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):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 1

Cis-Regulatory Activity and Expression of Transposable Elements and their KRAB Zinc Finger Repressors in Human Embryonic Stem Cells and Brain Organoids

Nina Haring
Elise van Bree
Marc Galland
Marten Smidt
Frank Jacobs
Abstract

Transposable elements (TEs) are mobile pieces of DNA that compose approximately half of the human genome. The evidence is mounting that they have played an important role in evolutionary innovation of gene regulatory networks. The main repressors of TE activity, the KRAB zinc finger (KZNF) proteins, evolved in parallel and have also been wired into gene regulation programs. Here we studied the cis-regulatory role of TEs in stem cell-derived human cortical organoids and midbrain dopamine (mDA) organoids by studying the transcriptional and H3K27ac enhancer profiles of TE subfamilies and KZNF proteins that repress them. We show that KZNF genes targeting the same TE family show similar expression dynamics and follow the expression pattern of their target TE. From the highly cell type specific H3K27ac profile on TEs we distilled that mainly young LTR elements are H3K27ac positive in hESCs, whereas mainly old TEs from several classes are H3K27ac enriched in neuronal tissues. We identified that LTR10 and MER4 elements show a hESC specific H3K27ac profile and contain motifs for pluripotency associated transcription factors (TFs). MER61F and LTR5B, on the other hand, were H3K27ac enriched in mDA neurons and LTR12E in cortical neurons. They showed enrichment of neuronal (subtype) specific TF motifs. By comparing different neuronal tissues derived from one stem cell lineage, our analysis provides a platform for discovering the evolutionary impact of TEs and KZNFs in rewiring neuronal cell type specific gene regulatory networks in the developing human brain.
Introduction

Almost half of the human genome is composed of transposable elements (TEs), mobile DNA sequences that can spread through copy and paste mechanisms. For long this highly repetitive part of our genome was viewed as ‘junk DNA’, but in recent years the evidence is mounting that TEs are widely utilized by the host species for evolutionary innovation of gene regulatory networks (Kunarso et al. 2010; Chuong et al. 2013; Sundaram et al. 2014; Chuong et al. 2016; Sundaram et al. 2017; Trizzino et al. 2017; Marnetto et al. 2018). Many examples have emerged showing how co-option of TE sequences have changed the regulation of important genes in a highly species-specific manner. By carrying various tissue-specific TF binding sites in their DNA sequence, the invasion of TEs can establish new networks of co-regulated gene expression (Wang et al. 2007; Kunarso et al. 2010; Sundaram et al. 2014), bringing previously unrelated genes under control of the same TF-bound TEs. This fosters the formation of new gene pathways and eventually the evolution of new structures and functions. Large scale analyses of TF binding sites and chromatin status provided insights into the extent at which TEs functions as cell type- and species-specific gene regulators (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; Trizzino et al. 2017, 2018; Pontis et al. 2019). Early embryonic development is an excellent example of how TEs are co-opted as regulatory elements and drive evolution. HERV-H/LTR7 and HERV-K/ LTR5-Hs show an epigenetic signature associated with enhancers 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 facilitate TF binding to the genomic HERV-H sequences and are essential for maintaining pluripotency of human stem cells (Wang et al. 2014; Lu et al. 2014). Another member of the HERV family, MER41, rewired the innate immune response by providing STAT1 binding sites essential for the expression of interferon-gamma dependent genes (Chuong et al. 2016). Transcriptomic analysis and P300/H3K27ac profiling suggest that specific TE classes are also co-opted as gene regulators in the developing brain (Notwell et al. 2015; Brattås et al. 2017; Trizzino et al. 2018; Li et al. 2019; Pontis et al. 2019). As the human brain is one of the most evolutionarily divergent organs, providing us with higher cognitive functions that distinguish us from other mammals, TEs could have played a big role in reorganizing gene expression
programs underlying this. However, our understanding of the role of TEs in neuronal gene regulatory pathways is still in its infancy.

TEs are epigenetically repressed through binding of specific KZNF genes of the KRAB zinc finger (KZNF) gene family. The high expression of KZNFs in the brain suggests that gene expression in the brain may be particularly sensitive to the regulatory activity of TEs (Imbeault et al. 2017). KZNF proteins are composed of a tandem array of zinc fingers that recognize specific DNA sequences and a KRAB domain that recruits co-factor KAP1 (TRIM28) which is a docking site for repressive epigenetic modifiers (Nielsen et al. 1999; Schultz et al. 2001, 2002; Sripathy et al. 2006; Wolf and Goff 2007; Turelli et al. 2014). KZNF genes evolved in parallel with TEs, driven by the occurrence of novel TEs after an expansion wave (Thomas and Schneider 2011; Jacobs et al. 2014; Imbeault et al. 2017). Given that TEs bound by a KZNF are frequently positive for either heterochromatin or enhancer associated histone marks suggests that KZNF proteins are regulating the availability of TE derived cis-regulatory elements (Imbeault et al. 2017). Additionally, the highly dynamic expression of KZNF genes across different human tissues is pointing towards a cell type specific regulation of TEs (Imbeault et al. 2017). However, direct evidence connecting KZNF expression levels to the presence/absence of heterochromatin/enhancer associated histone marks on the TE that they bind is lacking. To test the hypothesis that TE insertions have shaped gene regulatory networks in the human brain, we used human embryonic stem cells (hESCs) and differentiated them into cortical and midbrain dopaminergic (mDA) organoids. In these two neuronal tissues, derived from one parental cell line, we indeed find evidence for cell type specific expression of KZNF genes as well as the regulatory activity of TEs.

