



LUND
UNIVERSITY

Epigenetic and transcriptional effects of full-length LINE-1 retrotransposons in human neural progenitor cells

by
Ninoslav Pandiloski
May 2021

Supervisor: Dr. Christopher Douse
Principal Investigator: Prof. Dr. Johan Jakobsson
Molecular Neurogenetics Laboratory (A11)

Table of Contents

1.	Abstract	3
2.	Background	4
	iPSC-derived neural progenitors and their use in studying neural development	4
	Transposable elements influence genome function.....	6
	TE classification and LINE-1 elements.....	7
	Epigenetic regulation in neurodevelopment and TE control.....	9
3.	Aim	10
3.	Materials and Methods.....	10
	Cut&Run and Bulk-RNA sequencing on NPCs	10
	Mapping sequencing reads over transposable elements	11
	Genomic locations of TEs.....	12
	Visualization of L1 epigenetics and differential peak calling	13
	Differential gene expression with featureCounts and DESeq2	13
4.	Results.....	14
	Genomic locations of full-length L1 and HERV elements	14
	Quality Control.....	14
	Defining the epigenetic status of FL-TEs in NPCs.....	16
	Quantitative transcriptional analysis over L1HS regions	26
5.	Discussion.....	33
6.	References	37

1. Abstract

Transposable elements (TE) are parasitic genetic entities that have the ability to alter their position within a genome and have the potential to influence its functionality. Although TEs play beneficial roles in the evolution of their host, they are active in various human disorders including neurological diseases, where may contribute to the progress of the disease. Long Interspersed Nuclear Element-1(LINE-1, L1) is the dominant, autonomously-transposing TE family in all mammals and some the evolutionary-young LINE-1s are thought to be active in human neural progenitor cells (NPCs). We defined the epigenetic environment and transcriptional activity of the seven evolutionary-youngest L1 subfamilies in human NPCs, and compared this profile to human endogenous retroviruses (HERVs). Our analysis of two heterochromatin-associated histone methylation marks (H3K9me3, H3K27me3) suggested that H3K9me3 is responsible for silencing a majority of the studied TEs, whereas H3K27me3 is significantly less present. Profiling of two epigenetic marks associated with transcriptional activity (H3K4me1, H3K4me3) revealed that the presence of H3K4me3 is a better predictor of LINE-1 transcription than H3K4me1 for those elements that evade repression. Furthermore, we investigated the transcriptional differences between human and chimpanzee NPCs, focusing in particular on genes containing human-specific LINE-1 elements (L1HS). We found 11 genes containing a full-length L1HS sequence that showed downregulation in human compared to the chimpanzee NPCs. A Chi-square test suggested dependency between the presence of L1HS and the downregulation of genes, perhaps due to induction of H3K9me3-heterochromatin. Looking ahead, this project provides a base for further exploration of functional roles of LINE-1 elements in human NPCs and a template to investigate mechanisms of LINE-1 and HERV regulation by chromatin regulators.

2. Background

iPSC-derived neural progenitors and their use in studying neural development

The brain is the center of every mammalian nervous system. Its transcriptional complexity and the epigenetic mechanisms that influence transcriptional programs during development, are being uncovered now more than ever [1]. The cerebral cortex, as the largest site of neural integration in humans, allows for unique human functions such as speech, perception, awareness, thought, and consciousness [2]. Because of cortex's central role in cognitive and emotional processing, understanding the organization and the mechanisms that govern its development remains the focus of neuroscience [3]. That knowledge allows us to not only understand why our brain differs from other mammals, but also explore the mechanisms behind neurological disorders.

As the overall lifespan of humans has increased over the years, so has the prevalence of neurodegenerative disorders, many of which are age dependent and greatly vary in their pathophysiology. Thus, effective ways to study and treat them are essential. Animal models present powerful experimental tools when elucidating the mechanisms and the genes involved in the disease, but may fall short when discovering-human specific therapeutics [4,5]. Recent advancements of *in vitro* methods such as reprogramming of adult somatic cells (*e.g.*, fibroblasts) to human induced pluripotent stem cells (iPSCs) have great potential in furthering our understanding of these disorders.

In the iPSC protocol, a cocktail of transcription factors can be used to generate stem cells from patients, which allows for patient-specific experiments without the use of genetic modifications, thus excluding any variation induced by the genetic modification step [6]. This represents an advantage when investigating neurological disorders, as most of them have a strong genetic component which is captured completely by

the iPSCs. Furthermore, iPSCs have the capacity of self-renewal and can be differentiated into any cell type using specific combinations of growth factors and culturing conditions. These features make iPSCs an unlimited source of patient-derived cells that may be used for studying human specific neurodevelopment [7].

One approach to study the development and disorders of the human brain is reprogramming human derived iPSCs into neural neural progenitor cells (NPCs) [5]. These NPCs are of particular interest as they can provide an expandable and well characterized source for specific neural and glial cell subtypes, including the ones found in the cerebral cortex. That states the important potential of NPCs when investigating the neurogenesis of the human cortex and it's disorders, as well as discovering cell therapies [8].

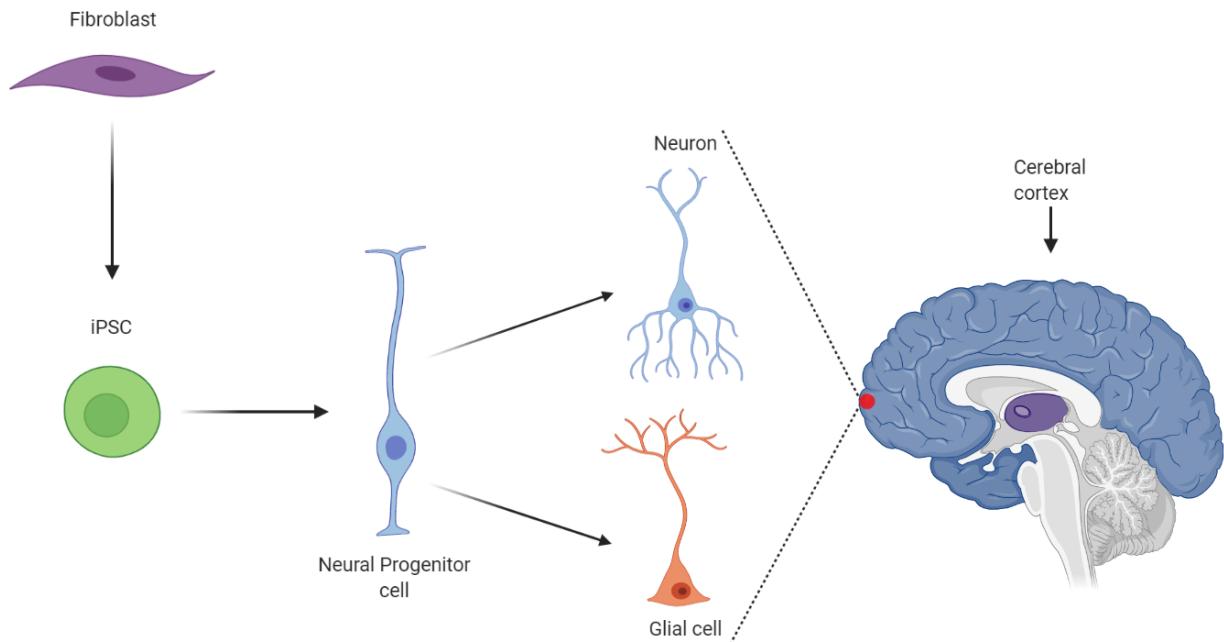


Figure 1. Schematic representation of differentiating iPSCs into NPCs, and their role in the neurogenesis of the cerebral cortex

Transposable elements influence genome function

First discovered by studying maize kernel mosaicism, TEs are repetitive DNA sequences that have the ability to alter their position within genomes. They are present in nearly all eukaryotes, and can be transferred either to offspring through inheritance or between non-mating organisms via horizontal transfer [9,10]. The initial sequencing of the human genome suggests that 45% is comprised of TEs, which is likely a conservative estimate since some elements may have diverged beyond recognition. In contrast, only 2% of the human genome is comprised by sequences encoding for proteins, illustrating the scale of potential contribution TEs might have on the structure and function of the genome [11].

It is well established that genetic neuronal alterations give rise to various neurodegenerative disorders and that environmental exposures contribute to their pathogeneses. Epigenetic dysregulations also play a key role in the disease processes and indicate that transposable elements (TEs) display aberrant activation, which may contribute to the progress of the disease [12]. Corroborating this idea is Rett syndrome, where a mutation in the TE regulator MECP2 enables mobilization of LINE-1 transposons and disease-related genetic mutations can influence the frequency of its retrotransposition [13]. Additionally, with evidence growing on TE activity in psychiatric disorders, it is becoming increasingly clear that they might play a role in human disease [14,15].

The primary threat TEs pose for organisms is their transposition in both somatic and germline cells, which could lead to mutagenic insertions upon retrotransposition or cause element polymorphism. Both events can result in detrimental and/or pathological consequences, particularly when they occur in the organism's early development [16,17]. To avoid this, most TEs are kept under transcriptional repression, for example by recruiting epigenetic silencing complexes, such as the TRIM28-KRAB-ZNF system or the HUSH complex [18]. However, TE repression may also affect the genome's functionality, as changes to the local

chromatin environment may hinder the access of the cell's transcriptional machinery on nearby genes [19,20].

Furthermore, as some elements display transcriptional activity in humans, they can also be a source for regulatory noncoding RNAs or be translated into TE-derived peptides. TE derived peptides or nucleic acids have the potential to trigger an innate immune response, which could contribute to the pathogenesis of human diseases [21,22]. Intriguingly, despite the diverse negative genomic effects caused by TEs, they have also been recognized as beneficial gene regulatory elements in specific pathways [23]. Besides being noncoding RNAs and TE-peptides, the nature of the elements also allows them to be a functional site for gene regulatory factors, thus serving as an alternative promoter or a gene enhancer. In fact, even the genomic mosaicism caused by somatic TE transposition has been suggested to contribute to individual differences and genomic functional diversity [16].

TE classification and LINE-1 elements

Found in an incredibly large number of forms and varieties, TEs display deep evolutionary origins with the host organism and constant diversification. However, despite their diversity, we classify TEs into two broad classes defined by their mechanism of transposition: DNA transposons (class II) and retrotransposons (class I) [24]. DNA transposons replicate with a “cut-and-paste” mechanism where a circular DNA intermediate binds to a target site, both specifically and unspecifically. Though evidence of DNA transposon activity has been reported in the primate lineage, it was estimated to have ceased ~37 mya, therefore DNA transposons are sometimes referred to as fossils in human genomes [25].

Retrotransposons transpose via a “copy-and-paste” mechanism – an RNA intermediate is reverse transcribed into a cDNA and integrated into different locus of the genome. Retrotransposons can be further

divided into those with a long terminal repeat (LTR) and those without (non-LTR). LTR retrotransposons generally depend on viral integrases for their transposition, displaying characteristics similar to retroviruses, both in replicative the mechanism and structure [26]. Prominent examples of LTR retrotransposons include several families of endogenous retroviruses (ERVs). In contrast, non-LTRs rely on target-primed reverse transcription, where the cDNA is reverse-transcribed from an RNA template directly onto a chromosomal site [27]. Far outnumbering other TEs, non-LTR retrotransposons are the most common human transposon. One family, the long nuclear interspersed elements (LINEs), comprises about 20% of the human genome. Although LINEs also contain degenerated copies, it is estimated that 80 – 100 elements of the LINE-1 (L1) class are retrotransposition-competent in humans [28].

