The battle inside our genome

Controlling transposable elements and the evolution of human gene regulatory networks

Haring, N.L.

Publication date
2020

Document Version
Other version

License
Other

Citation for published version (APA):
Introduction

Controlling Transposable Elements and the Evolution of Human Gene Regulatory Networks
Introduction
Transposable elements (TEs) are parasitic and mobile DNA sequences that contribute to genome expansion through copy/cut-and-paste activity. The hypothesis that these elements play a fundamental role in evolution reaches back to the pioneering, and for long unrecognized, work of Barbara McClintock on TE derived controlling elements in maize (McClintock 1950). She attributed changes in the color of maize kernels to the (in)activation of genes as a consequence of mobilized TEs. A few decades later this was put in a broader evolutionary perspective by Britten and Davidson. They acknowledged that major events in evolution require significant changes in gene regulation and proposed a model in which TEs provide novel regulatory sequences on a large scale (Britten and Davidson 1971). Today, the evidence in support of this theory is mounting. Although the amplification of TE sequences is purely ‘selfish’ replication with potentially detrimental consequences for the host genome stability, it is now becoming increasingly clear that the host genome also co-opted them for its own benefits. The past decade genome wide analyses of transcription factor (TF) binding sites and epigenetic landscape at TEs in combination with transcriptome profiling in various species has contributed to our understanding of the extent to which TEs contribute to gene regulation in the animal kingdom. The identification of highly dynamic genomic defense strategies, targeted at specific TE families, suggest that they can be (de)repressed in a controlled fashion. Here we discuss findings that enhanced our understanding of the integration of TEs in gene regulatory networks and point out open questions.

The classification of Transposable elements
Due to their origin early in evolution there is a high level of variety of TEs in eukaryotic species. They can be subdivided into DNA transposons that mobilize through a cut-and-paste mechanism and retrotransposons that use an RNA intermediate and thus amplify through a copy-and-paste strategy (Greenblatt and Brink 1963; Boeke et al. 1985) (Fig 1). Retrotransposons are further classified based on the mechanism of chromosomal integration; LTR retrotransposons and non-LTR retrotransposons. Each of these subclasses are subdivided into superfamilies based on a common genetic organization and a monophyletic origin (e.g. LINE elements). The majority of superfamilies are present in all eukaryotes (Malik and Eickbush 2001; Ohshima and Okada 2005; Feschotte and Pritham 2007). The most detailed level of classification, before referring to single TE insertions, are the families and subfamilies (e.g. superfamily: LINE, family: LINE1 and subfamily: L1PA), which are closely related elements that are descendants of a single ancestral insertion. At this level the
Introduction

Figure 1. Hierarchical classification of transposable elements.
Transposable elements can be subdivided into two classes: retrotransposons and DNA transposons based on their transposition intermediate. A further subclassification is based on their integration intermediate or mechanism. Further subclassification in superfamilies/families/subfamilies is done based on phylogeny. All superfamilies are indicated, but examples of families and subfamilies are given.

highest diversity between species is expected, although the lack of thorough annotation of TE families in many genomes hampers such comparative analyses (Hoen et al. 2015; Platt et al. 2016). Noteworthy is that most TE families are inactive, meaning that they are unable to mobilize, and thus their evolutionary path ended. Although they may still be a burden for the host at many levels, they will not generate bursts of novel insertions accompanied with potential detrimental consequences for genome stability. The few TE families that are still capable of retrotransposition in the human genome, LINE1, SVA and Alu elements, have been the focus of many studies in the past years. LINE1 elements, for example,
are autonomous elements that contain all domains required for retrotransposition in their sequence (Box 1). Their retrotransposition activity has been associated with genetic mosaicism, DNA damage, genome instability and insertional mutagenesis in neurons, cancer, and aging (Muotri et al. 2005; Baillie et al. 2011; Coufal et al. 2011; Miglio et al. 2018; De Cecco et al. 2019; Simon et al. 2019). The past decade it has become increasingly clear that epigenetically active TEs, regardless of their capability of retrotransposition, can also function as strong gene regulators (Wang et al. 2007; Kunarso et al. 2010; Savage et al. 2013, 2014; Jacobs et al. 2014; Sundaram et al. 2014; Notwell et al. 2015; Chuong et al. 2016; Ito et al. 2017; Sundaram et al. 2017; Trizzino et al. 2017, 2018; Pontis et al. 2019). Therefore, in this thesis TE activity refers to an active epigenetic signature that may, but does not have to, include transcription and retrotransposition.

