Transposable elements as hidden neuronal gene regulators in health and disease

van Bree, E.J.

Publication date
2022

General discussion

In this thesis we showed the involvement of transposable elements (TEs) and KRAB zinc fingers (KZNFs) in gene regulation.

Incorporation of transposable elements in neuronal gene-regulatory networks

Barbara McClintock's findings in the 1950's were the first to highlight the importance of mobile DNA elements in gene regulation (McClintock 1950, 1956). Now, decades later, extensive studies on TEs revealed the many different ways in which TEs have driven genome evolution. With numerous transcription factor binding sites residing within TE sequences, transposition events spread regulatory units across the genome (Thornburg et al. 2006; Bourque et al. 2008; Sundaram et al. 2014). Unravelling developmental time-point and cell-type-specific gene regulation mediated by TEs can provide fundamental information on species and cell-type-specific gene expression. With studies showing somatic mosaicism in neurons and reduced repression of TEs upon ageing (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; Pontis et al. 2019; Li et al. 2013), the brain is an intriguing organ to study which TEs have the potential to function as enhancer elements. On top of that, the brain has undergone profound changes in size and complexity throughout primate evolution, both features of the human brain suggested to be fundamental to the development of our cognitive and intellectual skills (Roth and Dicke 2005).

In chapter one and two, we analyse the gene-regulatory potential of TEs in different developing and adult brain regions. We show that the repertoire of TEs with an enhancer potential differs greatly between human embryonic stem cells (hESCs) and hESC-derived developing brain tissues. This is supportive for cell-type specific epigenetic regulation mediated by TEs (Pontis et al. 2019). Recently emerged TEs (evolutionarily young of age) displayed H3K27ac activating epigenetic marks in hESCs, which is in accordance with recent studies analysing the activity of TEs in early developmental stages (Pontis et al. 2019; Gao et al. 2018). Surprisingly, evolutionary young TE classes became depleted for H3K27ac upon differentiation. We observed that H3K27ac-enriched classes of TEs were often shared amongst many different adult neuronal tissues, despite the high level of cellular and functional diversity between brain regions. This suggests that these TE classes are mainly incorporated in pan-neuronal regulation of gene expression. This is an interesting concept that warrants further investigations, since much insight can still be gained on the level of pan-neuronal gene regulation.

We also find indications for inter-individual differences in H3K27ac at TEs in specific adult donor brain regions. Whereas these findings are potentially very interesting, the first question we need to ask is whether these differences are potentially reflecting technical artefacts related to the isolation and storage of donor brain tissues. It has been shown that differences in isolation method, post-mortem times and storage conditions can affect histone marks and the identification thereof (Vermunt et al. 2014). Furthermore, differences in cell composition between different isolates, might underlie variations in epigenetic signatures at TEs that we observed between individuals. In the future, a better control of cell types should be ensured using for example single-cell sequencing. Even if the inter-individual differences we observed are real and not due to technical or

experimental biases, the relatively small scale of our pilot analysis in AD/PD patients and healthy controls of varying ages, has prevented us from drawing any conclusions about whether the differences in TE-activation profiles are in any way related to ageing and disease. Despite the need for better models to study ageing, modelling ageing of cells or tissues *in vitro* has remained a great challenge in the field. Aspects of ageing cells can be induced with small compounds, such as (1) DNA replication stress inducers, (2) DNA-damaging agents, (3) epigenetic modifiers, (4) inhibitors of telomerase activity, (5) cyclin-dependent kinase inhibitors, (6) p53 activators, (7) activators of protein kinase C, and (8) ROS inducers (Petrova et al. 2016). Recently, Shaker and colleagues suggested that human brain organoids cultured for prolonged times (e.g. 10 or 13 weeks) show hallmarks of senescence (Shaker et al. 2021). However, it remains unclear how accurate these artificial methods are modelling all aspects of ageing, including possible aberrant TE activity. Alternatively, direct conversion of neurons from fibroblasts of aged individuals could also prove useful. Forced direction of fibroblasts into a neuronal phenotype prevents the loss of age-associated transcriptomic and epigenetic signatures seen when reprogramming fibroblasts into induced pluripotent cells (iPSCs) (Mertens et al. 2015; Tang et al. 2017; Victor et al. 2018). Attempts have also been made to generate organoid models of Alzheimer's disease, which might prove useful in assessing TE activity in neurodegenerative tissues (Choi et al. 2014; Gonzalez et al. 2018).