LINE-1 is the dominant, autonomously replicating non-LTR in all mammals and its activity has been correlated to various malignancies and hereditary diseases in humans [29]. Phylogenetic analyses suggest that several families of L1s have competed for host factors, however only the primate specific L1PA families remained active within the evolution of the anthropoid lineage [30]. A full-length (FL) element is ~6kb long and contains a unique 5' untranslated region (UTR) with regulatory features, and a bicistronic RNA encoding for two open reading frames, ORF1 and ORF2 [31]. The resulting ORF proteins form a ribonucleoprotein complex, facilitating the autonomous transposition of L1 elements.

Transcription of L1s is controlled via an internal promoter located in the 5'UTR, making the elements less dependent on the integration site than other transposons that lack their own promoter. Moreover, the 5'UTR contains an antisense promoter, driving the transcription of adjacent genes, along with binding sites for factors that are essential for transcription initiation and enhancer function [32,33]. Upon insertion of L1s, sequence mutations accumulate at a neutral rate, suggesting that older subfamilies display a higher mutation rate, thus showing lesser activity as conserved functional sites degenerate [34]. Understanding these L1 characteristics allows for a more concise estimation of the evolutionary age of L1PA subfamilies and discovery of the youngest, human-specific L1 subfamily (L1HS).

Epigenetic regulation in neurodevelopment and TE control

Epigenetics studies heritable changes to genome function that arise without altering the DNA sequence: for example, by chemical modifications to histones or the DNA itself [35]. Changes in the epigenomic landscape correlates with changes in the chromatin structure that can affect transcriptional activity. In humans, such regulatory mechanisms are crucial for proper neural development [36]. Indeed, accurate differentiation of neural progenitors into mature neurons or glial cells depends on epigenetic mechanisms [37].

Furthermore, epigenetic mechanisms have evolved in eukaryotic cells to silence the expression of TEs and reduce their mobility [38]. Genomic regions containing TEs are also usually enriched with particular chromatin modifications, for example DNA methylation and the trimethylation of histone H3 lysine 9 (H3K9me3). Together, these epigenetic marks attract a repressive chromatin structure that is less accessible to the transcriptional machinery. Mutations to genes maintain such a repressive epigenetic status leads to re-activation of TEs. For instance, inducing global DNA de-methylation in human NPCs, through knockout of maintenance DNA methyltransferase DNMT1 leads to upregulation of evolutionarily young L1s [39].

Dysregulation of the epigenome can lead to devastating consequences, both during embryonic and adult neurogenesis. Apart from downregulating genes crucial for proper neuronal functioning, new evidence correlates epigenomic dysregulation with several neurodegenerative disorders [40]. Since specific epigenetic modifications can be associated with transcriptional activation or repression, studying these modifications in neuronal cells may provide a better understanding of epigenetic and TE functions in neurodegenerative pathologies and neurogenesis. In the long term, understanding the underlying mechanisms is necessary to uncover new therapeutic intervention strategies.

In this project, we define the epigenetic modification marks found around all full-length LINE-1 and HERV elements in human NPCs and investigate the transcriptional activity of the elements themselves and the surrounding genes. We also compare the expression found around human specific L1HS elements to the expression found in NPCs of our closest living evolutionary relative – the chimpanzee. Together, the project tests a central hypothesis that transposable elements are important gene regulatory elements in neuron progenitor cells.

3. Aim

With this project we aim to locate full-length (FL)-LINE1 elements in the human genome and analyze their epigenetic status and transcriptional activity in human neural progenitor cells (NPCs). This was achieved by defining the enrichment of four different epigenetic marks over the LINE-1 elements and comparing the resulting profiles to the RNA-seq signal which detects transcriptional activity in the same regions. Furthermore, we examine whether human-specific FL-L1HS can cause changes in the transcriptional activity of human genes. We approached this by performing differential expression analyses between human and chimpanzee NPCs and specifically looking at conserved genes that contain L1HS. In summary, we define the transcriptional activity around human L1s in NPCs and observe the effects of the human specific L1HS on human NPC gene expression.

3. Materials and Methods

Cut&Run and Bulk-RNA sequencing on NPCs

iPSCs were derived from four samples, two humans ($n=2$; H1 and H2) and two chimpanzees ($n=2$; Ch1 and Ch2). The pluripotent stem cells were further differentiated into NPCs, following a protocol established in the Molecular Neurogenetics Laboratory at Lund University [41].

Profiling of histone methylation marks in the NPCs was achieved using the Cleavage Under Targets and Release Using Nuclease (Cut&Run) sequencing. A technique that combines antibody-targeted cleavage with paired-end sequencing to identify binding sites of DNA associated proteins. Cut&Run negative controls are obtained using an unspecific Immunoglobulin G (IgG) to establish a reference background for peak calling and set a baseline for the experiment. It differentiates from other techniques by producing data with low background signal and a higher sensitivity. Libraries for the experiments were designed using Hyperprep kit (KAPA) and further sequenced with Illumina NextSeq, 75bp paired-end reads (150 cycles).

Bulk-RNA sequencing was performed to reveal the gene quantitative expression in both chimp and human NPCs. As the name suggests, bulk-RNA is performed on a population of cells often used when testing for gene specific differences between conditions. Subsequently to RNA purification and extraction, cDNAs of fragmented RNA was synthesized and later used for library preparation. In this experiment, bulk-RNA library was prepared using TruSeq RNA Library Prep kit v2, and reads were sequenced with Illumina NextSeq. The sequencing of bulk-RNA, however, differentiates from Cut&Run as it uses 150bp paired-end reads (300 cycles), thus increasing the sequencing depth and specificity.

Mapping sequencing reads over transposable elements

Because of their abundance, as well as their similarities in the repeats, TEs have always presented technical challenges for sequence alignment programs, especially when exploring families, such as the L1s. That is because members of the family display highly similar, if not identical sequences in multiple regions of the genome, thus creating ambiguities when mapping reads. This problem was resolved computationally, by either setting a limit on the quality of alignment or performing unique mapping – where reads map to exactly one location within the reference genome. Prior to any alignment, the sequencing quality of samples was

tested with FastQC (v. 0.11.8) and MultiQC (v. 1.2). For comparative purposes, both human and chimp samples were mapped to the human GRCh38 (hg38) genome.

Because Cut&Run has a different sensitivity than the usual alignment dataset, mapping of the methylation marks was carried out using Bowtie2 (v. 2.3.4.2) with a few deviations from the default parameters. Besides setting a sequencing accuracy greater than 99.9%, all singletons and improper pairs were removed from our dataset using the following flags: `--local --very-sensitive-local --no-unal --no-mixed --no-discordant --phred33 -I 10 -X 700`. Subsequently, an alignment score of MAPQ ≥ 10 was set removing all hits that have a chance greater than 10% to originate elsewhere.

The bulk-RNA mapping was performed with STAR (v. 2.6.0c), using Gencode35 GTF as annotation guide. STAR reads the guide file and extracts the splice junctions of genes, significantly improving the mapping accuracy. Furthermore, to ensure unique mapping, sequenced reads could map to a maximum of 1 locus and the allowed mismatch ratio was decreased by 90% from the default. The mapping output of both approaches was subsequently transformed into other file types, with which the alignments could be observed as a signal, where a higher number of aligned reads corresponds to a more intense signal.

Genomic locations of TEs

The coordinates of the seven evolutionarily youngest, full-length L1s, were extracted from RepeatMasker (v. 4.1.1) – program screening genomes for DNA sequences with low complexity and interspersed repeats. The genomic locations of full-length human ERVs (FL-HERVs) on the other hand, were a kind gift from a collaborator – Patric Jern. It is possible to describe the epigenetic status of the extracted elements by intersecting them with the mapping signal with the sites of interest. Additionally, file manipulation with

\BEDTools (v. 2.26.0) and SAMTools (v. 1.9) allowed for subsetting the L1 genomic regions into categories *e.g.*, intronic, transcriptionally active or repressed.

Visualization of L1 epigenetics and differential peak calling

The L1 and HERV coordinates were intersected with the signal files to compute matrices describing signal scores over their genomic regions. DeepTools (v. 2.5.4) was used not only in creating the signal matrices, but also to visualize them. Depending on the observed enrichment, some matrices were sorted according to signal intensity, and others according to genomic coordinates, 5' – 3'. Each genomic region was covering 16kb – 6kb covering the body of the element, and 5kb spanning each side of the transposon. The epigenetic status was also observed around regions where genes and L1s coincide.

To identify genomic regions enriched with the methylation marks, differential peak calling was carried out using Sparse Enrichment Analysis for Cut&Run (SEACR). This method identifies areas that are significantly enriched with a specific methylation mark in comparison to its respective IgG background. SEACR is able to calibrate its peak calling threshold using the IgG background, thus discriminating with near-perfect specificity between true and false positives. The coordinates of methylation peaks were also intersected L1s and visualized using the DeepTools method stated above.

Differential gene expression with featureCounts and DESeq2

To quantify the gene RNA expression in human and chimpanzee NPCs, the bulk-RNA mapped signal was quantified using with featureCounts. It also implements chromosome hashing and feature blocking along with other strategies to efficiently assign reads. FeatureCounts produces a count file containing the number of aligned reads onto a specific genomic feature. The count matrix is later then used perform differential

expression analysis with DESeq2 – an R package designed for normalization, visualization, and differential expression analysis of the given count data. It employs Bayes estimations to estimate the dispersion and log fold changes, and it is considered the standard for RNAseq differential analysis.

4. Results

Genomic locations of full-length L1 and HERV elements

RepeatMasker was browsed to identify L1 elements in humans and extract their genomic coordinates, along with other information such as length, subfamily, and orientation. To ensure that our database contained only full-length L1s (FL-L1s), all extracted elements with sequences shorter than 6kb were removed from further analysis. We then extracted the members of the seven evolutionary-youngest LINE-1 subfamilies specific to primates (L1PA1-7), resulting in a total of 6,156 FL-L1 elements. The genomic coordinates of 7,036 full-length human endogenous retroviruses (FL-HERVs) were provided by Patric Jern (Uppsala University). To capture the status of the chromatin surrounding the transposable elements themselves, the length of each element was extended by 5kb upstream and downstream of the element body. Additionally, since young L1s have shown activity in NPCs specifically, we further separated the seven youngest subfamilies into groups of their own. This represented our database of genomic coordinates for all full-length L1s and HERVs (FL-TEs).

Quality Control

The CUT&RUN sequencing method [42] was employed to profile the presence of the four epigenetic marks in human NPCs (n=2; H1 and H2): H3K4me1, H3K4me3, H3K9me3 and H3K27me3. Subsequent to each sequencing step, FastQC and MultiQC analyses were performed which tested the quality of the raw data and revealed no sequencing issues. The reads were then mapped to a reference human genome (hg38) using

bowtie2, producing Sequence Alignment Maps (SAMs), which after several transformations finally take the form of a Binary Alignment Map (BAMs). These files contain information about the reads and their alignment against hg38, including an alignment score, MAPQ. The MAPQ scores represent transformation the probability a read is mapped incorrectly, calculated by $-10 \times \log_{10}(p)$. Since transposons are repetitive and abundant in the genome, one major challenge associated with profiling their epigenetic status and transcriptional activity is that reads map to multiple locations, decreasing the confidence of each alignment to unique loci. We tested the alignment quality of sequencing data for all epigenetic marks over the elements by reporting the average MAPQ score of aligned reads. We then tested the alignment quality over all HERVs and separately for the seven L1 subfamilies.