**Mechanisms of TE mediated gene regulation**

Genome-wide profiling of histone marks, open chromatin, TF binding sites and transcriptomics has provided compelling evidence that TEs are frequently activated and function as enhancers in various cell types (Fig 3A). This shows that TEs are not simply junk DNA that is systematically and robustly silenced. In fact they have the ability to regulate gene expression both in cis and in trans. The availability of chromatin immunoprecipitation followed by sequencing (ChIP seq) data sets for an extensive list of histone marks provided by, among others, the ENCODE project has provided insights into the extent to which TEs display characteristics of cis-regulatory elements (Jacques et al. 2013; Notwell et al. 2015; Trizzino et al. 2017, 2018; Pontis et al. 2019). This is further supported by an extensive list of TFs that bind to TEs, which frequently co-localizes with the presence of enhancer associated histone marks (Wang et al. 2007; Kunarso et al. 2010; Xie et al. 2010; Lynch et al. 2011; Chuong et al. 2013; Sundaram et al. 2014; Notwell et al. 2015; Chuong et al. 2016; Ito et al. 2017; Sundaram et al. 2017; Pontis et al. 2019). In human cells mostly evolutionarily young (i.e. primate or human-specific) TEs belonging to the LINE, LTR/HERV and SVA family are co-opted as cis-regulatory units (Wang et al. 2007; Kunarso et al. 2010; Brattås et al. 2017; Chuong et al. 2016; Ito et al. 2017; Trizzino et al. 2017; Pontis et al. 2019) (Box 1). Wang et al (2007) were the first to show that >30% of p53 binding sites localize to LTR elements and confirmed the regulatory potential of several elements in reporter assays. Recently, comprehensive analyses including ChIP seq data of 97 TFs revealed that many LTR and HERV subfamilies are bound by groups of cell type specific TFs and that the youngest elements were mostly bound by pluripotency TFs (Ito et al. 2017). This is in line with earlier studies showing the integration of several HERV families in the regulatory network orchestrating early embryonic development (Kunarso
Box 1. LINE1, HERV/LTR and SVA elements
Mainly evolutionary young members of the LINE1, HERV/LTR and SVA families are found to adopt properties of enhancers in various cell types. Below we describe the most essential information about each of these families.

**LINE1**
Full length LINE1 elements are approximately 6000bp long and contain the following domains: an internal promoter, two open reading frames (ORFs), and a poly-A tail. ORF1 and ORF2 encode two proteins that associate with the RNA they are encoded by (Fig 2A) (Hohjoh and Singer 1996; Kulpa and Moran 2006). ORF2 protein regulates the retrotransposition of LINE1 elements, as it has reverse transcriptase and endonuclease activity (Feng et al. 1996; Cost et al. 2002). With 500,000 insertions, LINE1 elements account for approximately 17% of the human genome. LINE1 elements are subdivided into families L1PA17-1 (with 1 being the youngest) that evolved linearly from each other, meaning that the youngest family replaced the older one in terms of activity (Khan et al. 2006). Indeed, most L1PA elements are inactive mostly due to 5' truncations. In total between 3000-4000 elements are full length, of which 60-100 of L1PA1 (also known as L1Hs) elements are still able to retrotranspose and are therefore called “hot LINE” elements (Sassaman et al. 1997; Brouha et al. 2003). LINE1 elements are shown to be a source of genetic variation between and within individuals in several species (Muotri et al. 2005; Coufal et al. 2009; Beck et al. 2010; Baillie et al. 2011; Jacob-Hirsch et al. 2018).

**SVA**
SVA elements are a primate-specific TE family called after its composite domains: SINE-VNTR-Alu. A full length SVA element can be divided into five components from 5' to 3' end: A hexamer repeat, an Alu-like sequence which is a composition of two antisense Alu fragments, a variable number of tandem repeats (VNTR) generally 400-900 bp long, a short interspersed element (SINE) region derived from HERV-K10 of about 500bp long and a poly-A tail (Fig 2B) (Wang et al. 2005). SVA elements may vary in length between 700-4000bp, but a canonical SVA element is around 2000bp long. There are roughly 2700 fixed insertions in the human genome, half of which inserted in our genome after the split with the last common ancestor with chimpanzees and are therefore unique to humans. Their activity results in a new germ line insertion approximately once in every 900 births (Xing et al. 2009). A subclassification based on the evolutionary age of the SINE region has divided the SVA family into six subfamilies (A-F, with A being the oldest and F the youngest subclass) (Wang et al. 2005). They are non-autonomous TEs that hijack the LINE1 retrotransposition machinery for their mobilization (Raiz et al. 2012).
HERV

Human endogenous retroviruses (HERVs) and solitary LTR elements represent ~8% of the human genome. HERVs share several domains with retroviruses such as 5’LTR, GAG, POL, ENV and 3’LTR (Fig 2C). The ~30,000 HERV elements have been classified into into Class I (HERV-W and HERV-H), Class II (HERV-K) and Class III (HERV-L and HERV-S) based on phylogenetic analyses of the POL gene in comparison to the classification of animal retroviruses (De Parseval and Heidmann 2005). In humans HERV/LTR elements are not capable of retrotransposition, but they are frequently identified as strong gene regulators in cis and trans (Macfarlan et al. 2012; Lu et al. 2014; Wang et al. 2014; Göke et al. 2015; Grow et al. 2015; Brattås et al. 2017; Li et al. 2019; Min et al. 2019) (Macfarlan et al. 2012; Lu et al. 2014; Wang et al. 2014; Göke et al. 2015; Grow et al. 2015; Brattås et al. 2017; Li et al. 2019; Min et al. 2019).