**Results**

**3D in vitro differentiation of cortical and midbrain dopamine organoids**

To identify TE subfamilies that adopted cell type specific cis-regulatory properties we differentiated H9 hESCs into cortical and mDA organoids using existing protocols (Doi et al. 2014; Field et al. 2019). For mDA organoid differentiation, an mDA fate was induced during the first 12 days in an adherent culture (Fig 1A). At day 12 (D12) spheres were generated that were cultured up to D42. mDA organoids were very uniform in size when isolated at D28 and D42 for subsequent RNA sequencing (RNA seq) and chromatin immunoprecipitation
Figure 1. 3D differentiation of midbrain dopamine neurons.
A) Schematic overview of differentiation protocol to generate mDA neurons from H9 hESCs adapted from Doi et al (2014). B) Bright field pictures of mDA spheres at day 28 (D28) and D42. C) Immunofluorescent staining of cryosectioned D42 mDA spheres showing expression of general neuronal (NeuN) and mDA specific (TH and FOXA2) markers.
followed by sequencing (ChIP seq) (**Fig 1B**). Immunofluorescent staining of cryosectioned mDA organoids showed high expression of neuronal and mDA markers at D42 (**Fig 1C**). Note that in the center of the organoids mDA neuron density is lower. Possibly the large size of the organoids hampered penetration of all medium components into the center. However, expression of general neuronal marker NeuN and DAPI suggest that these cells were alive and adopted a neuronal fate (**Fig 1C**). When mDA organoids were plated around D28 we observed neurite outgrowth and expression of mDA markers at D42 (**Suppl Fig1B+C**).

Cortical organoids were differentiated as described by Field et al (2019) (**Fig 2A**). At D35 organoids contained neural rosette structures positive for PAX6 surrounded by neurons characteristic of different cortical layers (**Field et al.**).
Fig. 2B). RNA sequencing was performed on three replicates of hESCs, D28 and D42 mDA organoids and two replicates of D28 and D42 cortical organoids (hESC and cortical organoid data generated by Field et al. 2019). In addition, H3K27ac ChIP seq was performed on two replicates of hESCs and D42 mDA organoids and three replicates of D35 cortical organoids. Principal component analysis and pairwise sample correlation revealed that the neuronal tissues are more similar to each other than to hESCs, but still display a high level of cell type specificity of gene expression and H3K27ac epigenetic marks (Suppl Fig 1A+B). Cortical and mDA identity was confirmed by showing expression of known cell type specific markers (Suppl Fig 1C).

Similarity in dynamics between KZNF gene expression and TE-derived transcripts
We assessed the expression dynamics of the KZNF gene family and TE subfamilies to get a better understanding of their role in acquiring cell type identity. We restricted our analysis to primate specific KZNF genes and TE families, as they may reveal the most recent evolutionary rewiring of neuronal gene expression. Interestingly, KZNF genes show a highly cell type specific expression pattern which was supported by similar specificity in the post mortem human brain (Fig 3A, Suppl Fig 3A). Compared to hESCs, KZNF expression levels were overall higher in cortical organoids and lower in mDA organoids (Suppl Fig 3B).

We made use of publicly available KZNF ChIP data (Imbeault et al. 2017) to investigate the expression dynamics of KZNF genes that share a target TE. We found that KZNF genes that target the same TE subfamily show a highly similar expression profile (Fig 3B, Suppl Fig 4). The concerted expression of KZNF genes seemed to be directly linked to the TE they repress because randomly
selected primate KZNF gene groups did not show similar patterns of co-expression (Suppl Fig 5). Next we studied the expression profile of TEs in relation to KZNF expression. Because TE-derived transcripts often lack annotation and have variable transcript structures, we used de novo transcript assembly to obtain their start and end positions. We considered a transcript TE-derived when a transcript start falls within a TE. Principal component analysis including only TE transcripts not overlapping exons revealed a strong cell type specific TE expression pattern, mostly separating hESCs from both neuronal tissues (Fig 4A). Although in hESCs collective TE expression levels are elevated compared to neuronal tissues, the diversity of expressed TE subfamilies appears to be higher in cortical organoids (Fig 3C, D). Finally, we combined expression dynamics of KZNF genes with the TE subfamilies that they target. Surprisingly, we found that some TE subfamilies showed an expression profile highly similar to the KZNF that represses them (Fig 3E).

Evolutionary young LTR elements are H3K27ac positive in hESCs, whereas old TE subfamilies of all classes are enriched for H3K27ac in neurons

To study the co-option of TEs as regulatory elements we obtained the H3K27ac profile of TE subfamilies in hESCs, cortical organoids and mDA organoids. In line with the transcriptomics data was the highly cell type specific H3K27ac profile on TEs (Fig 4B). Next we identified specific TE subfamilies that were significantly enriched in each cell type. H3K27ac peaks overlapping a TE were used as input for the TE-analysis pipeline to determine significant enrichment of observed over the average of 1000 expected overlaps (https://github.com/4ureliek/TEanalysis)(Kapusta et al. 2013). Enrichment (or depletion) was calculated by taking the log2 of the number of H3K27ac positive insertions from a given TE subfamily (e.g. LTR5-Hs, MER61F etc) over the average expected number of positive hits detected by chance. Only TE subfamilies belonging to the superfamilies DNA, LINE, LTR or SINE and with FDR<0.05 were included in downstream analyses. Cortical organoids showed enrichment of 140 TE subfamilies, mDA organoids of 132 and hESCs of 88 (Fig 5A). In the neuronal tissues all TE classes are enriched to a similar extent and DNA and LTR TEs are enriched most frequently (Fig 5B). Top ranking enriched TE subfamilies belong to all TE classes and are shared frequently between cortical organoids and mDA organoids, whereas they are depleted among H3K27ac positive TEs in hESCs (Fig 5C). MER126 is the DNA TE subfamily with the strongest enrichment in cortical organoids and mDA organoids and LTR12E the strongest enriched LTR (Fig 5B+C). MER126 is an ancient element originating in amniota, whereas
LTR12E is specific to Hominoidea.
In hESCs LTRs are the predominant and most strongly enriched TE class and the top-ranking LTR subfamilies are mostly depleted in neuronal tissues (Fig 5B+C). That we find LTR5-Hs and LTR7 among the highest enriched TE subfamilies is in line with other studies that associated these elements with early
embryonic development and regulation of pluripotency (Göke et al. 2015; Grow et al. 2015). This shows the robustness of our method. Noteworthy is that in hESCs we observe that mainly evolutionary young TE subfamilies (specific to the primate lineages towards human) are enriched, whereas they are depleted in neuronal tissues (Fig 5D). On the contrary, evolutionary old TE subfamilies (shared between ancient phylogenetic classes such as the tetrapoda and vertebrata) are enriched in neuronal cells (Fig 5D).