In chapter three and four we study the gene-regulatory potential of SVA elements, a great-ape specific class of TEs of which half of the elements are unique to humans (Wang et al. 2005; Quinn and Bubb 2014). We find that unrepressed SVAs can act as strong enhancers, but also find SVA variants that can reduce nearby gene expression. Additionally, we find that genes near or with an intragenic SVA have increased in expression compared to rhesus macaque, a species without SVAs in its genome. These data are important further support for the regulatory potential of SVAs and their impact on the evolution of gene expression dynamics, as has been discussed in other studies (Savage et al. 2013, 2014; Jacobs et al. 2014; Pontis et al. 2019). We show that repeat expansions within the VNTR regions of SVAs can affect the regulatory potential of the elements. In chapter three we show that ZNF91 binds SVAs at the VNTR region and its borders, which is in line with previous data showing the importance of the VNTR region for ZNF91-mediated repression (Jacobs et al. 2014). Interestingly, in chapter four we show that the gene-regulatory potential of structurally variable SVAs with varying VNTR lengths differs in the presence of ZNF91. This may therefore be mediated by more or less ZNF91 binding, depending on the VNTR length. Another fascinating mechanism that might be involved in the differential gene-regulatory potential of SVAs is the formation of guanine quadruplex (G4) structures in the VNTR. Stabilisation or disruption of G4s has been shown to alter gene expression (Siddiqui-Jain et al. 2002; Waller et al. 2009; reviewed by Lejault et al. 2021). They can form in regions with a high GC content, a characteristic of SVA VNTR and hexamer regions (Savage et al. 2013; Lexa et al. 2014). In fact, SVAs make up a substantial portion of predicted G4-forming DNA across the genome, given their relative size (Gianfrancesco et al. 2017). They can also facilitate binding of transcription factors, thereby promoting transcription (Spiegel et al. 2021). Whether the structural variations observed in SV-SVAs affect G4 formations is currently under investigation by others in our lab.

Future perspectives

The widespread involvement of TEs in gene-regulation is nowadays generally accepted. Studying the epigenetic landscape of TEs can give much insight into the potential of TEs to contribute to gene regulation in a cell-type specific manner. Recently, elaborate studies have focused on this; Pehrsson and colleagues (2019) mapped the epigenetic landscape of TEs throughout normal human development and in different human tissues. But epigenetic marks are at most predictors of gene-regulatory influences, and further functional analyses would be necessary to show the extent to which TEs are actually incorporated in gene-regulatory networks. This may be done by for example CRISPR interference (CRISPRi) or activation (CRISPRa) to repress or activate specific classes of TEs, as was very elegantly done by Fuentes et al. (2018) to reveal the genome-wide regulatory contribution of LTR5-HS elements to gene expression in human cells.

For SVAs, we did perform functional analyses, which opened up new questions related to how the differential regulatory potential of SVAs with varying VNTR lengths is mediated. Whether G4s, ZNF91 or other transcription factors are involved remains to be elucidated. Computational approaches to identify possible varying TF binding motifs within the VNTRs may be used, followed by subsequent ChIP-seq experiments to confirm binding of the TF to SVAs. By knocking-out *ZNF91* in hESCs, we generated a cell line that may be used for generating neuronal organoids, to assess if activated SVAs can influence neuronal gene expression.

