillie et al. 2011; Evrony et al. 2012; Upton et al. 2015; Erwin et al. 2016; Macia et al. 2017). For L1Hs, a human-specific TE family capable of autonomous transposition, single-cell sequencing analysis estimated that a unique somatic insertion can be found in every ~1.7 to 25 neurons (Evrony et al. 2012). This is much higher than the reported germline retrotransposition rate (Feusier et al. 2019). L1 somatic insertions found in the brain were often located in genes associated with neurogenesis and synaptic function (Baillie et al. 2011). One could think that this might be due to L1s preferentially inserting in active genomic regions, e.g. brain-associated genes in neuronal tissues, which is a feature of for example the murine leukaemia virus and HIV (Gogol-Döring et al. 2016; Sultana et al. 2017, 2019). However, while multiple studies show integration of a substantial number of L1s in or near genes (Gilbert et al. 2002, 2005; Symer et al. 2002; Beck et al. 2010), recent evidence does not point to a strong preference for active enhancers and transcribed genes (Sultana et al. 2019). With genes containing L1 insertions being more likely to show elevated expression levels, this could suggest that the brain is indeed tolerable towards TE-induced regulatory alterations. Somatic *Alu* and SVA insertions were also observed in the adult human brain, but the relative insertional activity of *Alu* elements seems much lower compared to L1s (Baillie et al. 2011). The relative expression level of TEs over the rest of the genome is also higher in the brain compared to many other tissues (Pehrsson et al. 2019). Furthermore, reduced repression of TEs is indicated by activating epigenetic marks observed at specific TEs in developing brain tissues and the adult brain (Pontis et al. 2019).

Similarly, KZNFs are expressed throughout the human developing brain and a higher number of KZNFs is expressed in the human brain compared to other adult tissues and cell types (Imbeault et al. 2017; Farmiloe et al. 2020; Turelli et al. 2020; Playfoot et al. 2021). Further support for a role for KZNFs and their targets in normal brain functioning was provided by Playfoot and colleagues revealing neuronal gene regulation by primate-specific KZNFs through TE-embedded regulatory sequences (Playfoot et al. 2021). Correlative expression analysis by our lab on KZNFs and their gene targets also suggests a direct regulatory effect of KZNFs on gene expression during human neurogenesis (Farmiloe et al. 2020). Interestingly, comparison of KZNF expression in multiple human and chimpanzee tissues revealed that a selection of KZNFs is specifically upregulated in the human brain relative to our closest living relative, the chimpanzee (Nowick et al. 2009). Via their effect on TEs and TE-mediated gene regulation, together with their direct gene-regulatory potential, KZNFs have likely contributed to human brain evolution, through regulation of transcriptional networks in the brain.

Multiple loci containing KZNFs have also been implicated in developmental malformation of the brain (Al-Naama et al. 2020; Hassan et al. 2008; Pramparo et al. 2011; Gana et al. 2012; Chien et al. 2012; Stevens et al. 2016). Furthermore, a number of loci containing KZNFs have been named in the context of disease causality for neurological disorders (reviewed by Al-Naama et al. 2020). These include intellectual disability (Hassan et al. 2008; Ramaswamy et al. 2010; Castillo et al. 2014; Agha et al. 2014; Ahmed et al. 2015),

schizophrenia (Yue et al. 2011), autism spectrum disorders (Takase et al. 2001; Metsu et al. 2014; Butler et al. 2015), X-linked mental retardation (Kleefstra et al. 2004; Lugtenberg et al. 2006; Shoichet et al. 2003), depression (Matsunami et al. 2016), and epilepsy (Spreiz et al. 2014). Functional analyses are however necessary to unravel if KZNFs are involved in inter-individual differences in TE control and subsequent neuronal gene-regulation that could explain neurological disease susceptibility and various brain developmental phenotypes.

1.9 hESC-derived neuronal organoids

A large part of this thesis will focus on the role of TEs and KZNFs in a neuronal context, because of the prominent activity that has been observed previously in this tissue. For this, we utilise human stem cell-derived 3D neuronal organoids as a model to study the role of TEs during brain development. Cell lines and animal models such as fruit flies, zebrafish and mice have been widely used in biomedical research. And while these models have contributed to a better understanding of many human diseases, translating findings from these models to humans can be difficult due to human-specific biological processes. The human developing brain, for example, differs on a structural and physiological level from mice, and the complexity of the brain cannot be captured with 2D cell culture (Lui et al. 2011). By aggregating and differentiating stem cells and exposing them to molecular cues, *in vitro* 3D tissues can be formed that mimic human organ structures. Tissues that have the same genetic background as patients can even be generated using the induced pluripotent stem cell (iPSC) technology (Takahashi et al. 2007). With this technique fibroblast cells can be reprogrammed to a pluripotent state, thereby enabling unlimited culturing of patient-specific stem cells and stem cell-derived tissues (Kim et al. 2020). Of course, also 3D organoid cultures have their limitations. One of these is the cell type heterogeneity seen in organoid tissues, which limits reproducibility. This is for example seen in a self-directed culture method developed by Lancaster and colleagues, where 'whole-brain' organoids could contain cells from the forebrain, choroid plexus, hippocampus and retina (Lancaster et al. 2013; Lancaster and Knoblich 2014). The variability between organoids can be reduced by more directed culturing protocols, where specific brain regions such as the fore- or midbrain are generated (Eiraku et al. 2008; Doi et al. 2014).