Next we generated H3K27ac profile plots of TE subfamilies that were significantly enriched in a cell type specific fashion according the TE-analysis pipeline. In addition, we verified that the H3K27ac peak was truly element specific and not an extension of a peak elsewhere. H3K27ac positive elements were used for a motif screen, to identify if and where in their sequence they contain binding sites for cell type specific TFs which potentially bind these TE-derived enhancers. We used LTR5-Hs and LTR7 as positive controls to show the trustworthiness of our method. Both show an element specific H3K27ac signal that is strongly enriched in hESCs (Suppl Fig 6). We found that a high fraction of H3K27ac positive elements (56-92%) contain motifs for TFs essential for pluripotency (Suppl Fig 6B+D). ChIP data analysis of several of these TFs in hESCs indeed showed that they bind LTR5-Hs and LTR7 elements (Kunarso et al 2010, Grow et al 2015, Pontis et al 2019). In addition, we found that LTR10C/D and MER4D (MER4D/MER4D0/MER4D1) display an hESC specific H3K27ac profile (Fig 6A+C, Suppl Fig 7A+B). LTR10 and MER4 are evolutionary young TE subfamilies originating in the LCA of old world monkeys (OWM) and primates respectively. In total we identified 33 H3K27ac positive LTR10C+D elements in hESCs. We used 17 LTR10D elements, all similar in length, to screen for pluripotency TF motifs. Multiple sequence alignment revealed that they were sometimes 5’ side truncated, but aligned well at the 3’ end. Therefore, we extended all elements from the 3’ end to the 5’ side of the shortest element, as CentriMo requires all input sequences to be the same length. We detected the enrichment of several pluripotency TF motifs at multiple locations spread out across the element: 53% contained a NANOG motif, 65% a SOX2 motifs, 35% an OCT4 motif and 41% a KLF5 motif (Fig 6B). In total we identified 142 H3K27ac positive MER4D/D0/D1 elements in hESCs. We used 42 elements (MER4D/D0/D1) that were approximately equal in length for a motif screen. 69% contained an OCT4 motif, 43% a SOX2 motif and 50% a NANOG motif (Fig 6D). All three motifs were detected ~80bp downstream of the center, which seems to coincide with the dip in H3K27ac signal (Fig 6C+D). MER61F and LTR5B are primate specific TE subfamilies that showed enrichment of H3K27ac in mDA organoids (Fig 6E+G,
Cis-Regulatory Activity and Expression of Transposable Elements and their KRAB Zinc Finger Repressors in Human Embryonic Stem Cells and Brain Organoids

Figure 5. H3K27ac enrichment on specific TE subfamilies in hESCs, cortical and mDA organoids

A) Venn diagram showing overlap of TE subfamilies that are overrepresented among H3K27ac positive TEs between different cell types. B) Boxplots showing log Odds ratio (Log2(observed/ave.expected)) of number of H3K27ac positive TEs (observed) over average of 1000 positive hits expected by chance. Only significantly enriched (FDR<0.05) TEs are displayed and grouped per TE Superfamily. Each dot represents a single TE subfamily. This data is obtained with the TE-analysis pipeline (Kapusta et al 2013, see methods). C) Heatmap showing Log2 Odds ratio of top 15 enriched TE subfamilies per cell type. Black indicates that zero elements were H3K27ac positive (no Log2 Odds ratio could be calculated). C) Heatmap showing Log2 Odds ratio of top 15 enriched TE subfamilies per cell type. Black indicates that zero elements were H3K27ac positive (no Log2 Odds ratio could be calculated). The colors of the TE names refer to the superfamilies displayed in B (e.g. grey TE names indicate that they belong to the LTR class). TE subfamilies that showed less than 10 insertions positive for H3K27ac were excluded from the analysis. D) Bar plots showing log2 Odd ratio (Log2(observed/ave.expected)) of number of H3K27ac positive TEs (observed) over the average of 1000 expected positive hits detected by chance. TE subfamilies were grouped based on evolutionary age. Stars indicate significant enrichment or depletion (FDR<0.05).
Suppl Fig 7C+D. We used 11 of 18 H3K27ac positive MER61F elements for a motif screen and showed enrichment of mDA-specific TF motif PITX3 in 71% of elements and neuronal-specific ASCL1 and PAX6 motifs in 47% and 59% of the elements (Fig 6F). In total we detected 35 LTR5B elements positive for H3K27ac in mDA organoids. Note that although the profile plot of LTR5B suggests that the H3K27ac profile is not highly mDA specific H3K27ac, the number of LTR5B elements that pass the MACS peak threshold in the other cell types is limited (Fig 6H). We used 12 H3K27ac positive LTR5B elements of similar sequence length in a motif screen. Several mDA-specific TF motifs were overrepresented in the majority of elements and the PITX3 motif was present in all of them (Fig 6H). Finally, we identified 12 Hominioidea-specific LTR12E elements that were H3K27ac positive in cortical organoids. Although 8 of the 12 elements were also H3K27ac positive in mDA organoids and hESCs, the H3K27ac level was on average higher in cortical organoids (Fig 6I). A motif screen across 9 full length LTR12E elements extended 1000bp upstream from their 3' end showed the enrichment of several general (NFYB and SMCA1) and neuronal-specific (BRN4) and cortical-specific (TBR1 and CUX2) TF motifs between the center and 3' end of the elements, the TBR1 motif was present in all elements (Fig 6J).