Figure 2. Domains that compose LINE, SVA and HERV elements.
Poli = polymerase II promoter, UTR = untranslated region, ORF = open reading frame, Poly-A = Poly-A tail.

et al. 2010; Pontis et al. 2019). In human embryonic stem cells (hESCs) between 15% and 21% of binding sites of pluripotency factors OCT4 and NANOG reside in primate-specific TEs in vicinity of highly conserved target genes, suggesting species specific regulation of stem cell related gene expression (Kunarso et al. 2010). A recent study suggested that binding of KLF4 to these elements is essential to recruit OCT4 (Pontis et al. 2019). Note that the majority of these studies are descriptive and correlational, which complicates the identification of target genes controlled by specific TE insertions or TE subfamilies. The availability of easy-to-use (epi)genetic tools such as CRISPR/Cas9 will lead to the generation of more causal evidence for specific TEs and their target genes in
the coming years (Box 2). Recently, CRISPR/Cas9 mediated deletion of MER41 elements in immune cells revealed that they are STAT1 targets essential for the regulation of specific immunity genes (Chuong et al. 2016). Alternatively, it has been suggested that the integration of novel TE insertions can secondarily regulate gene expression through the induction of local heterochromatin, as most TEs are continuously repressed (Fig 3B). However, evidence showing support for this phenomenon and the scale at which it affects gene expression genome wide is relatively scarce. One example is a comparison of different mouse ESC lines that contain polymorphic TE insertions to show that some TEs induce H3K9me3 and H4K20me3 in the flanking sequence, leading to alterations in gene expression (Rebollo et al. 2011).

Besides functioning as cis-regulatory elements, TEs are shown to regulate gene expression as alternative promoters or post-transcriptionally, mostly in early embryonic development but also in other tissues (Fig 3C+D). RNA seq of mouse oocytes and 2-cell (2C)-stage embryos revealed that many 2C em-

**Figure 3. TEs can regulate gene expression in *cis* and *trans.*

A) TEs can provide TF binding sites and show an epigenetic signature of enhancers that allow it to regulate the expression of genes in close vicinity. B) The insertion of TEs is usually followed by KZNF-mediated induction of local heterochromatin, which may spread into the flanking genomic regions and affect gene expression. C) HERV-derived transcripts are shown to function as scaffolds for TFs and facilitate the cis-regulatory function of TEs in *trans.* D) Transcriptionally active TEs provide alternative promoters that lead to the generation of chimeric transcripts consisting of the TE and a downstream gene.
bryo-specific genes were driven by muERV-L elements, generating chimeric TE-gene fusion transcripts (Fig 3D) (Macfarlan et al. 2012). In human preimplantation embryos dynamic transcription of distinct HERVs marks specific populations of early embryonic cells (Göke et al. 2015; Grow et al. 2015). OCT4 and NANOG binding at LTR elements, together with local DNA hypomethylation facilitate HERV transcription in these cells (Göke et al. 2015, Grow et al. 2015). Strikingly, expression of these HERV transcripts are shown to be essential for maintaining pluripotency and self renewal of embryonic stem cells, suggesting a post-transcriptional regulatory role of these non-coding RNAs (Wang et al. 2014; Lu et al. 2014). In fact, these HERV derived long noncoding RNAs (lncRNAs) function as scaffold for TFs and chromatin remodelers and their transcriptional activity is required to maintain chromatin architecture during early embryonic development, which in its turn ensures proper gene regulation (Fig 3C) (Lu et al. 2014; Zhang et al. 2019). At the same time these scaffolds are facilitating OCT4 binding at LTR7 elements and thus control the strength of the cis-regulatory properties of these elements (Lu et al. 2014). To summarize, TEs add an extra layer of complexity to the gene regulatory program through various mechanisms of regulation.

TE driven evolutionary innovation of gene regulatory networks
Active TEs are fuel for evolutionary innovation of gene regulatory networks. Through waves of TE invasions they bring along TF binding sites which can establish co-regulation of gene groups that were previously unrelated (Fig 4). In support of this is the observation that nearly all primate-specific cis-regulatory elements are TE derived (Jacques et al. 2013). A comparison of enhancer associated histone marks (H3K4me1 and H3K27ac) in liver cells from six primate species (including humans) revealed that SVA and LTR12 elements most frequently overlap ape- and human-specific enhancers, affecting gene expression in the close vicinity (Trizzino et al. 2017). The primate specific expansion wave of MER41 elements provided many novel STAT1 binding sites and rewired gene networks essential for orchestrating an immune response (Chuong et al. 2016). A comprehensive analysis of the binding sites of 26 TFs, in human and mouse showed that the vast majority of TE-derived TF peaks was species specific (98% and 99% respectively) (Sundaram et al. 2014). This may be explained by novel TE expansion waves after the human-mouse split resulting in new TE families that provided TF binding sites.