Function of KZNFs in the regulation of TEs and gene expression

KRAB zinc finger genes (KZNFs) belong to the largest family of DNA binding proteins. We and others previously proposed that the continuous expansion of KZNF genes in mammalian genomes is explained by the evolutionary arms race model: new KZNFs evolve through segmental duplication and subsequent diversification, which can optimise them for recognizing and repressing new TEs once these invade the genome (Thomas and Schneider 2011; Jacobs et al. 2014; Castro-Diaz et al. 2014). This seems like a never-ending story, because in time some TE element can circumvent repression by accumulating mutations that interfere with KZNF repression, resulting in new waves of TE invasions. Newly emerged KZNF genes will need to be recruited to repress the novel TEs once they have managed to escape repression. Whereas the evolutionary arms race model is very clear for a number of KZNFs and TEs (Jacobs et al. 2014), there is another model that could explain the continuous growth of the KZNF gene family in mammalian genomes. This model suggests TEs and KZNFs become co-opted for gene-regulatory roles after the KZNFs have lost their original purpose to repress TEs that lost their transposition capacity (Wolf et al. 2020). In this model, the emergence of KZNFs as TE controllers can actually facilitate the domestication of TEs, as the activity of TEs can be dynamically regulated for the hosts' benefit. Importantly, the evolutionary arms race model is not at all incompatible with the model of KZNF/TE co-option, and we argue that both processes can happen subsequently or simultaneously. To make things even more complex, there are also KZNFs that do not seem to have any TE targets, but are conserved nevertheless. These might be co-opted for other, non-TE-mediated functions. The characterization of KZNFs was greatly advanced by the ChIP-mediated profiling of binding sites of over

280 KZNFs (Schmitges et al. 2016; Imbeault et al. 2017). Although these studies used effective techniques (ChIP-seq and ChIP-exo) to assess all possible targets of these DNA binding proteins, it remains to be established whether the KZNF binding sites are correct representations of the actual binding of endogenous KZNFs *in vivo* in various different cell types. For some TEs, more than one KZNF is identified as an interacting factor, something we also show in chapter two and three. This raises the question which KZNFs are actually necessary and crucial for repression of the target TEs. It also stresses the importance of downstream functional analyses, something we addressed in chapter two, three and four.

ZNF519 has a gene-regulatory effect independent of its TE target

In chapter two we study ZNF519, a recently evolved KZNF, and elaborate on its regulatory potential. Based on the gnomAD database, which contains almost 200,000 exomes and genomes, no homozygous loss-of-function (LoF) mutations are present in the *ZNF519* gene locus. Rare low-confidence heterozygous LoF are reported, but according to gnomAD's filtering the annotation or quality of these calls is dubious. This suggests that ZNF519 is an essential gene. Despite genome-wide binding of ZNF519 to MER52 TEs, we did not find support for a role for ZNF519 in the control of MER52 elements in our experiments. This suggests that ZNF519 may be a good example of a KZNF that was co-opted for other functions after its role of TE repression became redundant. Indeed, we find widespread binding of ZNF519 to gene promoters independent of MER52 elements, and our data support a gene-regulatory effect of ZNF519 on these genes. The binding of ZNF519 to gene promoters is not a unique phenomenon; around one third of all KZNFs that were tested in large binding assays had non-TE targets that include promoters (Schmitges et al. 2016; Imbeault et al. 2017). Most of them were ancient KZNFs and it was suggested that accumulating mutations may have masked any original TE targets, leaving only the KZNF-recruiting motif to be conserved. However, our investigation into the role of ZNF519 described in chapter two proposes an alternative explanation for KZNF binding to gene promoters. With its much later appearance in primate evolution, ZNF519 gives us a great snapshot of how co-option of KZNFs may actually develop. Comparative analysis between the binding sites in MER52 elements and gene promoters reveals a very high similarity in binding motifs. This suggests that the binding of ZNF519 to gene promoters may have been co-incidental and simultaneous with its optimization to MER52 elements. In that regard, it is important to realise that in order for retroviruses and TEs to be effective, they evolve to mimic mammalian gene promoters (or segments thereof) in order to recruit a specific set of transcription factors that are widely used in the tissue the retrovirus choose as a target. In the case of ZNF519, the virus ancestral to the MER52 elements in our genome may have mimicked the promoters that are now bound by ZNF519, even after the virus itself has ceased to exist. The work on ZNF519 therefore may provide a snapshot of the early stages of completed co-option, and it emphasises the long-term and multi-layered impact of an invasion of a retrovirus and the response of the host to control it.