To summarize, we show that evolutionary young LTR elements are H3K27ac positive and contain pluripotency TF motifs in hESCs. Both neuronal subtypes generally display H3K27ac on older elements, although several primate specific subfamilies were overrepresented among H3K27ac peaks in cortical or mDA organoids and enriched for general neuronal and neuronal subtype-specific TF motifs.

Figure 6. TE subfamilies that potentially function as cell type specific cis-regulatory elements.
A+C+E+G+I) Profile plots showing average coverage of H3K27ac ChIP in RPKM on several TE subfamilies that showed enrichment in one cell type (LTR10C/D and MER4D/0/1 in hESCs, MER61F and LTR5B in mDA neurons, LTR12E in cortical neurons). All full length elements from one subfamily were included. Grey box indicates the approximate size of an element. B+D+F+H+I) Motif screen to detect enrichment and the location of TF motifs in H3K27ac positive TEs that may predict TF binding. Only TEs of approximately the same length were included (LTR10D: 483-522bp, MER4D/0/1: 850-922bp, MER61F: 585-619bp, LTR5B: 958-989bp, LTR12E: 1060-1127bp). Based on multiple sequence alignments TEs were extended from the 3' end or the center to the length of the shortest element to obtain sequences of exactly the same length. They were used as input for a MEME CentriMo local motif enrichment analysis to screen for the presence of cell type specific TF motifs. The plots show the probability of the best match to a given motif occurring at a given position in the input sequences. The percentages indicate the fraction of screened elements that contained this motif. Venn diagrams show the total number of H3K27ac positive TE insertions per cell type and the amount of overlap between cell types.
Cis-Regulatory Activity and Expression of Transposable Elements and their KRAB Zinc Finger Repressors in Human Embryonic Stem Cells and Brain Organoids
Discussion

Here we studied the co-option of TEs as cis-regulatory elements in hESCs and developing human neurons. The cell type specific expression of both TEs and KZNF genes is in agreement with findings in other tissues and supports the hypothesis that through coordinated activity and expression of TEs and KZNF genes, they have contributed to our genome’s transcriptional and epigenetic landscapes. The functional relevance of this coordinated regulation appears to be supported by the observation that KZNF genes targeting the same TE family show similar expression dynamics. Indeed, a recent study showed that KLF4 overexpression in hESCs induces the expression of HERV-H targeting KZNF genes, ultimately leading to H3K9me3 deposition on these elements (Pontis et al. 2019). The fact that several TE subfamilies show similar expression dynamics as their cognate KZNF suggests that KZNF genes are regulated in response to TE activity. It deserves some attention that generally KZNF and TE expression was lower in mDA organoids compared to cortical organoids and hESCs. Although this could reflect a neuron subtype specific regulation, we cannot rule out that this is a consequence of differences in the library preparation for RNA sequencing. Since we detected an equal number of H3K27ac positive TE subfamilies in mDA organoids and cortical organoids, it seems unlikely that TEs are co-opted less frequently as cis-regulatory elements in mDA neurons.

The highly cell type specific H3K27ac enrichment on TEs suggest that they play a role in developing and/or maintaining cell identity. Note that the neuronal tissues show much more similarities, suggesting the integration of TEs into pan-neuronal regulatory networks. We found that evolutionary old TEs are overrepresented among H3K27ac positive TEs in neuronal tissues. This is in line with a study describing a strong enrichment of p300 on MER130 and UCON elements, originating in the LCA of tetrapods, in the developing mouse cortex (Notwell et al. 2015). Yet, we also identified younger TE subfamilies that showed signs of neuron specific cis-regulatory activity. LTR5B and MER61F, H3K27ac enriched in mDA neurons, both originate in the LCA of humans and new world monkeys and the latter is bound by p53 (Wang et al. 2007). Interesting is that a large scale TF ChIP analysis revealed FOXA2 and OTX2 binding at LTR5B elements in various cell lines, supporting a potential role for this TE subfamily in mDA development (Ito et al. 2017: http://herv-tfbs.com). LTR12 elements, showing high H3K27ac in cortical organoids, have been identified as strong transcriptional drivers in various cell types (Yu et al. 2005; Beyer et al. 2016; Krönung et al. 2016; Sokol et al. 2016; Brocks et al. 2017). They are regulated by the ubiquitously expressed TF NFYB and a source of alternative
TSS’s, leading to the generation of LTR12-gene chimeric transcripts linked to cellular processes such as apoptosis (Krönung et al. 2016; Brocks et al. 2017). Very little is known, however, about their role in gene regulation in the brain.

The strong H3K27ac enrichment on evolutionary young LTRs in hESCs is in line with earlier data demonstrating that binding sites for pluripotency regulating TFs are poorly conserved and frequently TE derived (Kunarso et al. 2010). Here we identified that LTR10 and MER4D elements, in addition to earlier identified LTR7 and LTR5-HS, display hESC specific H3K27ac enrichment and contain binding motifs for pluripotency TFs. LTR10 elements emerged in the LCA of new and old world monkeys and a subset of them are bound and regulated by p53 (Wang et al. 2007). MER4D elements may play an important role in maintaining stem cell identity, as they are bound by SOX2, NANOG, and OCT4 in hESCs (Ito et al. 2017: http://herv-tfbs.com).

Note that the vast majority of TF ChIP experiments are executed in cell lines, so little is known about TE derived TF binding sites in the human brain. Large scale histone ChIP experiments, on the other hand, have provided insight into the epigenetic landscape at TEs in post mortem tissue derived from different brain regions and put forward specific TEs as potential cis-regulatory elements (Trizzino et al. 2018; Pontis et al. 2019). However, further mechanistic studies linking specific TE-derived enhancers to their target gene(s) is hampered by the fact that it is impossible to use these tissues for downstream experiments. The advantage of using hESC derived neuronal tissues is the possibility to experimentally assess the impact of the identified TE enhancers on gene expression using CRISPR or CRISRPi/a techniques.