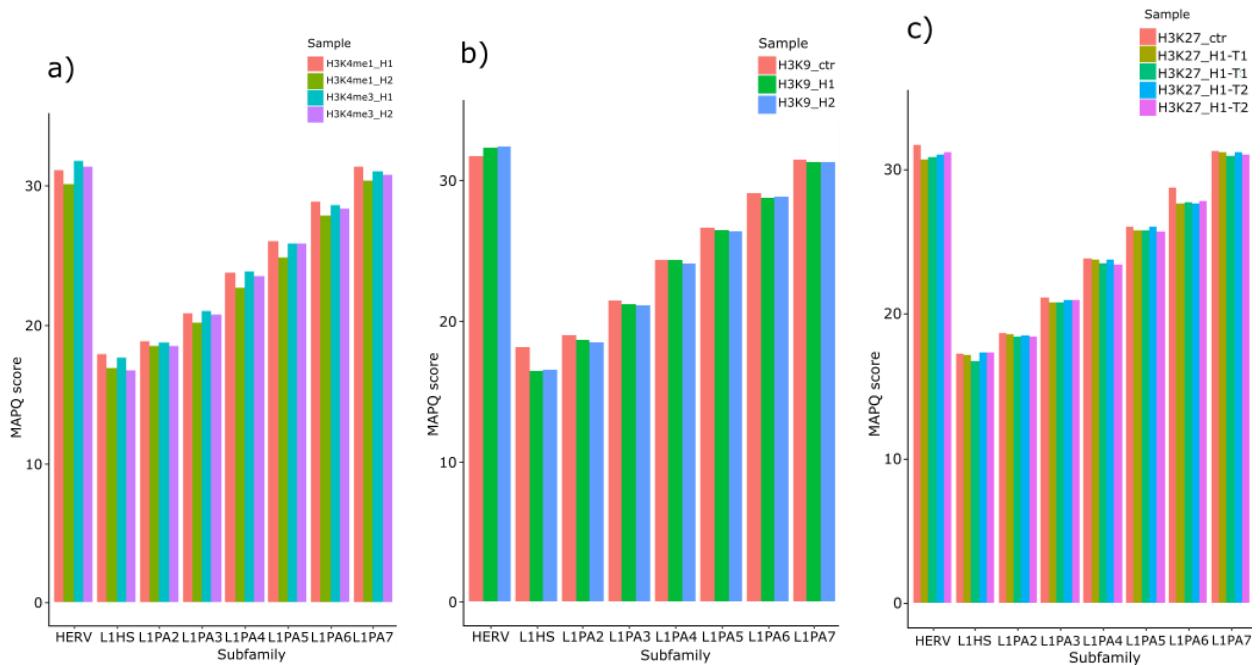


Figure 2. Average alignment quality of (a) two H3K4 activation marks, (b) H3K9me3 repressive mark, and (c) H3K27me3 repressive mark over L1s and HERVs. L1 elements are sorted by evolutionary age, starting with youngest – L1HS.

A similar quality of alignment is observed between the four epigenetic marks when mapping them over L1s and HERVs. Namely, steady improvement of alignment score was observed as we move towards older L1s, revealing a correlation between the evolutionary age of L1 elements and their mappability (Figure 2). This is consistent with transposons accumulating mutations over time that increase the uniqueness of each

sequence. Accordingly, reads for all epigenetic marks plotted over HERVs (which are generally older than the L1PA families) show an average alignment score of 30 or higher (Figure 2). Bulk-RNA sequencing was performed on identical NPCs to detect the transcriptional activity present around the elements, to compare the observations made in the epigenetic environment. No filtering or transformation steps were required as STAR allows unique mapping and directly outputs a BAM file.

Defining the epigenetic status of FL-TEs in NPCs

To define the epigenetic status of L1s and HERVs in NPCs, we retained reads with MAPQ score > 10 (i.e. a high confidence of unique mapping), and transformed the BAM files into BigWigs – displaying dense, continuous data on a genome browser. The four marks and Bulk-RNA signal were plotted over each of the seven subfamilies of FL-L1s and their extended 5kb flanking regions (figure 3). DeepTools was used both for calculating the signal matrices and visualizing them.

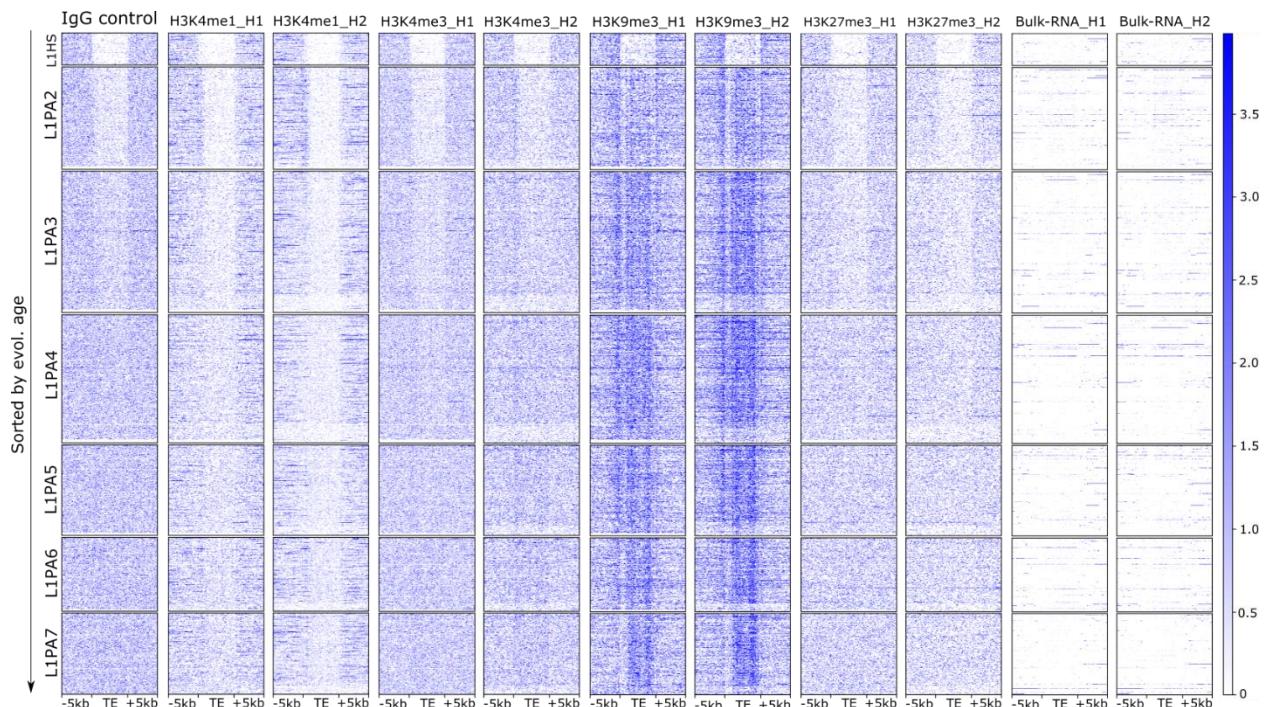


Figure 3. Relative presence of the four epigenetic marks over the youngest FL-L1s and their 5kb flanking regions. The transcriptional expression, obtained from the Bulk-RNAseq, is also shown.

One can get an overall look of the data with a heatmap showing the relative presence of the two activating marks H3K4me1/me3, the repressive H3K9me3 and H3K27me3, and the observed transcriptional activity over FL-L1s. The rows are grouped in seven blocks representing each of the L1 subfamilies, sorted by evolutionary age, starting with the human specific L1HS at the top. Each row within a block is a unique location of an L1 element that belongs to a specific subfamily. Each column depicts 16kb of 6,156 unique L1 locations (5kb + 6kb FL-L1 + 5kb) in the human genome (figure 3).

Both the epigenetic marks and the Bulk-RNA sequencing data are depicted over two columns of the heatmap, representing H1 and H2. IgG is displayed on the very left column of the heatmap, serving as a negative control. The IgG track showed no signal in any of the replicates, thus this is just presented once. Still, interesting to observe is that in the IgG control there was a decline of aligned reads over the body of young FL-L1 elements, to almost fully disappearing over L1HS. This reflects the mappability issues that alignment tools experience with the repetitive nature of transposable elements. H3K4me1 showed higher overall enrichment from the activating marks but only over the flanking regions of the elements. Notably, there was a decrease in H3K4me1 signal over the body of the elements, which was not solely attributable to the lower mappability since it could be seen even in the L1PA4-7 subfamilies. H3K4me3 was observed over a small number of elements, especially over the 5'UTR of some younger L1s. Most strikingly, the presence of H3K9me3 can be found throughout the body of almost all L1 elements apart from L1Hs, likely due to low mappability of these elements (see Fig. 2). Interestingly, H3K27me3 displayed almost no enrichment throughout all regions.

It is important to note that despite the prevalence of the repressive H3K9me3, transcriptional activity was sometimes observed in each subfamily, both over the element body and the flanking regions.



Figure 4. Relative presence of the three epigenetic marks over FL-HERVs and their 5kb flanking regions. Transcriptional status obtained from bulk-RNAseq data is also shown on the heatmap.

HERVs were not further classified, thus the heatmap in Figure 4 displays the entire population of 7,036 HERVs from our database sorted by the H3K4me1 signal intensity (Figure 4). Although H3K4me3 activation was found in fewer locations, many of them represented regions of co-localization between the two activating marks. Again, H3K9me3 shows the highest enrichment over the elements, however it appears to be localized over fragments of the genomic regions rather than accumulating over the entire length like seen over L1s. Furthermore, a clear anti-correlation was observed between H3K9 and H3K4me1, with very little to no presence of the repressive mark over regions with high transcriptional potential (Figure 4). Transcriptional activity was also observed for some elements, some of which present transcription over the entire 16kb regions. These depictions of the epigenetic status are recurring throughout the project, both when L1 and HERV data is presented (Figure 4).

To define genomic regions as transcriptionally active, SEACR was used to perform differential peak calling on the activating H3K4me1/me3 marks. This peak calling method compares enrichments of the samples against their respective IgG controls, reporting back all genomic regions in which the sample shows significantly higher signal. These regions were subsequently intersected with the FL-L1s and FL-HERVs to identify TEs embedded in locations rich in activating epigenetic modifications.