ZNF91 mediates the gene-regulatory potential of SVAs

ZNF91, studied in chapter three, is a KZNF that we previously showed to fit well with the evolutionary arms race model (Jacobs et al. 2014). Its preferential targets are SVA (SINE-VNTR-*Alu*) elements, and ZNF91 binds the majority of SVAs present in the human genome.

We show that upon deletion of *ZNF91* in hESCs, around a third of the SVAs gain activating epigenetic marks. While we show that ZNF91 binds to all SVA subclasses, mainly evolutionarily young SVA elements become active in *ZNF91* knockout (KO) cells. A study by Pontis and colleagues (2019) showed that another KZNF, ZNF611, was able to repress SVAs in naive hESCs. We show that deletion of *ZNF611* does not activate SVAs in primed hESCs, suggesting that the regulation and control of TEs might be differentially regulated between cell types and developmental stages, potentially dependent on endogenous KZNF expression levels.

Along with the activation of SVAs, KO of *ZNF91* also resulted in the activation of a selection of genes in the vicinity of SVAs. This confirms that SVAs can alter gene expression by functioning as cis-regulatory enhancers. Our data is in line with previous research where a trans-chromosomal mouse cell line, naturally lacking *ZNF91* expression, was used to study SVAs on human chromosome 11 (Jacobs et al. 2014). Ectopic expression of *ZNF91* in these cells repressed SVAs and genes near SVAs. The distance to an SVA was the main factor determining the repressive effect seen on gene expression, supportive for a gene-regulatory effect mediated by SVAs. Besides the cis-regulatory activity of SVAs, we also show support for a gene-regulatory effect of SVAs by functioning as alternative transcription start sites (TSSs). As a result, chimeric SVA-gene transcripts were generated after *ZNF91* KO. Therefore, SVAs might be domesticated as essential components for transcription initiation in specific cell types or developmental stages (Miao et al. 2020; Damert et al. 2004). On the other hand, their exonization might also contribute to disease development (Knebelmann et al. 1995; Vervoort et al. 1998; Meischl et al. 2000; Zarnack et al. 2013). ZNF91 therefore fits well with both the evolutionary arms race model and the model of KZNF co-option and the combined integration of KZNF/TE-couples in gene-regulatory networks.

While we also observe ZNF91 binding at numerous gene promoters upon over-expression of the gene, we did not find support for a direct effect of ZNF91 on the genes it may bind to. Therefore, we assume that the main function of ZNF91 is the control of SVA elements and SVA-mediated gene regulation. While this makes it seem like *ZNF91* and *ZNF519* have taken different evolutionary paths to genome conservation, we may only have seen part of ZNF91's ongoing evolution so far. In a few million years, when the transposition activity of SVAs may have diminished, ZNF91, as another gene promoter-binding KZNF, might have become co-opted for similar TE-independent gene-regulatory functions like ZNF519.

Upregulation of KZNFs upon TE activation

In *ZNF91* KO cells, we observed a collective upregulation of SVA-rich KZNF gene clusters. One hypothesis is that these KZNFs, as guardians of the genome, are activated as a response to protect against the adverse effects of SVA activity. In mice, the activation of TEs by KO of the KZNF co-repressor *KAP1* was also observed concomitantly with activation of KZNF clusters (Kauzlaric et al. 2017). The distance of KZNF genes to a KAP1 peak did not play a role in the gene activation, pointing to an indirect effect similar to what we see upon *ZNF91* KO (Kauzlaric et al. 2017). Of note is that some KZNFs are direct KAP1 targets, and upregulated in human KAP1 KO cells (Tie et al. 2018). It remains elusive what the

effects of elevated KZNF expression are. One hypothesis is that the expression level of KZNFs is important for its binding-site specificity: higher expression would lead to less specific binding patterns and occupation of new genomic target sites by KZNFs. Therefore, increased expression of KZNFs might be a mechanism to allow KZNFs to explore new binding sites, thereby increasing the chances of suppressing the newly emerged TE family that elicited the elevated KZNF levels in the first place. In line with this reasoning, high expression of KZNFs is also seen in tissues with high TE activity, such as embryonic stem cells (Kauzlaric et al. 2017; Pontis et al. 2019) and the brain (Imbeault et al. 2017).