Importantly, for TE propagation to have any meaning on an evolutionary scale, insertions need to be passed on to the next generation and therefore need to take place in germ cells or the early embryo. Interestingly, many of the TE silencing mechanisms are released during early embryonic development in the
interest of the proper activation of host gene regulatory programs. An example is the genome wide loss of DNA methylation to reset imprinted genes (Miyoshi et al. 2016). This provides a window of opportunity for TEs to become active, including retrotransposition and exposure of their gene regulatory potential (Kunarso et al. 2010; Pontis et al. 2019). In both humans and mice species-specific TEs have rewired the core gene expression program in embryonic stem cells by providing binding sites for pluripotency factors (Kunarso et al. 2010; Pontis et al. 2019). HERV-H/LTR7 and HERV-K/LTR5-Hs show an epigenetic signature associated with enhancers (e.g. H3K27ac and H3K4me1) and become transcriptionally active in a narrow developmental time window of the preimplantation embryo (Wang et al. 2014; Grow et al. 2015; Göke et al. 2015; Pontis et al. 2019). They are regulatory units that provide binding sites for pluripotency TFs on a large scale and demarcate boundaries of topologically associated domains (Kunarso et al. 2010; Wang et al. 2014; Zhang et al. 2019). On top of that, HERV-H transcripts are essential for maintaining pluripotency of human stem cells (Wang et al. 2014; Lu et al. 2014). There are even examples of the same gene(s) being regulated by different lineage-specific TEs in different species (Romanish et al. 2007; Emera and Wagner 2012; Chuong et al. 2016). An example is the NLR family apoptosis-inhibitory protein (NAIP) gene (member of the IAP gene family) that is driven by species-specific insertions of LTR elements in it’s promoter (Romanish et al. 2007). The enrichment of LTRs in promoters
of other IAP genes in several species showed that certain loci may be more sensitive to TE insertions. An alternative explanation is that the ancestral locus already contained a regulating TE that has been substituted by a new TE that took over the regulatory properties of the older one, leading to decay of that TE. This turnover model suggests that over time this would result in renewal of co-opted TEs in these loci. The parallel integration of TEs in gene regulatory networks during early development in both mice and humans suggest that this strategy is one of the main drivers of evolution of gene regulation (Kunarso et al. 2010; Grow et al. 2015; Pontis et al. 2019). With regard to recent evolution it is unmistakable that TE expansion waves have provided novel TF binding sites on a large scale. Yet, TF binding sites derived from ancient TE families are more difficult to identify as they have accumulated many (species-specific) mutations leading to degradation of the TE sequence around the TF binding site. Indeed, comparisons of non-exonic parts of the genomes of 29 mammalian species showed that thousands of conserved sequences are fragments from TEs (Lowe and Haussler 2012). Likely the current numbers are an underestimation of the true amount of TE-derived TF binding sites in the genome of a given species. Taken together, several lines of evidence showed that TE expansion waves provide an effective way to rewire gene regulatory networks and fuel evolution.

An evolutionary arms race: KRAB zinc finger proteins as genomic defense against TE invasions

So far we have mainly focused on the beneficial effects of TE activity for the host. However, we need to stress that their activity is probably even more likely to be harmful, causing genome instability, DNA damage and mutagenesis upon insertion. Note that for TEs to persist it is essential that they are tolerated enough to amplify, without compromising the fitness of the host. From the perspective of the host, except the few TEs that are occasionally bound by TFs and co-opted for gene regulatory purposes, TE activity needs to be repressed to maintain genome stability and normal gene expression. Although there are some indications for the existence of TE encoded mechanisms that control propagation (Lohe and Hartl 1996), their activity is mostly repressed by host factors that evolved in parallel to repress the enormous variety of TE sequences. The KRAB zinc finger (KZNF) gene cluster is a family of TFs that is specialized in binding and repressing specific TE families. KZNF proteins consist of a tandem array of zinc fingers that determine DNA binding specificity (Frankel et al. 1987; Thukral et al. 1991; Cook et al. 1994) and a KRAB domain that is bound by cofactor KAP1 (Nielsen et al. 1999; Sripathy et al. 2006) that recruits epigenetic modifiers to ensure repression through histone modifications (Schultz et al. 2001, 2002b) and DNA methylation (Turelli et al. 2014) (Fig 5A, Box 3). Almost half of the 400 KZNF genes in the human genome are primate specific, indicating they played a role in recent evolution (Thomas and Schneider 2011).
Box 2. Experimental approaches to study TE mediated gene regulation

The optimization of several novel experimental techniques has enhanced the extent at which we can assess gene regulatory properties of specific sequences and identify their target genes. This way several TE families and individual insertions are proven to control the expression of specific genes. Below we discuss the major techniques that revolutionized the field.

Luciferase assays are an interesting approach to elucidate the regulatory potential of (specific parts of a) TE sequence outside it’s cellular and/or genetic context (Savage et al. 2013; Jacobs et al. 2014; Savage et al. 2014; Turelli et al. 2014; Sundaram et al. 2017; Trizzino et al. 2017). Jacobs et al (2014) used an SVA-D luciferase construct to show that in a mouse cellular environment, that lack primate-specific KZNF genes optimized to repress SVA elements, they harbor a strong gene regulatory potential that exceeded the OCT4 enhancer. Luciferase assays with fragments of another SVA-D showed that different parts of the SVA have a different regulatory potential in a cell type and orientation dependent manner (Savage et al. 2013). Alternatively, Sundaram et al (2017) used this technique to show that (computationally reconstructed) ancestral versions of cis-regulatory TEs have a stronger regulatory potential than their descendants. Although these studies provided insights into the gene regulatory potential of various TE families, it remained unclear to what extent this potential is unleashed in a cellular context and what genes are targeted.

CRISPR/Cas9 mediated deletion of specific TE insertions followed by the profiling of gene expression changes allows to directly address this (Chuong et al. 2016; Sundaram et al. 2017; Fuentes et al. 2018; Zhang et al. 2019). In human immune cells, various MER41 elements are bound by immune-related TFs (Chuong et al. 2016). Mutating these binding sites with CRISPR/Cas9 resulted in a decrease of IFNG-induced genes closeby these elements. In hESCs, deletion of two HERV-H elements affected chromatin architecture and expression of genes at their 5' side (Zhang et al. 2019). Similarly, the deletion of an intronic RLTR9E element resulted in downregulation of the Akap12 gene in mESCs (Sundaram et al. 2017).