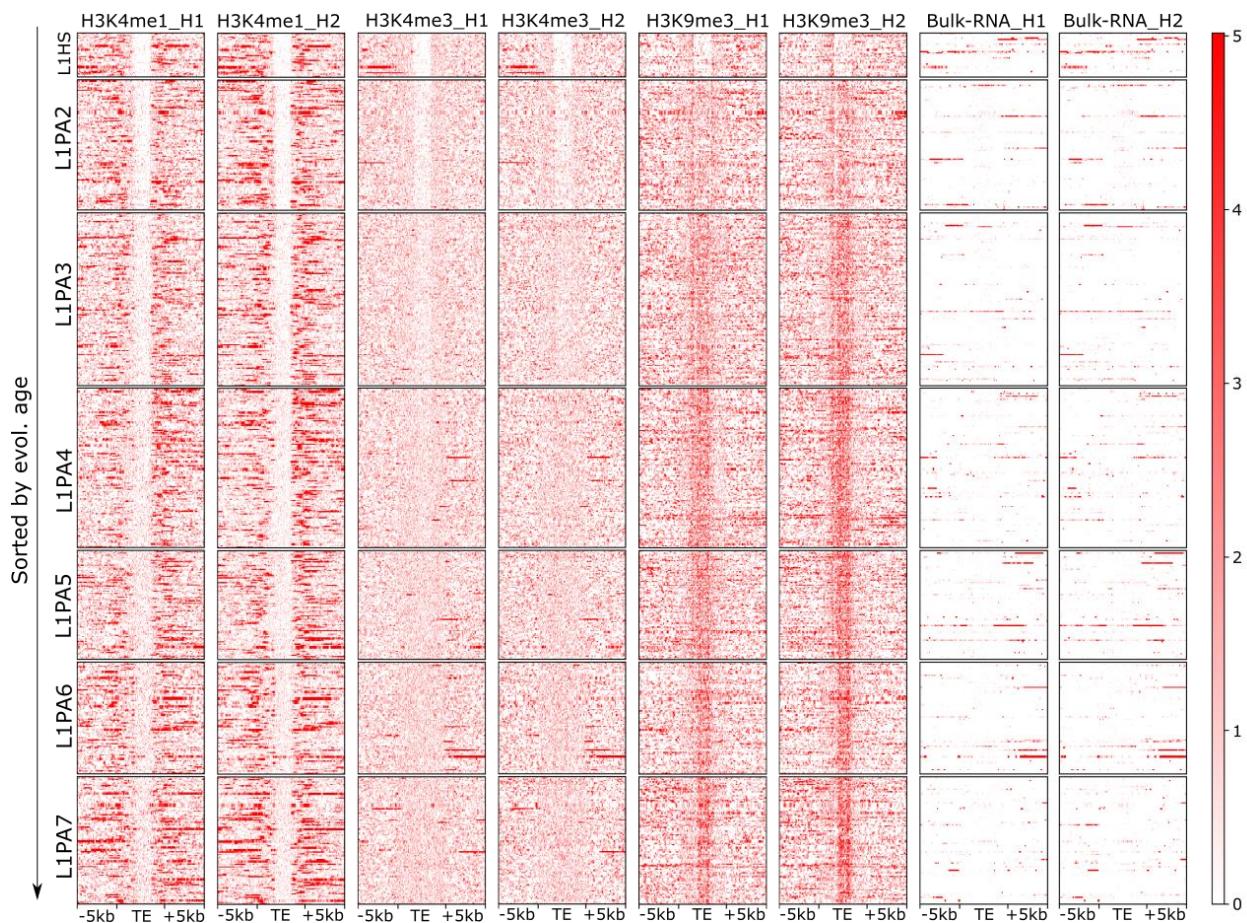


Figure 5. Epigenetic status of L1 elements found in H3K4me1 enriched regions in human NPCs. Described using relative presence of the three epigenetic marks and the transcriptional expression observed there.

A total of 697 FL-L1 elements were found within 5kb of a H3K4me1 peak. H3K4me3 again displays significantly less presence in comparison to H3K4me1, but still shows some enrichment in almost every

subfamily, including the human specific L1HS. H3K9me3 behaves very similarly as in Figure 3, covering most of the element bodies, showing transcriptional repression throughout. The Bulk-RNA confirms what the epigenetic marks depict, majority of L1s overlapping with the H3K4me1 enhancer modification are repressed by H3K9me3. However, transcriptional activity is exhibited over some the bodies of the elements, including the youngest L1HS.

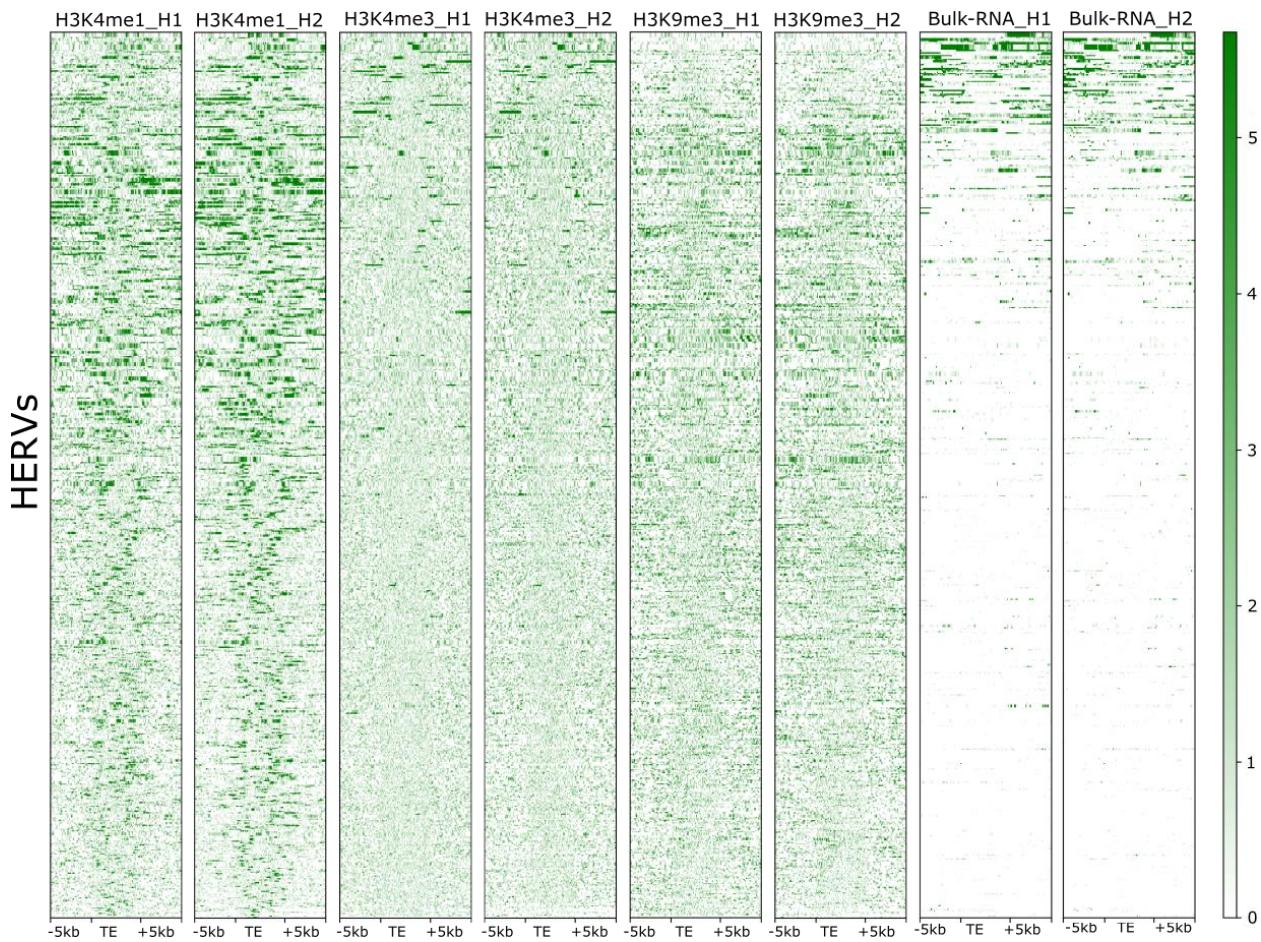


Figure 6. Epigenetic status of HERV elements found in H3K4me1 enriched regions in human NPCs. Described using relative presence of the three epigenetic marks and the transcriptional expression observed there.

686 FL-HERVs were found around H3K4me1 enriched regions. The activating H3K4me1 behaves similarly in HERVs as it does in L1s, however we could observe clear accumulation of the enhancer mark

not only over the flanking regions, but also over the element bodies. The other two epigenetic marks on the other hand, show almost identical patterns as in Figure 4: regional co-localization of H3K4me3 and shorter spanning H3K9me3 enrichments. However, it should be noticed that by sorting HERVs according to the Bulk-RNA signal intensity, as in Figure 6, transcriptional activity is exhibited by elements coinciding with both H3K4 methylation marks and seemed to be anti-correlated with H3K9me3.

Having observed the nature of H3K4me1, and to fully define the activating regions of the elements, we focused on analyzing the epigenetic status in regions enriched with H3K4me3. Despite showing enrichment at fewer elements than H3K4me1, H3K4me3 appears to be more accurate in describing activity around the TEs.

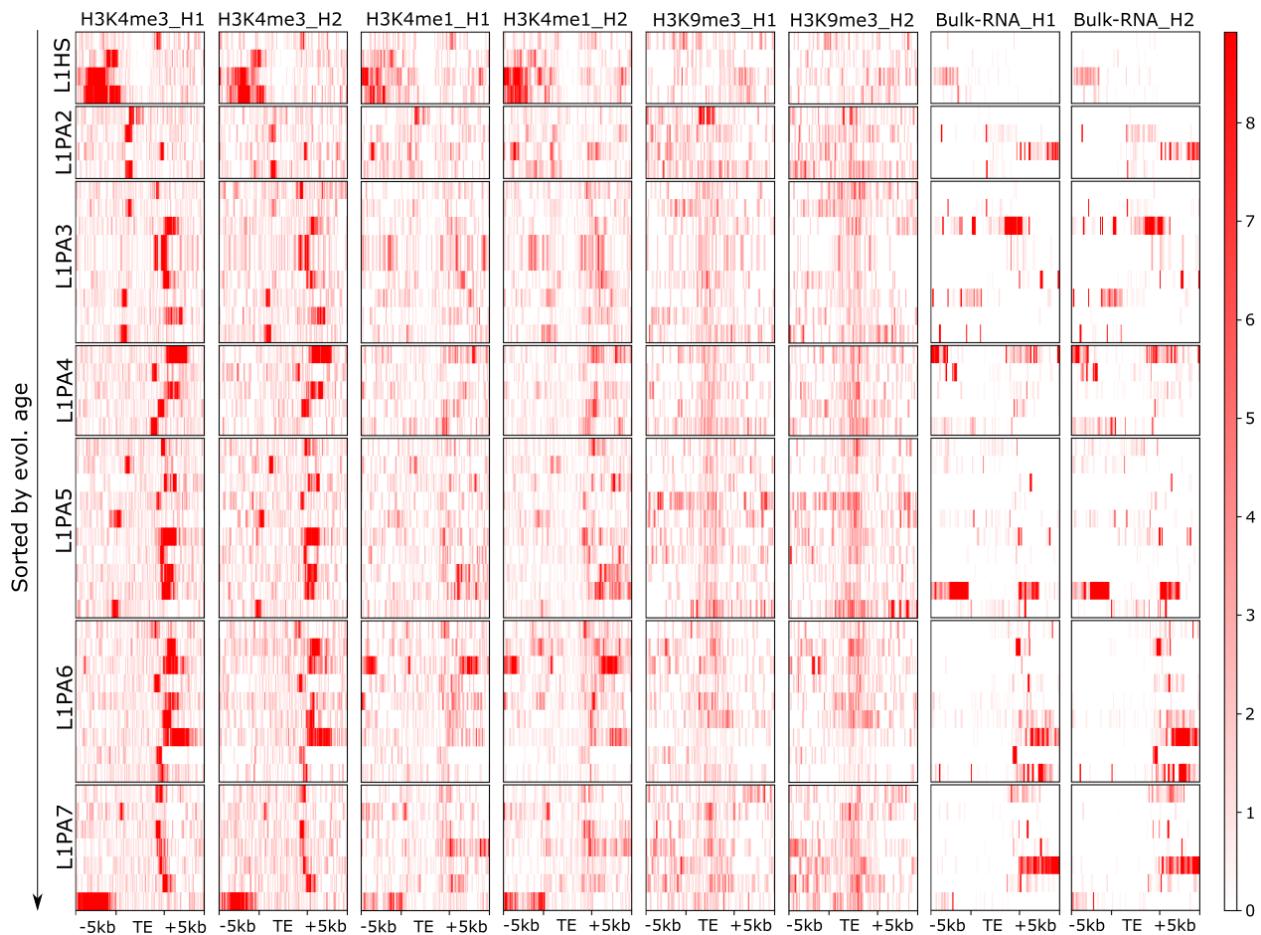


Figure 7. Epigenetic status of L1 elements found in H3K4me3 enriched regions in humans. Described using relative presence of the three epigenetic marks and the transcriptional expression observed there.