In summary, we provide insights into how TEs are regulated and potentially function as gene regulators during human neuronal development. We suggest that the regulatory properties of TEs are controlled by dynamic expression of members of the KZNF gene cluster. Moreover, we propose several TE subfamilies that may play an important role in brain region specific gene regulation. These findings lay ground for future studies unveiling the role of TEs in the evolution of gene regulatory networks of the human brain.
Materials and Methods

hESC culturing

H9 human embryonic stem cells were grown on mouse embryonic fibroblasts (MEFs, amsbio) or matrigel (Corning) coated dishes. They were cultured in hESC medium consisting of: DMEM/F12 supplemented with 2 mM L-glutamine (Invitrogen), 20% knockout serum replacement (Gibco), penicillin/streptomycin (Invitrogen), non-essential amino acids (Invitrogen), 0.1 mM - 2-mercaptoethanol (Invitrogen). hESCs cultured on matrigel received hESC medium that was applied to MEFs for 24 hours before. Medium was supplemented with basic fibroblast growth factor (sigma, 8 ng/μl) and changed daily to secure pluripotency of hESCs. For maintenance of the culture, cells were grown in colonies and passaged manually by cutting the colonies with a needle.

Neuronal differentiations

mDA differentiation

For mDA differentiation a partial 3D differentiation protocol by Doi et al (2014) was used with slight adaptations. In short, hESCs are seeded as single cells using accutase and grown to 100% confluency. Then hESC culture medium was changed to differentiation medium containing: GMEM (Invitrogen), 8% KOSR (Invitrogen), 0.1 mM non-essential amino acids (Invitrogen), sodium pyruvate (Invitrogen), and 0.1 mM 2-mercaptoethanol (Invitrogen). This was considered D0 of differentiation. To induce a neuronal faith LDN193189 (0.1 μM, Sigma) was added from D0-D11 and SB431542 (10 μM, Sigma) from D0-D5. A midbrain dopamine identity was stimulated by the addition of purmorphamine (2 μM, Sigma) and FGF8 (100 ng/ml, R&D) from D1-D7 and CHIR99021 (3 μM, Sigma) from D3-D12. From D12 onwards neuronal maturation medium was used consisting of: neurobasal medium (Invitrogen), B27 without vitamin A (Invitrogen), Ascorbic acid (0.2 mM, Sigma), BDNF (20 ng/ml, R&D), GDNF (10 ng/ml, R&D), Dibutyryl cAMP (0.4 mM, Sigma). At D12 cells were dissociated with accutase for 10 min at 37°C and collected in neuronal maturation medium supplemented with rock inhibitor Thiazovivin (5 μM, Sigma). 2 x 10^4 cells per well were collected in ultra low attachment 96 well plates with round bottom (Corning) to form organoids overnight. Medium was refreshed every other day, or if a new factor was added, during the first 12 days. During the maturation phase the medium was refreshed every two to three days. Organoids were harvested at D28 and D42.

To study neurite outgrowth organoids were plated around D28. Two organoids per well were mechanically dissociated and plated on 8 well chamber slides coated with PLO/fibronectin/Laminin. Subsequently, they were cultured with normal media refreshments until IF analysis around D42.
Cortical organoid differentiation

The cortical organoid differentiation protocol described by Eiraku et al. (2008) and Field et al. (2019) was used with slight adaptations. Full grown hESC colonies (2-3mm in size) were lifted off in one piece from the MEF dish using a cell scraper and transferred to an ultra low attachment plate (Corning) to form embryoid bodies (EBs). They were cultured in neuronal differentiation medium containing: hESC culture medium supplemented with sodium pyruvate (Invitrogen) without additives such as bFGF. After 24 hours EBs were exposed to neuronal differentiation medium supplemented with: SB431542 (10 μM, Sigma), Cyclopamine (1 μM, Sigma), Dorsomorphin (1 μM, Sigma), IWR (3 μM, Sigma), this was considered D0. After applying this medium until D18, the medium was switched to neurobasal medium (Gibco) supplemented with N2 factors (invitrogen). Cyclopamine was added to this medium until D25. From D26 onwards cortical organoids were cultured in neurobasal/N2 medium without factors added.

Immunofluorescence

Immunofluorescent (IF) staining was performed on D42 cryosectioned mDA organoids and mDA organoids plated at D28 on PLO/laminin/fibronectin coated 8 well chamber slides and isolated at D38. mDA organoids for cryosectioning were collected in an eppendorf tube. After removal of all medium they were fixated in 1ml of 4% PFA in PBS for 13 minutes at room temperature (RT). Then they were washed twice in 1ml of 0.1% BSA in PBS followed by an overnight incubation in 1ml 30% sucrose at 4 °C for cryopreservation. Organoids were frozen per 5 in M1 embedding matrix (Thermo scientific) in cryomolds and stored at -80 °C. The frozen blocks containing the organoids were cryosectioned at a thickness of 16μm at -21 °C and stored at -80 °C. For IF the slides were thawed at RT and washed with PBS. Then they were incubated 15 min in 4% PFA in PBS for 12 minutes at RT followed by 3 washes of 5 minutes in PBS.

Plated organoids were washed twice with PBS followed by fixation with 4% PFA in PBS for 13 minutes at RT. Then they were washed 3 times with PBS for 5 minutes on a rocking platform. A different combination of antibodies was used in each well of the 8 well chamber slides, including a negative control. Slides with sectioned and plated organoids were blocked for 1hr in blocking solution (5% FCS, 0.1% BSA, 0.3% Triton-X 100 in PBS). Primary antibodies were applied in blocking solution and incubated overnight at RT. The following primary antibodies were used: goat-anti-FOXA2 (1:300, Santa Cruz SC6554), rabbit-anti-TH (1:1000, Pelfreeze P40101-150), sheep-anti-TH (1:1000, Millipore Ab1542), mouse-anti-NeuN (1:500, Chemicon MAB377), rabbit-anti-ALDH1A1 (1:100, Abcam Ab24343). Slides were washed 3 times 10 minutes with PBS followed by
a 2 hour incubation with secondary antibodies in blocking solution in the dark. The following secondary antibodies were used as a 1:1000 dilution: Goat-anti-rabbit 488 (Invitrogen, A11008), donkey-anti-rabbit 488 (Invitrogen, A21206), donkey-anti-sheep 488 (Invitrogen, A11015), goat-anti-mouse 594 (Invitrogen, A11005), donkey-anti-goat 594 (Invitrogen, A11058), donkey-anti-rabbit 594 (Invitrogen, A21207). Subsequently, the slides were washed with PBS for 10 minutes followed by a 3 minute incubation with DAPI (0.3μg/ml) in PBS. Then slides were washed 4 times 5 minutes in PBS and wells were removed from 8 well chamber slides. Slides were embedded with Fluorsave (Calbiochem) and covered with coverslips (24x60mm, Menzel Glaser).