Alternative strategies are aimed at unleashing the gene regulatory potential of TEs in a targeted or genome wide fashion. The ablation of repressive co-factor KAP1 has been used to revive cis-regulatory properties of TE families on a global scale (Rowe et al. 2013; Fasching et al. 2015; Turelli et al. 2014; Brattås et al. 2017; Coluccio et al. 2018). In human cells this resulted in transcriptional and
epigenetic activation of TEs belonging to HERV, LINE and SVA families and activation of neighboring genes (Turelli et al. 2014; Brattas et al. 2016). Note that this widespread activation of TE families in combination with the ability of KAP1 to regulate genes directly (O’Geen et al. 2007), makes it difficult to attribute gene expression changes directly to specific TE families or insertions.

Recent adaptations of the CRISPR/Cas9 system allow for the enforcement of targeted chromatin changes using gRNAs and a nuclease dead Cas9 (dCas9) fused to a transcriptional repressor (CRISPRi) or activator (CRISPRa). By designing gRNAs that target the majority of TEs within a family, several studies have manipulated the chromatin status of LTR5-Hs, SVA, and HERV-H elements and showed that this affected gene expression, even over long genomic distances (Fuentes et al. 2018; Pontis et al. 2019; Zhang et al. 2019). This way causality between enhancer LTR5-Hs elements and their target genes was demonstrated (Fuentes et al. 2018). Note that the design of gRNAs specifically targeting a specific TE (sub)family for CRISPRi/a can be challenging, as many TE families show a high degree of sequence similarity.

Profiling of chromatin architecture is yet another strategy to study the impact of TEs on genome regulation, including their role as cis-regulatory elements. Several variants of chromosome conformation capture (3C) techniques have been employed to identify connections between TEs and their target sequence(s) (Raviram et al. 2018; Pontis et al. 2019; Zhang et al. 2019). These studies confirmed that earlier described MER41 elements connect to the studied IF-NNG-responsive gene promoters (Raviram et al. 2018) and showed that LTR5-Hs and SVA elements physically interact with promoter of genes that were downregulated after CRISPRi mediated repression of these elements (Pontis et al. 2019). Collectively these experimental approaches provide ways to elucidate gene groups that are wired into gene regulatory networks through the co-optition of TEs as cis-regulatory units.

Several lines of evidence showed that KZNF genes evolved in response to TE expansion waves, suggesting the existence of an evolutionary arms race (Fig 5B) (Thomas and Schneider 2011; Jacobs et al. 2014; Imbeault et al. 2017). A study presenting a strong correlation between the number of ZNF domains and LTR elements including genomes of many species was the first to provide large scale evidence for this phenomenon (Thomas and Schneider 2011). A large scale analysis of KZNF binding sites revealed that KZNF genes are approximately the same age as their target TE (Najafabadi et al. 2015). Identification of individual examples of TE-KZNF co-evolution has provided concrete
Introduction

evidence of an evolutionary battle. An illustrative example is the loss of the 129bp ZNF93 binding site in the 5' UTR of L1PA3 elements that allowed them to escape repression resulting in a new invasion of L1PA elements (Jacobs et al. 2014; Imbeault et al. 2017). Another KZNF, ZNF91, underwent major structural changes that enabled it to bind and repress SVA elements that arose around the same time (Jacobs et al 2014). Despite or owing to the adaptability of KZNF genes to repress TEs, their propagation continues and generates genetic diversity in the human population. Yet, the fact that many dead TE families are still bound by KZNF proteins (Schmitges et al. 2016; Imbeault et al. 2017) suggests that their repression is essential to protect from other, non-retrotransposition, TE-mediated effects. This is supported by the fact that TEs that are bound by KZNF proteins and adorned with H3K9me3 are also most frequently marked with histone marks associated with enhancers (Imbeault et al. 2017).

It is becoming increasingly clear that the function of KZNF proteins is not limited to the control of TEs (Fig 5B). Several other roles have been attributed to specific KZNF genes, such as autoregulation by binding to itself and/or other members of the KZNF gene family (Frietze et al. 2010; Yan et al. 2017), gene regulation (Oleksiewicz et al. 2017; Yang et al. 2017a; Chen et al. 2019, Farmiloe et al. In press), and maintenance of DNA methylation at imprinting control regions (Li et al. 2008; Riso et al. 2016; Takahashi et al. 2016, 2019). Several KZNF proteins have become indispensable for regulation of gene expression in mice and humans. They are shown to function as transcriptional repressors through KAP1 mediated H3K9me3 deposition on promoters of pro-differentiation (Oleksiewicz et al. 2017) and placenta-specific (Yang et al. 2017) genes in early embryonic development. A few others recruit epigenetic modifiers to imprinting control regions to protect it from genome wide erasure of DNA methylation during early embryonic development (Li et al. 2008; Riso et al. 2016; Takahashi et al. 2016, 2019). Deletion of Zfp57 and Zfp445 in mice caused embryonic lethality, showing that they have become indispensable for normal development (Takahashi et al. 2019). Taken together, the KZNF gene cluster shows the adaptability of the host to control TE invasions and parallel introduction of novel regulatory factors.