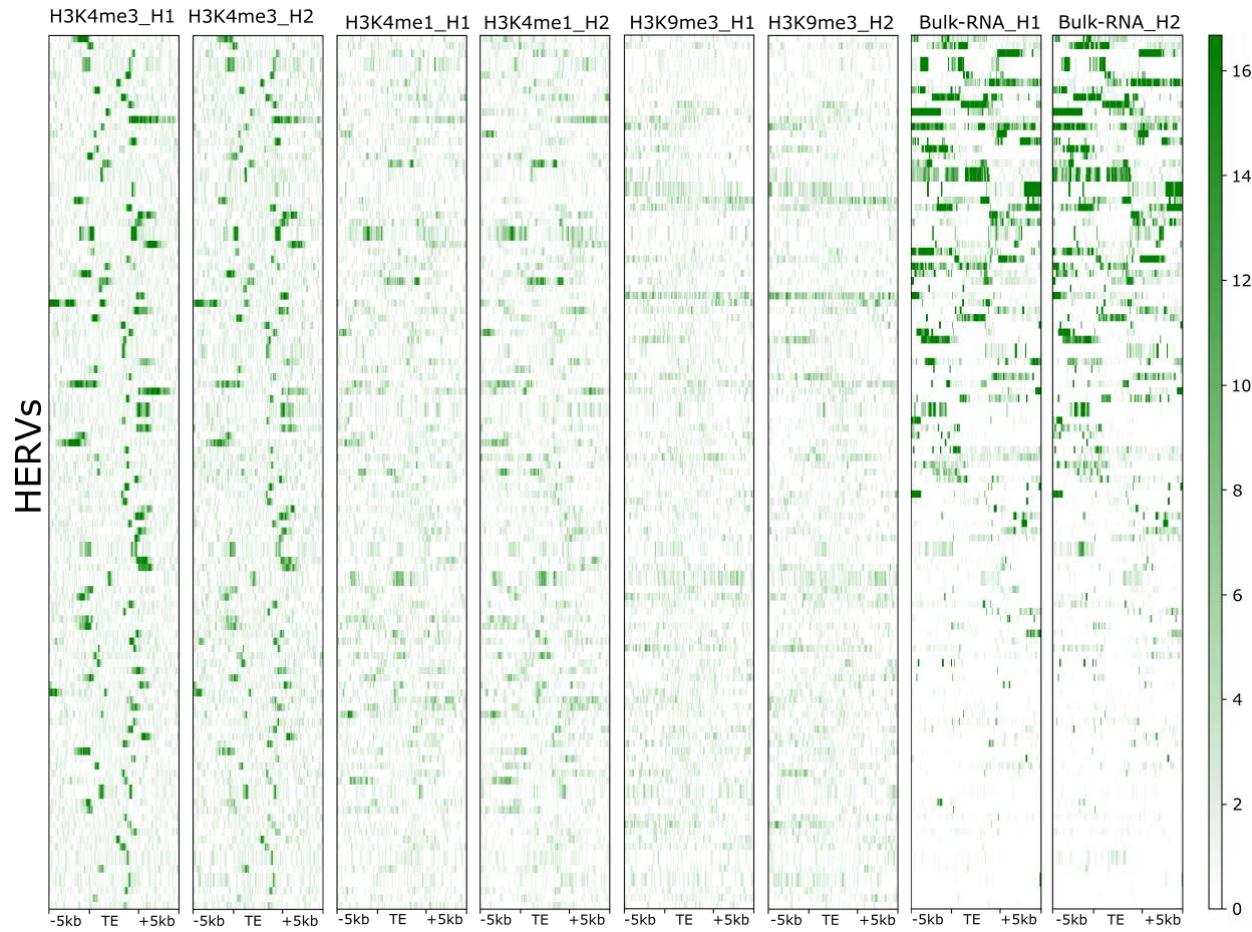


Figure 8. Epigenetic status of HERV elements found in H3K4me3 enriched regions in humans. Described using relative presence of the three epigenetic marks and the transcriptional expression observed there.

In figure 7, one can observe that there are only 48 H3K4me3-enriched LINE-1 elements, and that those peaks lie at the 5' or 3' of the element, co-localizing with H3K4me1 at 17 unique locations. The repressive H3K9me3 was present along the elements' bodies. However, H3K9me3 over the body of these elements was not always sufficient to maintain transcriptional repression: bulk-RNA sequencing corroborates observed epigenetic status by showing transcriptional activity in nearly all regions of H3K4me3 enrichment (Figure 7). Similarly, in HERVs, a co-localization of the activating marks was observed, however H3K9me3 was not found over HERV regions enriched with H3K4me3. Corroborating this was bulk-RNA sequencing, showing expression at 75 out of the 119 HERVs intersecting with H3K4me3 peaks (Figure 8).

FL-TEs intersect significantly less with H3K4me3 peaks, showing only 48 FL-L1s and 119 FL-HERVs elements around this activating mark.

Exploring the H3K9me3 peaks around FL-L1s further supported the observations made in figure 3. With 4,241 FL-L1 elements displaying significant enrichment of the repressive mark, almost 70% of all FL-L1s show accumulation of H3K9me3. The HERVs on the other hand have only 1,849 elements decorated with H3K9me3, suggesting that this repressive mark is not as important for silencing them as it is for L1s. Finally, genomic intersection between FL-TEs and genes was found to establish a list of all elements within in genes. Plotting the signal over these elements will reveal the epigenetic status over both element and the gene, helping in the assessment of the transcriptional effects one might have on the other.



Figure 9. Epigenetic status of L1 elements embedded in human genes. Described with the relative presence of the three epigenetic marks and transcriptional expression observed there.

A total of 3,316 L1s were found to intersect with human genes. Again, H3K4me1/me3 are comparable to each other when observed around genic FL-L1s, both in number of locations present and in quantity of enrichment. Heavy H3K9 repression is observed, and it appears to be contained mostly over the element body, with accumulations over the flanking regions in several locations.

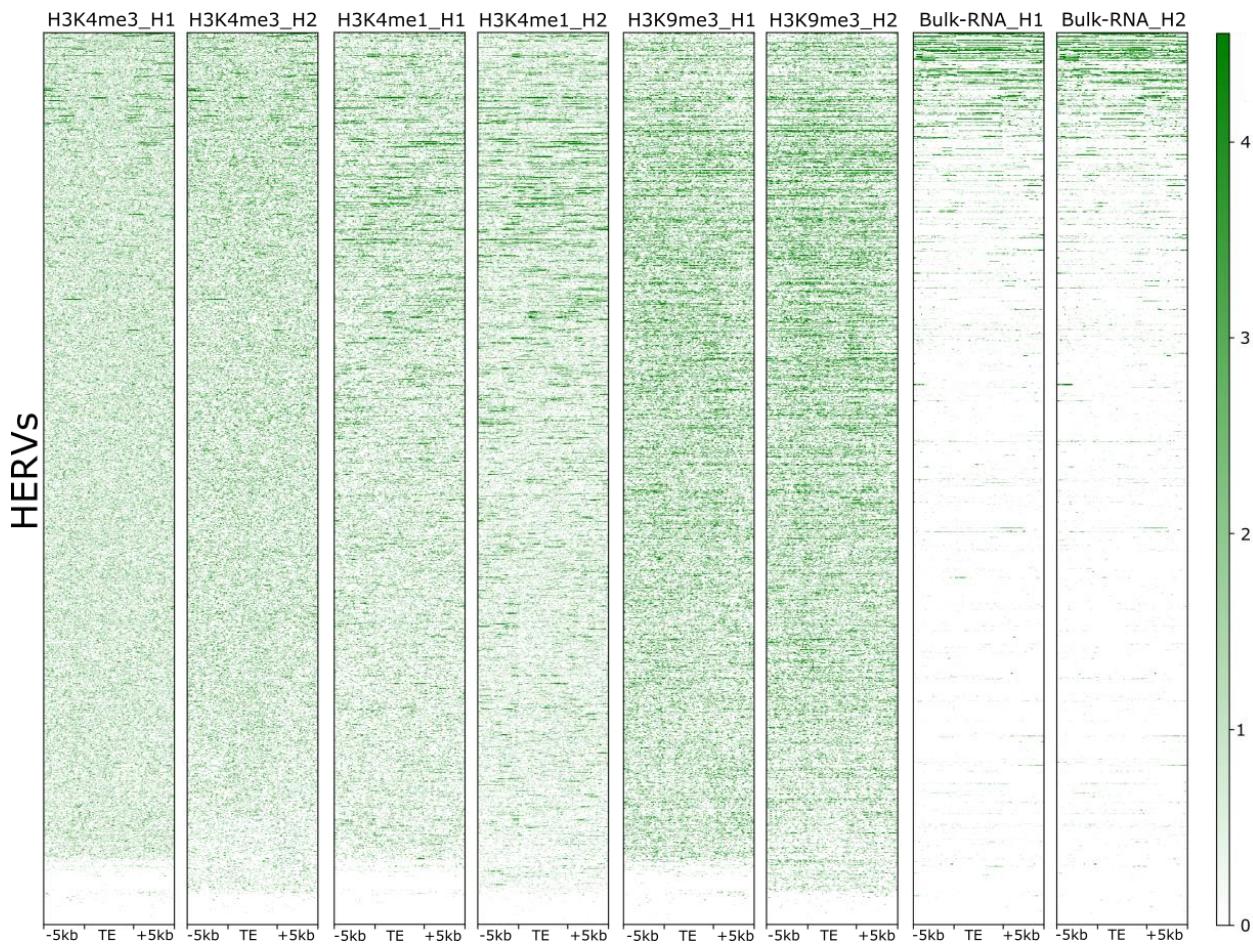


Figure 10. Epigenetic status of HERV elements embedded in human genes. Described using the relative presence of three epigenetic marks and the transcriptional expression observed there.

Genic HERVs display similar epigenetic environment to the one exhibited by HERVs around regions with H3K4me1 enrichment, shown in figure 6 (Figure 10). High enrichments of H3K4me1 were observed in nearly all genic HERVs, whereas the H3K4me3 was observed to be only scarcely co-localized with it. The

repressive H3K9me3 also displayed high enrichment, however only partially covering the co-localized regions of the activating marks. The transcriptional data supports the observation that, HERVs with enriched activating marks H3K4me1/me3 and low H3K9 enrichment show potential for transcriptional activity (Figure 10).

To further explore the environment of the seven youngest L1 elements embedded in genes, we extracted additional information about the genes they have transposed to. A list of genes was made to examine the profile of genes containing L1s, and we focused on analyzing genes with documented functions in NPCs. Additionally, to completely capture the epigenetic and transcriptional environment, the signals were visualized over entire genes, rather than just the elements. Two instances of transcriptionally active genes with functional roles are displayed in Figure 11, each containing an evolutionary young FL-L1 element. Similarly to the heatmaps, epigenetic status of these locations was described using the three epigenetic marks and the Bulk-RNA sequencing acquired at those regions. An advantage of observing the epigenome at an individual gene level, is the possibility to observe the signals' behavior over specific genomic features.