**RNA sequencing**

Three replicates of hESC samples and two replicates of cortical organoids D28 and D42 were harvested by Field et al (2019, GSE106245 for hESC and cortical organoid D28, cortical organoid D42 data unpublished). Three replicates consisting of a pool of 8-10 organoids each were isolated from D28 and D42 mDA organoids. Total RNA was isolated using Trizol according manufacturer’s protocol, followed by a DNAse treatment and RNA purification using the RNA clean and concentrator kit (Zymo research). Library preparation was done at Genomescan by first isolating the mRNA using oligo-dT magnetic beads. Subsequent adapter ligation and PCR amplification was done with NEB Next Ultra Directional RNA Library Prep Kit (NEB #E7420S/L). Samples were sequenced at 150 bp paired end at an illumina Hiseq 2500 device.

**Mapping and quantification of RNA seq data**

A snakemake (Köster and Rahmann 2012) pipeline was written to trim, map and quantify RNA seq data. The complete pipeline can be found at DOI:10.5281/zenodo.2581199. First read quality was assessed with FASTQC. Below a short description of the separate steps composing the pipeline. Paired end Illumina RNA seq fastq files were used as input. Trimmomatic (Bolger et al. 2014) was used to clip adapter sequences and trim low quality reads. Reads were aligned to the human genome (Hg19 version) using STAR (Dobin et al. 2013) with default settings except: outFilterMismatchNmax=2 and outFilterMultimapNmax=10, outWigType=bedGraph, outSAMtype=BAM SortedByCoordinate. Base-by-base coverage tracks were generated from each bedGraph file with bedGraphToBigWig. Data are displayed on UCSC genome browser (session: http://genome-euro.ucsc.edu/s/ninaharing/Chapter_1). Raw read counts of genes were determined with FeatureCounts (Liao et al. 2014) using Hg19 KnownGenes.GTF from UCSC for annotation (downloaded 12th September...
2016). Only properly paired reads were counted (-B) and the library was reversely stranded (-s 2).
To identify TE derived transcripts Stringtie (Pertea et al. 2015) was used to build de novo transcript models from all samples mapped with STAR (bam files sorted by coordinate were used as input). Predicted transcripts with a minimum length of 100bp (-m) and minimum read coverage of 1.5 (-c) were included. One annotation file was generated from all separate stringtie transcript models that could be used by FeatureCounts to estimate raw counts of TE derived transcripts. Transcripts were considered TE derived if their start (plus and minus 10bp) showed overlap with a TE (Repeatmasker.GTF Hg19 from UCSC, excluding TEs that are located in exons). Only transcripts meeting the expression threshold (baseMean>10 in at least one cell type) were included in downstream analyses.
DEseq2 (Love et al. 2014) was used to normalize raw counts and perform principal component analysis of genes and TEs. R packages ggplot2 (Wickham 2009) and ggpubr were used to generate plots and perform statistical testing. Heatmaps were generated with MeV (vs 4.8.1).