**Regulation of the battle between TEs and KZNF proteins to control TE-derived cis-regulatory elements**

At this point it is clear that the host can occasionally take advantage of the battle between TEs and KZNF proteins by co-opting their regulatory properties to innovate gene regulatory networks. It seems intuitive that through the regulation of the expression of KZNF genes, the exposure of the cis-regulatory potential of TEs can be controlled. However, very little is known about the
Figure 5. The evolutionary arms race between KZNF genes and TEs.
A) KZNF proteins bind to specific TEs with a tandem array of zinc fingers. The KRAB domain recruits co-factor KAP1, which attracts chromatin remodelers such as histone methyl transferases (HMT) and DNA methyl transferases (DNMT) that ensure H3K9me3 and DNA methylation are deposited. B) Schematic representation of the co-evolution of TEs and KZNF genes. Over the course of evolution an enormous diversity of TE families (blue circle) and KZNF genes (pink circle) accumulated in the human genome. Although the majority of TE families lost their capacity to mobilize (skull sign), they are still actively repressed by their KZNF mate. For the few active TE families KZNF-mediated TE repression facilitates their evolution, as mutations (red stripes) allow the TE to escape repression. This on its turn drives structural innovation of the KZNF cluster to optimize a KZNF gene to repress the escaped TE. There is evidence showing that the KZNF gene cluster has internal regulatory mechanisms, exemplified by the binding of specific KZNF proteins to their own genomic sequence or to the promoters and ‘3 side of other KZNF genes. Finally, KZNF proteins have been shown to regulate various cellular processes by binding gene promoters.
Box 3. The structural evolution of a KRAB zinc finger genes

KZNF proteins are composed of KRAB domain and a tandem array of zinc fingers. A zinc finger is generally 28-30 amino acids long and contains two cysteines and two histidines that coordinate binding to a zinc ion to stabilize the finger and they adapt a structural conformation with two beta-strands and one alpha-helix (Frankel et al. 1987; Thukral et al. 1991; Cook et al. 1994). Amino acid residues at positions -1, +2, +3 and +6 of the alpha helix determine the three consecutive nucleotides plus one nucleotide on the other strand recognized by a finger, and are therefore called the fingerprint amino acids (Pavletich et al. 2019). KZNF proteins contain 3-40 zinc fingers, with an average of 12 zinc fingers in humans (Yang et al. 2017b).

Upon examination of the three nucleotides bound by specific zinc fingers, prediction tools were developed to predict DNA sequences bound by tandem arrays from specific ZNF genes (Molparia et al. 2010; Persikov and Singh 2014). However, the experimental identification of the target sequences of many KZNF genes showed frequent discrepancies with the predicted sequence (Jacobs et al. 2014; Najafabadi et al. 2015; Schmitges et al. 2016; Imbeault et al. 2017). Systematic deletion of ZNF domains showed that only a subset of zinc fingers are required for target recognition by ZNF91 and ZNF93 (Jacobs et al. 2014). On top of that, it has been shown that the recognized DNA code is not simply a sum of the 3 nucleotides recognized by each finger. A recent structural study showed that zinc fingers in the middle of the ZNF568 zinc finger array bind to two or four instead of three nucleotides (Patel et al. 2018).

The KRAB domain is around 75 amino acids long and functions as strong transcriptional repressor. Although most KZNF genes encode the functionally dominant A-box, some have additional KRAB motifs (B,BI,b and C) (Bellefroid et al. 1991; Mark et al. 1999; Huntley et al. 2006). The KRAB domain itself does not have any enzymatic activity, but it recruits co-repressor KAP1 (TRIM28) to the sequences bound by the zinc fingers (Friedman et al. 1996; Kim et al. 1996; Moosmann et al. 1996). KAP1 recruitment is universal and essential for TE repression, as KAP1 knockdown/knockout results in collective derepression of various TE subfamilies in both murine and human cells, characterized by loss of H3K9me3 and DNA methylation, transcriptional activation and gain of H4 acetylation (Rowe et al. 2010; Turelli et al. 2014; Fasching et al. 2015; Brattås et al. 2017; Coluccio et al. 2018). It is well established that KAP1 mediates TE repression by functioning as a scaffold for epigenetic modifiers such as H3K9 histone methyltransferase SET domain bifurcated histone lysine methyltransferase 1 (SETDB1), heterochromatin protein 1 (HP1), histone deacetylases, and DNA methyltransferases (Nielsen et al. 1999; Ryan et al. 1999; Schultz et al.
regulation of KZNF genes and the consequences of KZNF expression dynamics for TE repression. Although studies have shown that the expression of KZNF genes is highly cell type and developmental stage specific (Brattås et al. 2017; Oleksiewicz et al. 2017; Imbeault et al. 2017; Kauzlaric et al. 2017; Farmiloe et al. In press), direct connection to the activity of their target TE is scarce. A recent study proposed a feedback loop between TE-derived enhancers and KZNF genes in early embryonic development. The conversion of primed into naive hESCs resulted in a high number of H3K27ac positive young TEs along with higher expression levels of KZNF genes in the neighborhood, ultimately leading to their own silencing (Pontis et al. 2019). They showed that KZNF expression levels correlate with the amount of H3K9me3 at their TE mate. Other studies showed that expression of KZNF genes correlate with the expression of genes
bound by the KZNF at their promoters (Oleksiewicz et al. 2017; Farmiloe et al. in press).