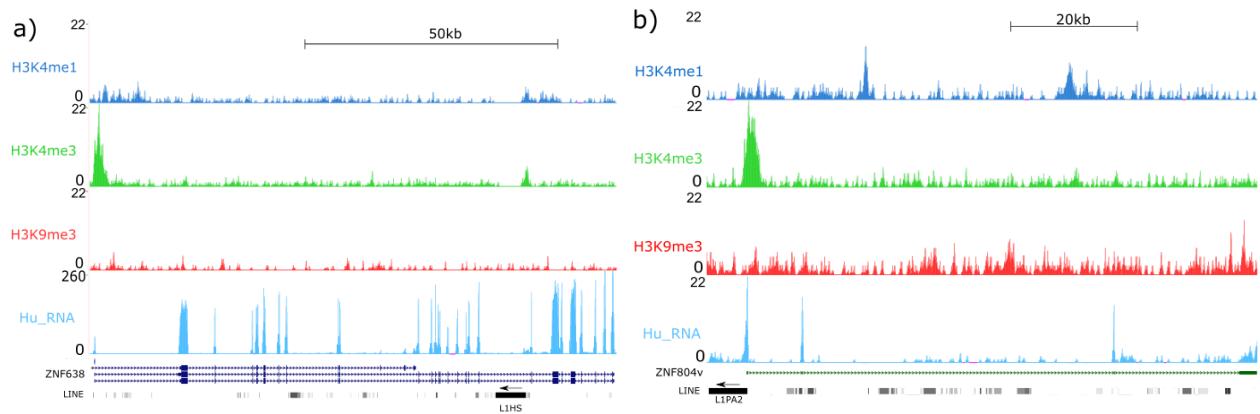


Figure 11. Epigenetic status of (a) ZNF638 gene and (b) ZNF804A variant which contain an L1HS and an L1PA2, respectively. Described using the relative presence of the epigenetic marks, IgG controls, and the transcriptional activity observed there.

Showing transcriptional activity in NPCs, ZNF638 mediates silencing of unintegrated retroviral DNA, as a

crucial protein in recruiting the HUSH complex along with other means of silencing (Figure 11a) [18]. Highly similar enrichment of both activation marks H3K4me1/me3 was observed, with a low presence throughout the entire gene except for two loci, one over the gene promoter region and the other over the 5' UTR of L1HS. In comparison, scarce signal with low presence was observed for H3K9, explaining the high bulk-RNA enrichment shown.

On the other hand, a more diverse enrichment is observed at a genetic variant of ZNF804A, which displayed regulation of schizophrenia-associated genes (Figure 11b). Namely, although enriched at several regions, the activating mark H3K4me1 fails to intersect with any genomic features. H3K4me3 however, shows significant enrichment directly over the 5' UTR of an L1PA2, and does not show any other enrichment within the gene. H3K9me3 also displays several enrichment peaks throughout the gene which appear to increase in some regions and usually not overlapping with the activating marks. It's important to observe from the bulk-RNA that H3K9me3 seems to repress the LINEs more specifically in comparison to the 3' UTR exons of the gene.

Quantitative transcriptional analysis over L1HS regions

Having defined the epigenetic patterns of FL-TEs and their potential for transcriptional expression in NPCs, we asked ourselves if the human specific FL-L1HS are able to trigger transcriptional changes around them. To reach this observation, bulk-RNA sequencing was performed on chimpanzee NPCs (n=2; Ch1 and Ch2), directly comparing them to the human NPCs already described (n=2; H1 and H2). The chimpanzee is the ideal species for this comparison because they share 99% protein coding similarity with humans, yet L1HS was shown to be specific to humans. Bulk-RNA sequencing samples from both species were uniquely mapped to the human genome assembly hg38. Gene quantification was performed using featureCounts, and the resulting count matrix was used to perform differential expression analysis (DEA) using DESeq2. We

first compare the different batches of human and chimp NPCs to test for transcriptional differences between them.

In order to answer if a FL-L1s can have cis-regulatory effects over a gene, we performed DEA on the genes that had a FL-L1 overlapping. Gene counts were normalized using the median of ratios method, calculating the ratio of each read to the geometric mean of all reads. DESeq2 searches for differentially expressed genes based on a negative binomial distribution and provides quality assessment values such as adjusted p-value (p-adj) and log2 fold change (LFC) for each read aligned. The p-values are attained by the Wald test and corrected using the Benjamini-Hochberg method. To display any underlying similarities or differences between the samples, a Principal Component Analysis (PCA) was performed on all genes as L1HS is human specific and that could produce a bias.

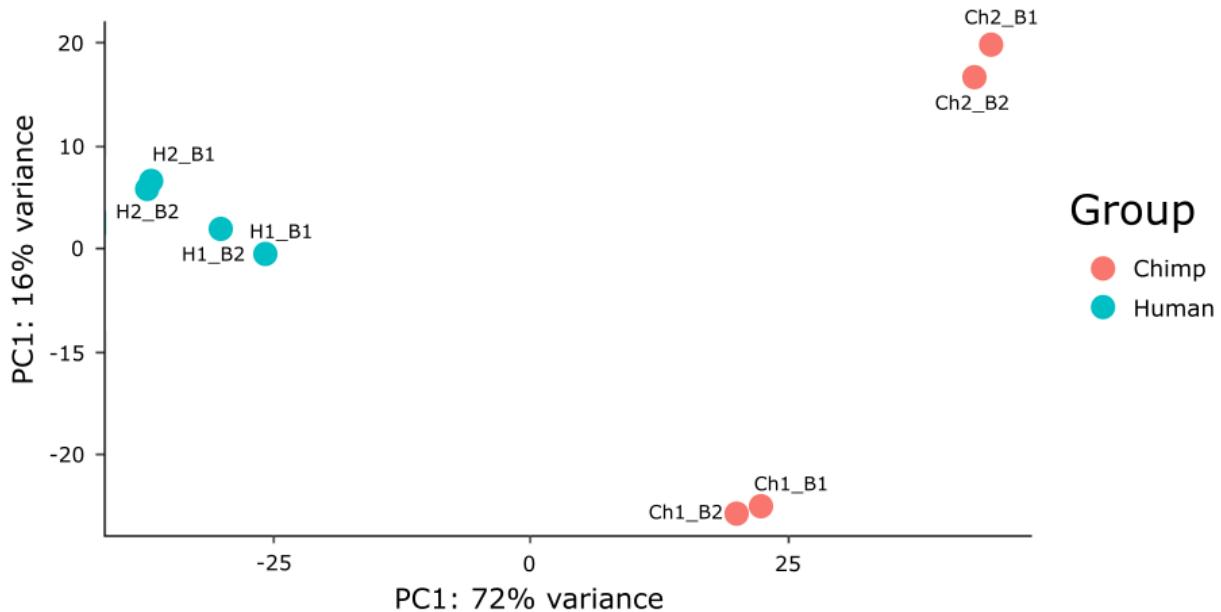


Figure 12. 2D principal component analysis of two different cell lines, each having 2 sequencing batches, from two species, Humans and Chimpanzees.

Clear grouping between all four human samples was observed on both principal components, not only when comparing the cell lines, but for the different sequencing batches as well. Unlike the human samples, chimp samples showed clustering at two different locations on the plot and appear to be better classified by the first principal component rather than the second one. Overall, the PCA plot showed that greatest divergence between the samples is observed due to species differences, rather than cell line or batch (Figure 12).

Furthermore, an MA plot was used to visualize the results from the differential expression analysis (Figure 13). This method transforms the data into a log fold change (LFC) and plots it against the mean average of the normalized counts, highlighting any genes showing significant difference between the two species. To estimate the LFC values more accurately and remove all noise when creating the MA plot, shrunken log fold values of the counts were calculated and presented (Figure 13).

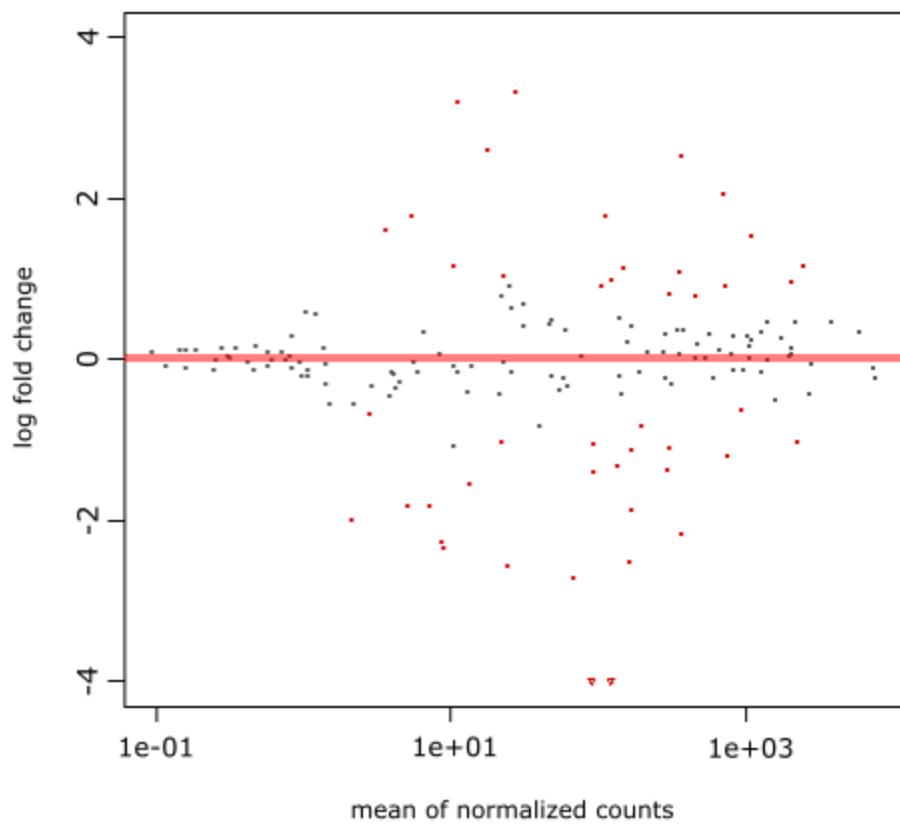


Figure 13. Mean average plot showing the mean of normalize2D principal component analysis of two different cell lines, each having 2 sequencing batches, from two species, Humans and Chimpanzees. Black points are non-significant, red points are significantly different genes ($p\text{-adj} < 0.1$).

The MA plot highlighted genes from our dataset that show significant differential expression, both in the positive and negative direction, with some of them having a high $p\text{-adj} \sim 0.1$. Thus, from the highlighted genes in Figure 13, we extracted all datapoints with $\text{LFC} > 2$ and a $p\text{-adj} < 0.05$, increasing the significance of the differential expression observed. By doing so, three groups of genes were created: non-significantly expressed, upregulated, and downregulated genes, which were further used to create expression heatmaps and a scatter plot, describing the LFC relationship between the species.