**H3K27ac ChIP**
Two replicates of hESCs and D42 mDA organoids and three replicates of D35 cortical organoids were harvested. Replicates consisted of: a confluent 60mm dish hESCs, 21 D42 mDA organoids, and 7 D35 cortical organoids. Samples were snap frozen, stored at -80°C and thawed on ice for crosslinking or crosslinked immediately after harvest by adding 1 ml of 11x crosslinking buffer (50mM Tris-HCL pH 7.5, 100mM NaCL, 1mM EDTA, 0.5 mM EGTA, 11% Formaldehyde) to 10ml of cold cell suspension for 10 minutes at room temperature on a rocking platform. To quench formaldehyde glycine was added at a final concentration of 0.11 M and samples were incubated for 5 minutes at room temperature on a rocking platform. Cells were pelleted through centrifugation at 2000 rcf and 4 °C for 5 minutes. To remove all traces of crosslinking buffer, cells were washed two times with 10 ml cold PBS. A 2 ml tissue dounce (Kontes Class Co.) was used to dissociate the neuronal tissues prior to cross-linking. For cell lysis, 1 ml lysis buffer 1 (50 mM Hepes-KOH pH 7.5, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP40, 0.25% Triton X-100) supplemented with protease inhibitor cocktail (Roche) was added to cell pellets and incubated on a rocking platform for 10 minutes at RT. To remove all traces of lysis buffer, cells were washed two times with 10 ml PBS. A 2 ml tissue dounce (Kontes Class Co.) was used to dissociate the neuronal tissues prior to cross-linking. For cell lysis, 1 ml lysis buffer 2 (10mM Tris-HCL pH 8, 200 mM NaCl, 1mM EDTA, 0.5 mM EGTA) supplemented with protease inhibitor cocktail (Roche) was added to cell pellets and incubated on a rocking platform for 10 minutes at RT.
Cells were pelleted again by centrifugation (5 min at 2000 rcf at 4 °C). The pellet was resuspended in 100 μl lysis buffer 3 (10 mM Tris-HCl pH 8, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% Na-Deoxycholate, 0.5% N-Lauroylsarcosine) supplemented with protease inhibitor cocktail (Roche). Chromatin was sheared to fragments of ~500bp in 12 cycles of sonication (intensity high, 30 sec on and 1 min off) in a bioruptor. 450 μl lysis buffer 3 and 50μl 10% Triton X-100 was added to the sonicated cell lysate. Samples were centrifuged max speed for 10 min at 4 °C. 50 μl supernatant was stored at -20 °C to be used as input sample and the rest was used for ChIP. 50 μl of Dynabeads M-280 sheep anti-rabbit IgG (invitrogen) were washed three times with 0.5% BSA in PBS before incubating with 5 μg rabbit anti-H3K27ac (abcam, ab4729, lot GR3205521-1) on a rotator at 4 °C for 4 hours. Excess antibody was removed by three wash steps with lysis buffer 3. Beads were resuspended in 500 μl 1% Triton X-100 in lysis buffer 3 and added to the lysate. IP was performed on a rotator at 4 °C overnight. Immuno complexes were washed four times with 1ml RIPA buffer (50 mM Hepes-KOH pH 7.5, 500 mM LiCl, 1 mM EDTA, 1% NP40, 0.7% Na-Deoxycholate) supplemented with protease inhibitor cocktail (Roche) and once with 1 ml cold TBS (50mM Tris-HCl pH7.6, 150mM NaCl) or TE (10 mM Tris-HCLpH8, 1 mM EDTA, H3K27ac) using a magnetic stand. After removal of all traces of TBS, the beads were resuspended in 200 μl elution buffer (50 mM Tris-HCLpH 8, 10mM EDTA, 1% SDS). 150 μl elution buffer was added to thawed Input sample (50 μl). Chromatin was eluted and crosslinking reversed by incubating at 65 °C on a rocking platform. 200μl TE (10 mM Tris-HCLpH8, 1 mM EDTA) was added to each sample to dilute SDS in elution buffer. ChIP DNA was treated with Ambion RNase Cocktail (Invitrogen) for 30 min at 37 °C and Proteinase K (Invitrogen) for two hours at 55 °C. DNA was extracted with one step phenol/chloroform and two steps chloroform. Ethanol (100%) was used to precipitate DNA for 45 min at -80 °C. DNA was pelleted by centrifugation at 14000rpm for 45 min at 4 °C. The Pellet was washed with 70% ethanol and centrifuged at 14000rpm for 10 min at 4 °C. DNA was dissolved in nuclease free water. Finally, extracted DNA was purified using DNA clean & concentrator-5 columns (Zymo research). qPCR using the Quantifast or Quantitect SYBR green PCR kit (Roche) was performed on a Roche Lightcycler 480 II to check ChIP enrichment. For all ChIPs one or two positive and one negative control region were included.

**H3K27ac ChIP seq library prep and sequencing**

25-30 ng of ChIP DNA was used as starting material for Truseq ChIP sample prep (Illumina, Lot 20194738). Provided protocol was followed with the following adjustments: The DNA clean & concentrator-5 kit (Zymo research) instead of
beads was used for DNA purification. DNA fragment size selection with 2% agarose size selections gels (Invitrogen) using E-Gel sizeSelect II (Invitrogen) was done before and after PCR amplification of indexed ChIP DNA to obtain the 400-500bp fraction. ChIP samples were pooled and subjected to 150bp paired end sequencing.

**Mapping and quantification of ChIP seq data**

A snakemake (Köster and Rahmann 2012) pipeline was written to trim and map ChIP seq data. The complete pipeline can be found at DOI:10.5281/zenodo.2581325. Below a short description of the separate steps of the pipeline, using paired end Illumina sequencing fastq files as input. Read quality was assessed with FASTQC. Trimmomatic (version 0.38) (Bolger et al. 2014) was used to clip adapter sequences and trim low quality reads. Bowtie2 (version 1.0.1) (Langmead and Salzberg 2012) in –end-to-end –very-sensitive mode was used to map reads to the human genome (Hg19 version). Fragment length for valid paired end fragments was set to a minimum of 80 (-l) and a maximum of 500 (-X). The resulting sam files were converted to bam files sorted by coordinate using Samtools (version 1.7) (Li et al. 2009). The rmdup function of Samtools was used to remove potential PCR duplicates. Bedtools genomcov and bedGraphToBigWig were used to generate unnormalized base-by-base bigwig files coverage tracks. BamCoverage from the deeptools package (version 2.5.7) (Ramírez et al. 2016) was used to generate RPKM normalized base-by-base (--binSize 1) coverage bigwig file. Peak calling was done with MACS2 (version 2.1.1.20160309) (Zhang et al. 2008). All tags at a given location were included (--keep-dup all) and subpeaks were determined (--call-summits). For hESC and mDA samples we included all MACS peaks that we called in both replicates for downstream analyses. For cortical organoids we included peaks with a MACS score >=100 occurring in all three replicates. ChIP data are displayed on the UCSC genome browser (session: http://genome-euro.ucsc.edu/s/ninaharing/Chapter_1).

**H3K27ac ChIP seq data analysis**

MultiBigwigsummary and PlotCorrelation were used to display the correlation of H3K27ac profile in general and on TEs using Pearson correlation (Galaxy deeptools version 3.1.2.0.1). The TE-analysis pipeline vs4.6 (https://github.com/4ureliek/TEanalysis) (Kapusta et al. 2013) was used to determine enrichment of specific TEs. For each cell type H3K27ac peaks showing overlap with a TE (using RepeatMasker track of UCSC, Hg19) in BED format were used as input. The TE-analysis pipeline assesses the overrepresentation of specific
TEs in this file (the observed frequency) compared to the average of 1000 expected overlaps based on shuffling the genomic positions of TEs. A two-sided binomial test followed by an FDR correction was performed to determine enrichment (or depletion) of each annotated TE present in the BED file used as input. Only TEs with FDR<0.05 were included for downstream analyses. The log odds ratio (Log2(observed/ave.expected)) was determined to show the level of enrichment. Bedtools v2.17.0 was used to run the TE-analysis_Shuffle_bed.pl script with the following parameters: -f H3K27ac_peaks.BED -q RepeatMasker.out -n 1000 -1 hg19.chrom.sizes -g TEage.txt (distributed with pipeline) -s rm. Profile plots were generated with deeptools computeMatrix and plotHeatmap (Galaxy deeptools version 3.1.2.0.1). RPKM normalized bigwig files of replicate 1 of each cell type were used to display the average H3K27ac profile around the center of a group of TEs that showed enrichment of H3K27ac. The plotted window was adjusted to the (average) size of the element and only full length elements were included.