The publication of binding sites of hundreds of KZNF proteins opens up new possibilities to explore the relation between KZNF gene expression and TE activity further (Najafabadi et al. 2015; Schmitges et al. 2016; Imbeault et al. 2017). Noteworthy is that KZNF expression is generally higher in stem cells, immune cells and the brain (Imbeault et al. 2017). Strikingly, the same cell types are known for their high number of TE-derived cis-regulatory elements (Kunarso et al. 2010; Chuong et al. 2016; Trizzino et al. 2018; Pontis et al. 2019). Together this suggests that some tissues are more sensitive to TE-mediated cis-regulation. Indeed, the past decade several studies pointed out that neurons are particularly sensitive to TE activity, sometimes leading to novel TE insertions generating genetic mosaicism in the brain (Muotri et al. 2005, Baillie et al. 2011). As our brain is one of the most evolutionary divergent organs, providing us with higher cognitive functions that distinguish us from other mammals, it is appealing to hypothesize that human-specific TE families contributed cis-regulatory elements that reshaped neuronal gene regulatory networks. H3K27ac profiling of fetal and adult brain tissue showed that SVA elements become activated in a region specific manner and are enriched among H3K27ac positive TEs (Pontis et al. 2019). Yet, our understanding of the extent of TE-mediated gene regulation, the TE families involved and their target genes in the human brain is still minimal. One of the reasons is the limited availability of post mortem human brain tissue, which hampers extensive profiling of tissue-specific TE-derived enhancers through analysis of key TFs. Being at the dawn of in vitro organoid differentiation, this provides us with novel strategies to study the role of TEs in differentiated neuronal tissues.

The role of TEs in neurodegenerative diseases

TE activity - transcriptional, epigenetic or retrotransposition - is generally not beneficial for the host and therefore limited in adult somatic cells. However, TE repression deteriorates with aging, leading to deleterious TE activation in various species (Li et al. 2013; Maxwell et al. 2011; Wood et al. 2016; De Cecco et al. 2019; Min et al. 2019). The brain seems particularly sensitive for this process. Noteworthy is that aberrant TE activity is increasingly associated with pathologies underlying neurodegenerative diseases. Transcriptional profiling of post mortem human brains showed transcriptional activity of various TE families in multiple sclerosis (MS) (Antony et al. 2004; Morandi et al. 2017), amyotrophic lateral sclerosis (ALS) (Douville et al. 2011; Li et al. 2012, 2015; Prudencio et al. 2017; Tam et al. 2019) and Alzheimer’s disease (AD) (Guo et al. 2018). Although it remains unclear whether they are cause or consequence of disease pathology, some observations seem to suggest that TE activation may
Controlling Transposable Elements and the Evolution of Human Gene Regulatory Networks

directly cause neuronal loss or dysfunction (Tan et al. 2012; Li et al. 2015; Krug et al. 2017). For example, overexpression of ALS-associated HERV-K env in the mouse brain resulted in several pathological characteristics also observed in ALS such as progressive motor dysfunction, decreased synaptic activity, dendritic spine abnormalities, and DNA damage (Li et al. 2015). The transcriptional upregulation of many young TE families in a subgroup of ALS patients is directly linked to aggregation of TPD-43 protein (Douville et al. 2012; Li et al. 2015; Krug et al. 2017; Tam et al. 2019), as ablation of TPD-43 protein resulted in TE transcription in mouse and human cells (Li et al. 2012; Tam et al. 2019). In Drosophila human TPD-43 promotes neurotoxicity through the activation of a specific TE family and inhibition of this TE is sufficient to rescue the phenotype (Krug et al. 2017). Guo et al (2018) used expression and H3K9ac data from AD brains in combination with a Tau overexpression fly model to show that TE activity is a direct consequence of Tau pathology. Earlier Tau pathology was already associated with heterochromatin loss and increased DNA damage in neurons (Khurana et al. 2012; Frost et al. 2014; Klein et al. 2019). The first may be causing TE activity, whereas the second can be the consequence of it. Although it is likely that genome wide TE derepression provides a burden for neurons, the exact mechanisms through which TEs execute their neurotoxic effects remain elusive. Possibly, the DNA damage observed in neurons from AD patients is a consequence of TE activity (Madabhushi et al. 2014). Retrotansposition activity is known to cause DNA damage, but derepressed TEs may also provide hotspots for non-homologous recombination events (Robberecht et al. 2013; Vogt et al. 2014). Alternatively, aberrant TE activity is associated with the induction of an neuroinflammatory response ultimately leading to cell death (Antony et al. 2004; De Cecco et al. 2019). Noteworthy is that most of these studies are focused on transcriptional activity of TEs, but their cis-regulatory potential remains largely unaddressed in the context of neurodegenerative diseases. However, the awakening of the TE-derived enhancers could be another mechanism through which they can dysregulate gene expression with deadly consequences. In conclusion, there is strong evidence that TEs become derepressed in several neurodegenerative diseases and that their products can be neurotoxic. However, what place they take in pathological cascades remains to be established.