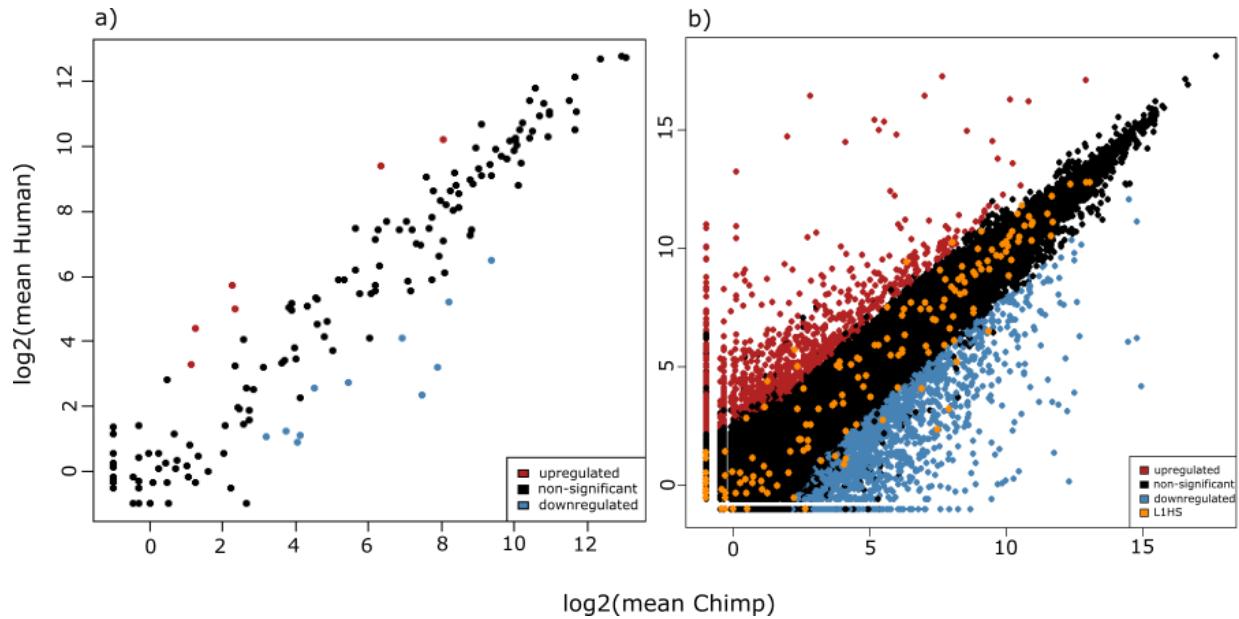


Figure 14. Scatterplot of the (a) mean human LFC values over the mean chimpanzee LFC values of genes containing L1HS elements ($p\text{-adj} < 0.1$; $\text{LFC} > 2$). (b) Mean human LFC values over the mean chimpanzee LFC values of all genes containing L1HS elements. ($p\text{-adj} < 0.1$; $\text{LFC} > 1$). Red and blue datapoints are signify upregulated and downregulated genes, respectively. Black datapoints are non-significant and orange points (only seen in 14b) are L1HS containing genes

Despite increasing the stringency on the quality of parameters set ($p\text{-adj}$ and LFC), genes that contain L1HS in their sequence still show significant differential expression. The scatterplot showed the calculated ratios of Human vs Chimp mean LFC values for every gene and displayed 11 downregulated and 6 upregulated genes, colored in blue and red, respectively (figure 14a). It is important to note that we mapped both species to the human genome, genes that show upregulation in human could prove to be false positives, thus we excluded them from further analysis.

To support our findings, we further increased the background of the DEA by performing DESeq2 over all NPC genes, rather than limiting to the ones that contain a FL-L1HS element (figure 14b). Following the DEA method already described, we observed 1,525 genes showing downregulation in human NPCs when compared to chimp NPCs. We then intersected these results with the locations of FL-L1HS to observe any differences between the differential expression analyses. Interestingly, the second DEA showed downregulation of 12 genes that contain an L1HS within their sequence, and included all genes observed

in the initial DEA (L1HS only). A table describing the number of genes showing downregulation and intersection with FL-L1HS was created (Table 1).

Table 1. Total number of downregulations and FL-L1HS intersections observed in all NPC genes.

	Downregulated	Non-significant
Intersect	11	172
Non-intersect	1,525	58,947

A total of 60,656 genes were found in human NPCs, out of which 184 intersected with FL-L1HS elements and only 12 of them show downregulation. On the other hand, 60,472 are not intersecting with FL-L1HS, from which 1,525 genes show downregulation. Using the intersection with L1 elements and downregulation as categorical variables, a Chi-square independence test was performed to check the dependency between presence of FL-L1HS and downregulation. Disproving the null hypothesis with a p-value of 0.0005, the two variables were found to show dependency between them. Finally, to visually observe the differences in expression between the species, we proceed by plotting expression heatmaps of genes downregulated in humans (Figure 15).

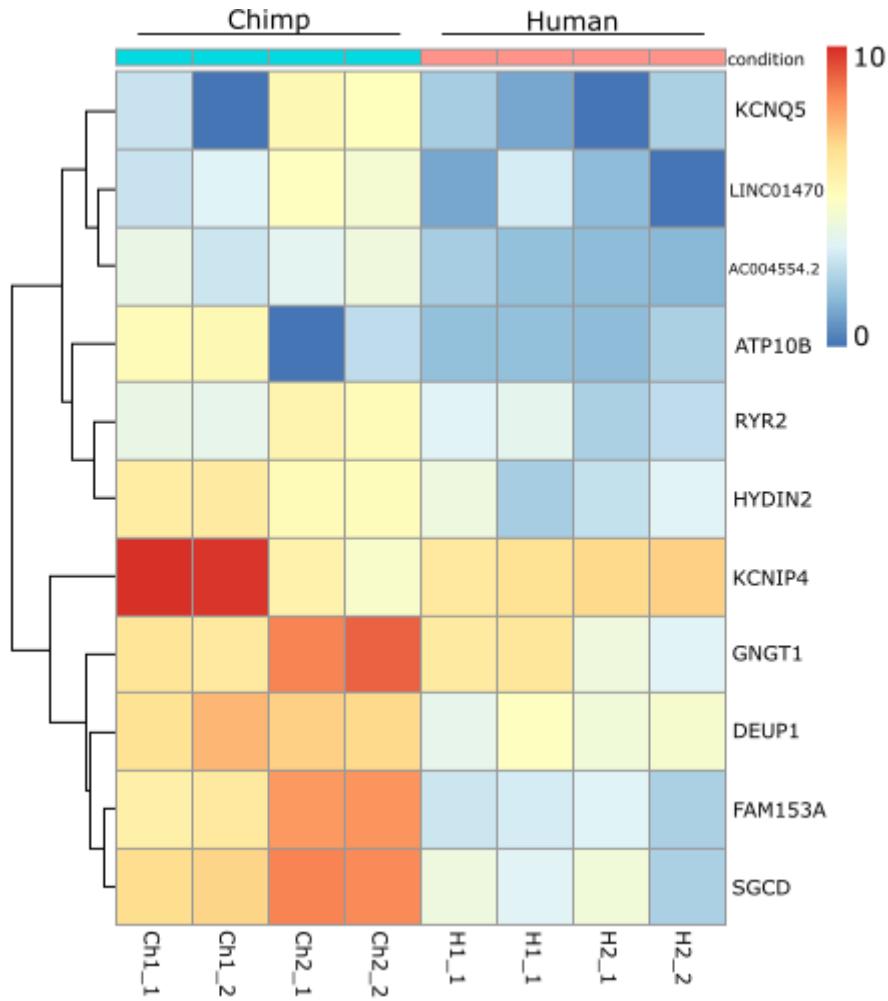


Figure 15. Expressional heatmap of genes that are transcribed significantly lower in humans in comparison to chimpanzees. Two Bulk-RNA sequencing batches of two different cell lines were used to count the expression.

Besides the obvious lower expression in the human samples, the heatmap shows that human cell lines display a more consistent expression of the genes, whereas the different chimp cell lines showed a greater variation between them. Furthermore, to investigate the molecular functions and biological processes in which these genes participate in, the downregulated list was used in a gene ontology (GO) software Panther, where we performed a statistical overrepresentation test.

Panther counts the number of genes that are related to specific GO terms and compares them to an expected value, if exceeded then they are considered to contribute to the functionality of the biological processes.

Because our analysis is performed on NPCs only, a suitable background for this test was the list of all genes that showed expression in our NPCs, rather than all genes from the hg38 genome. None of the results from the GO test had significant impact on regulatory networks.

5. Discussion

iPSCs provide an unlimited source of patient derived stem cells, which represent a cellular model for studying neurodevelopment as they can be differentiated into specific neural and glial subtypes. Genomic mosaicism that naturally occurs in the human brain has been suggested to contribute to functional diversifications within a neuronal population, thus contributing to individual differences and perhaps even triggering pathology.

Comprising almost half of our genome, TEs are one of the greatest contributors to the genetic mosaicism due to their transposability and are linked to functional regulation in brain development. In addition, the autonomous non-LTR LINE-1 elements can further contribute to genome function by providing binding sites for various regulatory elements, serving both as enhancers and repressors of transcriptional activity. LINE-1 elements show aberrant expression in various neurodegenerative disorders, supporting the idea that they are functional units in the neurons [9]. Epigenetic modifications seem to be specifically important for regulation of L1s in human NPCs. Here we defined the epigenetic status of full-length L1s and HERVs in human NPCs and we investigated the ability of the human specific L1HS to contribute to the regulation of expression.

Plotting the MAPQ of the read alignments served not only as a quality check, but also as a proof of concept for the correlation observed between the age of L1s and their mappability (figure 2). Specifically, we know that L1s accumulate neutral mutations over time which reduce the repetitive nature of elements, thus reducing the mappability problem. That is exactly why we see, a steady incline of alignment quality, starting from the youngest L1HS, to the older L1PA7 subfamily. In addition, the same pattern was observed when plotting the epigenetic marks over the FL-L1s, a decay of aligned reads over the body of younger L1s, shown even in the IgG negative control. Important to note is that because of the increased length of sequencing pairs (2x150nt) used in the Bulk-RNA samples compared with the CUT&RUN sequencing (2x75nt), the mappability issue decreases thus allowing us to partially observe signals found over the element body or at a repetitive sequence.

We observed that most young FL-L1s are covered in the repressive H3K9me3, however not by H3K27me3 (figure 3). Finally, RNA sequencing showed that some young FL-L1s evade repression, exhibiting transcripts from the element itself and adjacent regions. Similar observations were made for HERVs, however since they were not further classified into specific subfamilies, they were solely used to define the epigenetic status around them.

Narrowing down to the regions of expression, their peaks were compared, and it was established that H3K4me3 appears to find transcriptional activity of transposable elements more accurately than H3K4me1. That is because although H3K4me1 showed higher presence at more locations, most of it was repressed by H3K9me3 (figure 5), and regions omitted by the repressive mark were usually also enriched by H3K4me3 (figure 6). Oppositely, the H3K4me3 revealed locations with low repression and high transcriptional activity, both in L1s and HERVs. Finally, knowing the nature of L1 elements to affect local chromatin environment, we found that L1 and HERV elements embedded in genes showed similar epigenetic

signatures as non-genic elements, supporting the notion that these TEs could influence gene expression programs in NPCs.

Furthermore, defining the epigenetic status of TEs in NPCs is not solely exploratory as its use can be multifaceted, starting with comparison of epigenetics between human and chimpanzee, which could provide clues into how TEs have changed between the two species. This work can also be used as a template to investigate other mechanisms of TE regulation in NPCs, such as via KRAB-ZNFs, the HUSH complex or DNA methylation pathways. Interestingly, we found a human specific L1HS located in a transcriptionally active region, embedded in an intron of ZNF638, a gene important for transcriptional silencing alongside the HUSH complex [18]. Further investigation is required to investigate whether ZNF638 is involved in silencing L1s alongside HUSH, and therefore whether its intronic L1 is autoregulatory. Finally, plotting the epigenetic marks can provide the potential to identify transcriptionally active young LINE-1 elements, and further explore them.

In this project to further analyze FL-LINE1s, we tested the differential expression between human and chimpanzee over regions of genic L1HS elements, specific to the human and with the ability to cause regulatory changes. We postulated that both repression of genes due to accumulation of repressive marks brought by the element, and enhancement of genes because of the functional sites L1HS provides, may be possible. DEA was conducted to examine the differences in expression of genes overlapping L1HS elements in human and chimpanzee NPCs.