**Motif enrichment analysis**

We obtained a list of TEs overlapping a H3K27ac peak and focussed on the subfamilies that showed cell type specific enrichment according the TE-analysis pipeline. We screened these elements for the overrepresentation of cell type specific TF motifs with CentriMO of MEME-suite (version 5.0.4) (Bailey et al. 2009) CentriMO requires all input sequences to be of equal length, hence we only included TEs that were similar in length and extended them from the ‘3’ side or the center to the size of the shortest element. A multiple sequence alignment showed ‘5 truncation of H3K27ac positive of LTR10D, LTR5B, LTR12E, thus they were extended from the ‘3 end side. MER4D/0/1, LTR7, LTR5-Hs, and MER61F were extended from the center. CentriMO was run in local mode to also find uncentered enriched motifs, because we were interested in finding motifs across the entire element. To screen for TF motifs enriched in mDA specific H3K27ac positive TEs we included all peaks called for rep1. Correlation analysis showed that mDA rep2 is deviating from rep1 and the cortical samples, making it less trustworthy than rep1.
Supplemental figures and tables

A

<table>
<thead>
<tr>
<th>Days</th>
<th>H9 hESCs (single cells)</th>
<th>midbrain DA neuron</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Floor Plate induction</td>
<td>Neuronal differentiation &amp; Maturation</td>
</tr>
<tr>
<td>0</td>
<td>Adherent</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>Sphere</td>
<td>35</td>
</tr>
<tr>
<td>4</td>
<td>Sphere</td>
<td>42+</td>
</tr>
<tr>
<td>6</td>
<td>Sphere</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Sphere</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Sphere</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Sphere</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Sphere</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>Replate</td>
<td></td>
</tr>
<tr>
<td>42+</td>
<td>Replate</td>
<td></td>
</tr>
</tbody>
</table>

Culture

- LDN193189
- SB431542
- Pur+FGF8

Growth factors & small molecules

- CHIR99021
- BDNF+Ascorbic acid+GDNF+dibutyryl cAMP

B

Supplemental Figure 1. Plated mDA spheres show neurite outgrowth and expression of substantia nigra specific ALDH1A1.
A) Schematic overview of differentiation protocol to generate mDA neurons from H9 hESCs. Spheres were plated around D35 on PLO/laminin/fibronectin coated chamber slides. B) Immunofluorescent images of mDA neurons plated at D35 and stained at D42 showing expression of NeuN (general neuronal), TH (mDA specific) and ALDH1A (substantia nigra specific).
Supplemental Figure 2. RNA seq and H3K27ac ChIP of hESCs, cortical and mDA organoids show cell type specific gene expression and chromatin status.

A) Principal component analysis of gene expression  
B) Hierarchical clustering of Pearson’s correlation of H3K27ac ChIP samples across the entire genome (bins of 10kb) 
C) Hierarchical clustering of forebrain and mDA markers using z-score of normalized read counts of RNA seq data using the mean of replicates. 
D) Venn diagrams showing overlap of MACS peaks of H3K27ac ChIP replicates. For cortical organoid samples a MACS score cutoff of 100 was used.
Supplemental Figure 3. KZNF expression in different areas of the human brain in vitro and post mortem
B) Boxplots showing expression of KZNF genes (base-Mean>50 in at least one cell type in total 305 genes) across different replicates per cell type and time point. Wilcoxon rank sum test was used to compare all neuronal samples to hESCs. FDR was used to correct for multiple testing ****P<0.0001.
Supplemental Figure 4. Co-expression of KZNF genes that bind the same TE subfamily
Line plots showing expression dynamics (normalized read counts using mean across replicates) of primate KZNF genes that bind the same TE subfamily.
Supplemental Figure 5. Randomly selected primate KZNF genes do not show co-expression
Line plots showing expression dynamics (normalized read counts and z-score using mean across replicates) of randomly selected groups of KZNF genes.
Supplemental Figure 7. H3K27ac and (lack of) transcription at individual TE insertions
UCSC genome browser screenshots showing coverage tracks of H3K27ac ChIP of hESCs, mDA D42 and Cort D35 and RNA seq data of hESCs, mDA D42 and Cort D42 (pos= positive strand, neg=negative strand). Green box indicates TE.

Supplemental Figure 6. LTR5-Hs and LTR7 show hESC specific activity and are enriched for pluripotency transcription factor motifs.
A+C) Profile plots showing average coverage of H3K27ac ChIP in RPKM on LTR5-Hs and LTR7. Only full length elements were included. Grey box indicates the approximate size of an element. B+D) H3K27ac positive elements were extended from the center (only elements with the minimum size of the total window were included). They were used as input for a MEME CentriMo local motif enrichment analysis to screen for the presence of cell type specific TF motifs. The plots show the probability of the best match to a given motif occurring at a given position in the input sequences. The percentages indicate the fraction of elements that contained this motif. E+F) UCSC genome browser screenshot showing coverage tracks of H3K27ac ChIP of hESCs, mDA D42 and Cort D35 and RNA seq data of hESCs, mDA D42 and Cort D42 (pos= positive strand, neg=negative strand). Green box indicates TE.
References


Trizino M, Kapusta A, Brown CD. 2018. Transposable elements generate regulatory novelty in a tis-
sue-specific fashion. BMC Genomics **19**: 1–12.


Cis-Regulatory Activity and Expression of Transposable Elements and their KRAB Zinc Finger Repressors in Human Embryonic Stem Cells and Brain Organoids