Conclusions and perspectives
The past decade is characterized by an enormous progression in our understanding of how TEs may have contributed to the evolution of genomes and how they are controlled, largely facilitated by major technological advances allowing the genome wide profiling of TEs and the experimental assessment of the regulatory impact of single TE insertions. Suggestive evidence showing that TEs may have significantly rewired gene
regulatory networks over the course of evolution is mounting. So far, TEs have been found to influence the expression of genes during early embryonic development, and somatic cell lineages such as immune and liver cells. Extending these analyses to more tissues and cell types, such as different regions of the brain, will reveal if the co-option of TEs as cis-regulatory elements is limited to certain cell types/processes or reflects a more general mechanism. On top of that, most analyses are focused on the cis-regulatory properties of HERV/LTR elements, whereas the extent at which other TE families, such as SVA elements, rewired gene regulatory networks is still largely unclear.

Yet, we need to remember that TE activity is a two-edged sword. TEs are facilitators of evolutionary innovation on one side, but a cause of mutations, genomic instability and DNA damage on the other side. Therefore, their activity needs to be tightly controlled. The KZNF gene cluster represents a dynamic and highly adaptable defense system in charge of controlling TE activity in a time and place dependent manner. The dynamics of KZNF mediated TE repression are still largely unknown, as is the way they are regulated. Although there are several mechanisms proposed to be underlying their regulation, including auto-regulation by KZNF proteins and an TE-mediated immune response, more studies are necessary to truly understand the cause and consequences of KZNF gene control in relation to TE activity. An increasing number of studies link aberrant TE activation to pathogenic processes related to cancer, aging, neurodegenerative diseases and autoimmunity. Yet, if they are cause or consequences to disease etiology or progression remains an open question. The quickly advancing technology will increase our ability to functionally assess the impact of the battle between KZNF genes and TE families on gene regulatory networks and evolution of novel structures and functions. It will also open up new avenues for understanding the etiology of complex multifactorial conditions such as aging, cancer and neurodegenerative diseases.

**Aim of the thesis**

In this thesis we aimed to enhance our understanding of the regulatory role of KZNF proteins and TEs in human embryonic stem cells and *in vitro* differentiated neurons. In an unbiased genome wide analysis we set out to identify KZNF genes and TE families that show cell type specific activity and are potentially involved in region specific gene regulatory networks. In parallel, we focused on understanding the implications of the evolutionary arms race between SVA elements and KZNF genes. Through identification and subsequent deletion of SVA-binding KZNF proteins we aimed to elucidate their role in controlling SVA elements and their combined effect on human gene regulation.
Outline of the thesis
We hypothesized that TEs contributed to human evolution by shaping gene regulatory networks involved in brain development. In Chapter 1 we used hESC and hESC-derived cortical and midbrain dopaminergic (mDA) organoids to identify brain region specific regulatory activity of TE subfamilies by subjecting them to RNA sequencing and H3K27ac ChIP sequencing. First, we studied the global expression dynamics of KZNF genes and all TE subfamilies. In parallel, we focused on specific KZNF genes and their target TE to understand if and how their expression dynamics (anti)correlated. We studied the enrichment of H3K27ac on specific TE subfamilies and identified the presence of cell type specific TF motifs, to elucidate if they harbor the potential to be integrated in cell type specific gene regulatory networks.

SVA elements are shown to harbor a strong gene regulatory potential and can be repressed by primate-specific ZNF91 outside a human cellular context. However, it remained elusive to what extent ZNF91 binds SVA elements and if it is required and essential for their repression in human cells. Therefore we profiled ZNF91 binding sites and abolished ZNF91 expression to identify transcriptional and epigenetic consequences for SVA elements and genes.

In Chapter 2 we describe several approaches to delete ZNF91 expression using RNA interference. We used lentiviral transduction with an shRNA against ZNF91 to establish ZNF91 knockdown hESC lines. Alternatively, we applied transient lipid-based transfection of siRNAs against ZNF91 in various cell types and subjected ZNF91 knockdown hESCs and SK-N-SH cells to RNA sequencing to profile SVA and gene expression.

We started Chapter 3 by profiling ZNF91 binding sites through ZNF91-GFP overexpression in HEK293 cells and performing ChIP seq for GFP. Subsequently, we generated ZNF91 knockout hESC lines using CRISPR/Cas9 and subjected wild type and knockout clones to RNA sequencing and ChIP sequencing for histone marks associated with enhancers (H3K27ac) and active transcription (H3K4me3).

A large ChIP sequencing screen including hundreds of KZNF proteins showed that several other KZNF proteins also bind SVA elements. We hypothesized that multiple KZNF genes may be required for full repression of SVA elements. Chapter 4 was aimed at elucidating which KZNF proteins are essential for the control of SVA elements. We re-analyzed the ChIP sequencing data of ten KZNF proteins and identified that only ZNF611 also strongly binds to all SVA elements. Therefore, we reconstructed the evolutionary history of ZNF611 and compared it to the evolutionary timeline of SVA elements to understand when it acquired the ability to bind SVA elements. In parallel we generated ZNF611
Introduction

knockout and ZNF91/ZNF611 double knockout hESC lines, as they may both be essential for optimal SVA control.

In the Discussion we summarize our main findings and place them in a broader context. Also, we speculate about their implications and present future perspectives.
References


Controlling Transposable Elements and the Evolution of Human Gene Regulatory Networks