As apparently upregulated genes in humans could be false positives, we continued investigating only on the downregulated genes (i.e., with higher counts in chimpanzee samples). Namely, mapping chimp bulk-RNA sequencing to the human genome assembly represents a “disadvantage” for the chimp reads, as the target genome is not 100% identical. However, this does not apply to genes downregulated in humans, as

that would mean that human reads with an identical sequence to the reference genome align less compared to chimpanzee's samples. A solution for this is to map the bulk-RNA sequencing of both species onto the chimpanzee genome, thus confirming both upregulated and downregulated genes.

Seeing that all but one gene showed similar results between the two differential expression analyses of NPC genes served as a verification of the results observed in the DEA of L1HS genes. That is because increasing the number of genes used in the analysis increased the background, thus improving the statistical accuracy and significance of the test. The Chi-square test used these results to show co-dependency between the two variables, however further statistical tests need to be performed to reach more significant conclusions about the effects of L1HS in downregulation.

The conclusions of this project could be used to explore if the epigenetic environment is human specific or whether similar patterns are found in other organisms, such as primates. By doing so we could potentially find what causes the differences between humans and chimpanzees, since the protein coding regions share a 99% similarity. Finally, although the human specific L1HS element did not show any significant alterations in the gene ontology analysis, we can use the list of FL-L1 elements and genes found in transcriptionally active regions as potential targets for genetic manipulation, thus uncovering the effects of aberrant expression of a specific element.

Looking ahead, we see this project being a foundation for further exploring of the landscape these elements in human NPCs, serving both as database for targeting individual TEs or describing the epigenetic and transcriptional status of NPC elements.

6. References

1. Martynoga B, Drechsel D, Guillemot F. Molecular control of neurogenesis: A view from the mammalian cerebral cortex. *Cold Spring Harb Perspect Biol.* 2012;4. doi:10.1101/cshperspect.a008359
2. Human Anatomy - Kenneth S. Saladin - Google Books. [cited 25 Mar 2021]. Available: <https://books.google.se/books?id=DfESQgAACAAJ&dq=isbn:9780071222075&hl=en&sa=X&ved=2ahUKEwjDjorEpsvvAhXIAxAIHYC8CCoQ6AEwAHoECAAQAg>
3. Rubenstein JLR. Annual research review: Development of the cerebral cortex: Implications for neurodevelopmental disorders. *Journal of Child Psychology and Psychiatry and Allied Disciplines.* NIH Public Access; 2011. pp. 339–355. doi:10.1111/j.1469-7610.2010.02307.x
4. Gitler AD, Dhillon P, Shorter J. Neurodegenerative disease: Models, mechanisms, and a new hope. *DMM Disease Models and Mechanisms.* Company of Biologists Ltd; 2017. pp. 499–502. doi:10.1242/dmm.030205
5. Dolmetsch R, Geschwind DH. The human brain in a dish: The promise of iPSC-derived neurons. *Cell.* Cell Press; 2011. pp. 831–834. doi:10.1016/j.cell.2011.05.034
6. Kang S, Chen X, Gong S, Yu P, Yau S, Su Z, et al. Characteristic analyses of a neural differentiation model from iPSC-derived neuron according to morphology, physiology, and global gene expression pattern. *Sci Rep.* 2017;7. doi:10.1038/s41598-017-12452-x
7. Martínez-Cerdeño V, Noctor SC. Neural progenitor cell terminology. *Frontiers in Neuroanatomy.* Frontiers Media S.A.; 2018. doi:10.3389/fnana.2018.00104
8. Carpenter MK, Cui X, Hu ZY, Jackson J, Sherman S, Seiger Å, et al. In vitro expansion of a multipotent population of human neural progenitor cells. *Exp Neurol.* 1999;158: 265–278. doi:10.1006/exnr.1999.7098
9. Saleh A, Macia A, Muotri AR. Transposable elements, inflammation, and neurological disease. *Front Neurol.* 2019;10: 894. doi:10.3389/fneur.2019.00894
10. Gilbert C, Feschotte C. Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. *Current Opinion in Genetics and Development.* Elsevier Ltd; 2018. pp. 15–24. doi:10.1016/j.gde.2018.02.007
11. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics.* Nature Publishing Group; 2009. pp. 691–703. doi:10.1038/nrg2640
12. Bale TL, Baram TZ, Brown AS, Goldstein JM, Insel TR, McCarthy MM, et al. Early life programming and neurodevelopmental disorders. *Biological Psychiatry.* Elsevier; 2010. pp. 314–319. doi:10.1016/j.biopsych.2010.05.028
13. Muotri AR, Marchetto MCN, Coufal NG, Oefner R, Yeo G, Nakashima K, et al. L1 retrotransposition in neurons is modulated by MeCP2. *Nature.* 2010;468: 443–446. doi:10.1038/nature09544
14. Amir RE, Van Den Veyver IB, Wan M, Tran CQ, Francke U, Zoghbi HY. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl- CpG-binding protein 2. *Nat Genet.* 1999;23: 185–188. doi:10.1038/13810
15. Bedrosian TA, Quayle C, Novaresi N, Gage FH. Early life experience drives structural variation of

- neural genomes in mice. *Science* (80-). 2018;359: 1395–1399. doi:10.1126/science.aah3378
- 16. Erwin JA, Paquola ACM, Singer T, Gallina I, Novotny M, Quayle C, et al. L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nat Neurosci*. 2016;19: 1583–1591. doi:10.1038/nn.4388
 - 17. Ervony GD, Lee E, Mehta BK, Benjamini Y, Johnson RM, Cai X, et al. Cell Lineage Analysis in Human Brain Using Endogenous Retroelements. *Neuron*. 2015;85: 49–59. doi:10.1016/j.neuron.2014.12.028
 - 18. Robbez-Masson L, Tie CHC, Conde L, Tunbak H, Husovsky C, Tchasovnikarova IA, et al. The hush complex cooperates with trim28 to repress young retrotransposons and new genes. *Genome Res*. 2018;28: 836–845. doi:10.1101/gr.228171.117
 - 19. Imbeault M, Helleboid PY, Trono D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*. 2017;543: 550–554. doi:10.1038/nature21683
 - 20. Tunbak H, Enriquez-Gasca R, Tie CHC, Gould PA, Mlcochova P, Gupta RK, et al. The HUSH complex is a gatekeeper of type I interferon through epigenetic regulation of LINE-1s. [cited 3 May 2021]. doi:10.1038/s41467-020-19170-5
 - 21. Petri R, Brattås PL, Sharma Y, Jönsson ME, Pircs K, Bengzon J, et al. LINE-2 transposable elements are a source of functional human microRNAs and target sites. Feschotte C, editor. *PLOS Genet*. 2019;15: e1008036. doi:10.1371/journal.pgen.1008036
 - 22. Li W, Lee MH, Henderson L, Tyagi R, Bachani M, Steiner J, et al. Human endogenous retrovirus-K contributes to motor neuron disease. *Sci Transl Med*. 2015;7: 307ra153–307ra153. doi:10.1126/scitranslmed.aac8201
 - 23. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: From conflicts to benefits. *Nature Reviews Genetics*. Nature Publishing Group; 2017. pp. 71–86. doi:10.1038/nrg.2016.139
 - 24. Kapitonov V V., Jurka J. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature Reviews Genetics*. Nature Publishing Group; 2008. pp. 411–412. doi:10.1038/nrg2165-c1
 - 25. Pace JK, Feschotte C. The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage. *Genome Res*. 2007;17: 422–432. doi:10.1101/gr.5826307
 - 26. Boeke JD, Garfinkel DJ, Styles CA, Fink GR. Ty elements transpose through an RNA intermediate. *Cell*. 1985;40: 491–500. doi:10.1016/0092-8674(85)90197-7
 - 27. Malik HS, Burke WD, Eickbush TH. The age and evolution of non-LTR retrotransposable elements. *Mol Biol Evol*. 1999;16: 793–805. doi:10.1093/oxfordjournals.molbev.a026164
 - 28. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Morant J V., et al. Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A*. 2003;100: 5280–5285. doi:10.1073/pnas.0831042100
 - 29. Beck CR, Garcia-Perez JL, Badge RM, Moran J V. LINE-1 elements in structural variation and disease. *Annual Review of Genomics and Human Genetics*. Annual Reviews ; 2011. pp. 187–215. doi:10.1146/annurev-genom-082509-141802
 - 30. Konkel MK, Walker JA, Batzer MA. LINEs and SINEs of primate evolution. *Evol Anthropol*. 2010;19: 236–249. doi:10.1002/evan.20283

31. Sookdeo A, Hepp CM, Boissinot S. Contrasted patterns of evolution of the LINE-1 retrotransposon in perissodactyls: The history of a LINE-1 extinction. *Mob DNA*. 2018;9: 1–15. doi:10.1186/s13100-018-0117-4
32. Beck CR, Collier P, Macfarlane C, Malig M, Kidd JM, Eichler EE, et al. LINE-1 retrotransposition activity in human genomes. *Cell*. 2010;141: 1159–1170. doi:10.1016/j.cell.2010.05.021
33. Lavie L, Maldener E, Brouha B, Meese EU, Mayer J. The human L1 promoter: Variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. *Genome Res*. 2004;14: 2253–2260. doi:10.1101/gr.2745804
34. Associates S, Lynch M. THE ORIGINS OF GENOME ARCHITECTURE The Origin of Eukaryotes 1 Entry into the DNA World 4 A viral origin of DNA? 6 Membranes early or late? 7 The Stem Eukaryote 16 The Eukaryotic Radiation 18 Genome Repatterning and the Eukaryotic Radiation 23 A Synopsis of the First 2 Billion Years of Biology 26. 2007. Available: <https://codn.pw/678.pdf>
35. Martin C, Zhang Y. The diverse functions of histone lysine methylation. *Nature Reviews Molecular Cell Biology*. Nature Publishing Group; 2005. pp. 838–849. doi:10.1038/nrm1761
36. Hwang JY, Aromalaran KA, Zukin RS. The emerging field of epigenetics in neurodegeneration and neuroprotection. *Nature Reviews Neuroscience*. Nature Publishing Group; 2017. pp. 347–361. doi:10.1038/nrn.2017.46
37. Podobinska M, Szablowska-Gadomska I, Augustyniak J, Sandvig I, Sandvig A, Buzanska L. Epigenetic modulation of stem cells in neurodevelopment: The role of methylation and acetylation. *Frontiers in Cellular Neuroscience*. Frontiers Media S.A.; 2017. doi:10.3389/fncel.2017.00023
38. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*. Nature Publishing Group; 2007. pp. 272–285. doi:10.1038/nrg2072
39. Jönsson ME, Ludvik Brattås P, Gustafsson C, Petri R, Yudovich D, Pircs K, et al. Activation of neuronal genes via LINE-1 elements upon global DNA demethylation in human neural progenitors. *Nat Commun*. 2019;10: 1–11. doi:10.1038/s41467-019-11150-8
40. Gräff J, Mansuy IM. Epigenetic dysregulation in cognitive disorders. *European Journal of Neuroscience*. John Wiley & Sons, Ltd; 2009. pp. 1–8. doi:10.1111/j.1460-9568.2009.06787.x
41. Grassi DA, Brattås PL, Valdés JG, Rezeli M, Jönsson ME, Nolbrant S, et al. Post-transcriptional mechanisms distinguish human and chimp forebrain progenitor cells. *bioRxiv*. bioRxiv; 2019. p. 582197. doi:10.1101/582197
42. Skene PJ, Henikoff S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife*. 2017;6. doi:10.7554/eLife.21856