Future perspectives

A challenge that remains is obtaining an accurate profile of the genomic targets of KNZFs in different cell types. As explained above, overexpression of KZNFs might induce nonspecific binding, thereby resulting in false-positive target sites. At the moment, good quality and verified KZNF-specific antibodies are sparse, and their suitability for generating ChIP-seq datasets is very limited. An approach that might be useful to accurately pinpoint the primary targets of KZNFs at endogenous expression levels could make use of CRISPR-Cas9 to add epitopes for ChIP to the endogenous KZNF genes (Savic et al. 2015). While this was already proposed in 2015 by Wolf and colleagues (Wolf et al. 2015), so far, it has not been widely incorporated yet.

Another question that remains is whether ZNF519 was already co-opted in gene regulation in other primates, or whether this phenomenon is specific to humans. Similar studies like ours but in primate tissues can be performed to gain more insight in this. Other future developments will likely involve the functional assessment of other KZNFs, as the role for many of these proteins remains unexplored. Our work, together with previously published data shows that the function of KZNFs can be cell-type specific, and that their overall function therefore may not be easily assessed in any particular cell line. With the apparent widespread co-option of TEs in developmental and adult neuronal tissues, neurons may be an interesting cell type to further study the effect and co-option of KZNFs in gene regulation.

Structurally variable transposable elements in clinical genomics

In chapter four we show the importance of including structural variations in transposable elements in genome-wide association studies (GWAS) by linking structurally variable TEs to GWAS-defined disease risk loci. This is valuable work, since unravelling the genetic basis of disease still remains challenging to date. In 2002, the International HapMap Project was launched to provide a database of human genetic variation and insights into linkage disequilibrium between SNPs (International HapMap Consortium 2003). This facilitated the use of a set of 'tag' SNPs to predict genetic variants that are not directly assayed to indicate associations with phenotypes. The accuracy of this so-called genotype imputation however is low for rare variants, and is highly dependent on the size and diversity of the reference panel (Das et al. 2018). The 1000 Genomes Project, which ran from 2008 to 2015, provided a more comprehensive catalogue describing human variation (The 1000 Genomes Project Consortium 2010), and improved existing GWAS signals. Still, associations were made without having a clear and complete view of the

genome, and variation caused by for example TEs was not entirely taken into account. Further advancements came with efforts of The Structural Variation Analysis Group to uncover larger structural variations in phased (diploid) genomes (Sudmant et al. 2015). In the following years, diploid *de novo* genome assemblies of multiple individuals became available which were generated with long-read sequencing techniques (Chaisson et al. 2019; Porubsky et al. 2021; Ebert et al. 2021). These advances are critical for accurately studying genetic variations in TEs on a genome-wide level, because repetitive regions are difficult or impossible to detect with the paired-end-based output and SNP arrays currently used in GWAS. When we started our study in 2016, before all these advanced datasets were available, we took the laborious approach of predicting SVA structural polymorphisms with BLAST-mediated strategies on genomic DNA sequences, and verifying structurally variable SVAs by PCR and gel electrophoresis. The advent of long-read based phased genomes made it possible for us to confirm our findings and take a genome-wide approach for assessing structural variation in TEs.

Recently, scientists set a Guinness World Record by sequencing a complete genome in the astonishing time of only five hours and two minutes (Gorzynski et al. 2022). This ultra-rapid genome sequencing is based on the long-read Oxford Nanopore sequencing technology, which has proven efficient in detecting structural variation within TEs (Ewing et al. 2020). The tremendous developments in genome sequencing technologies and reduced price provide opportunities to assess the contribution of structural variation within TEs to complex diseases. It may soon change the way GWAS are performed.

The findings presented in chapter four stress the importance of including a wider variety of genetic variants in GWAS. I hope that in the near future, this thesis will help emphasise the need to develop affordable tools to have a better and more accurate mapping of structural variants in TEs. This will be a critical step to improve our understanding of genetic variation by structurally variable TEs in an evolutionary and medical context, and their application in clinical genetics.