Since many primate- and human-specific TEs are present in our genome, human brain organoids are highly beneficial for studies into the role of TEs during human brain development. While donor material can also be highly informative in biomedical studies, the genetic background of any two individuals differs at more than 4 million places (The 1000 Genomes Project Consortium 2015). TEs that are still actively transpositioning in humans are contributing to these inter-individual genetic variations. By using CRISPR-Cas9, a method used to generate specific genetic modifications in cells, together with organoid models, the effect of specific genetic features can be studied in tissues with the same genetic background.

Thesis aims and outline

In this thesis, we take several approaches to unravel the roles of TEs and KZNFs in gene regulation. By doing so, we aim to contribute to broadening the knowledge on their involvement in health and disease.

Chapter one provides insights into the gene-regulatory potential of TEs in different developing and adult brain regions. We focus on H3K27ac, the mark indicative of active enhancer regions, and use chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) to study this on a genome-wide level. Additionally, we assess if the enhancer activity of TEs is influenced by ageing and age-related neurodegenerative diseases. For this, donor brain material of individuals diagnosed with Alzheimer's and Parkinson's disease, or presumed healthy deceased individuals is used.

In chapter two we dive deeper into one specific class of TEs: the MER52 elements. These elements showed enhancer potential in the adult brain, indicating they may be incorporated into neuronal gene-regulatory networks. We further elaborate on their regulatory potential by assessing the epigenetic landscape at MER52 elements. For this, we use ChIP-seq on H3K9me3, H3K27ac, H3K4me3 and P300 indicative for either repressed or active chromatin. We also assess its gene-regulatory potential *in vivo*, and analyse the relationship with its repressor protein ZNF519. We finish this chapter with assessing the TE-independent regulatory potential of ZNF519 on genes. For this we study differential gene expression after over-expression of *ZNF519* in HEK293 cells, combined with ChIP-seq data informative of the genome-wide binding of ZNF519. Additionally we develop *ZNF519* knockout (KO) hESC cell lines using CRISPR-Cas9, and generate cortical organoids to study the effect of ZNF519 on gene-expression in a neuronal context. With this, chapter two contributes to a better understanding of KZNF-TE relationships and provides insights in the co-option of KZNFs in gene-regulatory networks.

In chapter three, we study another recently evolved KZNF, ZNF91, and assess its regulation of SVA elements. For this, we combine ChIP-seq on ZNF91 with transcriptional and epigenetic profiling of *ZNF91* KO hESCs. We also assess the potential of SVAs that are under control of ZNF91 to influence nearby gene expression. Additionally, we assess if ZNF611, another strong binder of SVA elements, is crucial for repression of the elements. Lastly we find evidence for a potential model in which activated SVAs cause an upregulation of KZNF genes, which could indicate a defensive response of the host genome to TE activation. Thereby, chapter three provides more insights into the complex relationships of KZNFs and their TE targets, and the impact of the evolutionary arms race in which they are involved on the evolution of gene-regulatory networks.

Chapter four focuses on the potential involvement of structural variation within TEs in disease risk; an important source for genetic variation that remains hidden in current genome wide association studies (GWAS). By using published long-read sequencing data and PCR analysis, we show that SVA elements are a major source of inter-individual genetic variation. We find a high number of structurally variable SVAs (SV-SVAs) in disorder-associated loci, and assess whether SV-SVAs may be potentially involved in disease susceptibility indicated by single nucleotide polymorphisms in GWAS. Finally, we genetically delete three SVAs in AD and PD disease-associated loci to assess multiple aspects of their gene-regulatory influence in a human neuronal context. Together, this study reveals a novel layer of genetic variation in transposable elements that may contribute to the identification of structural variants that are actual drivers of disease associations of GWAS loci and ease the discovery of novel therapeutic targets and

strategies for complex diseases.

In the Discussion, the important findings of this thesis are highlighted and put in perspective. Additionally, future perspectives will be discussed.