LINE-2 transposable elements shape post-transcriptional gene regulation in the human brain

Rebecca Petri[1], Per Ludvik Brattås[1], Marie E Jönsson[1], Karolina Pircs[1], Johan Bengzon[2] & Johan Jakobsson[1*]

[1] *Laboratory of Molecular Neurogenetics, Department of Experimental Medical Science, Wallenberg Neuroscience Center and Lund Stem Cell Center, BMC A11, Lund University, 221 84 Lund, Sweden.*

[2] *Laboratory of Glioma Cell Therapy, Lund Stem Cell Center, BMC B10, Lund University, 221 84, Lund, Sweden and Department of Clinical Sciences, Division of Neurosurgery, Lund University, 221 00, Lund, Sweden.*

[*]Correspondence:
Johan Jakobsson
Dept of Experimental Medical Science
Wallenberg Neuroscience Center
BMC A11
221 84
Lund
SWEDEN

Email: johan.jakobsson@med.lu.se
Phone: +46 46 2224225
Fax: +46 46 2220559

Abstract

Transposable elements (TEs) are dynamically expressed at high levels in multiple human tissues including the brain, but the function of TE-derived transcripts remains largely unknown. In this study we identify numerous miRNAs that are derived from TEs and expressed in the human brain by conducting AGO2-RIP, followed by small RNA sequencing on human brain tissue. Many of these miRNAs originated from L2 elements, which entered the human genome around 100-300 million years ago. We found that L2-miRNAs derive from the 3' end of the L2 consensus sequence and that they share very similar sequences, indicating that they could target transcripts with L2s in their 3'UTR. In line with this, we found that many protein-coding genes expressed in the brain carry fragments of L2-derived sequences in the 3'UTR, which serve as target sites for L2-derived miRNAs. Our findings uncover a TE-based post-transcriptional network that shapes transcriptional regulation in the human brain.

Introduction

The emergence and evolution of gene regulatory networks is thought to underlie biological adaptations and speciation. Transposable elements (TEs) have recently been implicated in these processes since they can amplify in numbers and move into new regions of the genome. Genomic analyses supports a role for TEs in gene regulatory networks, since a substantial fraction of TEs evolve under selective constraints despite being non-coding (Chuong et al. 2017). Still, the impact of TEs on human transcriptional networks remains poorly understood.

We have recently described that many transcripts expressed during human brain development contain TE-derived sequences (Brattas et al. 2017). These sequences appear to be indirectly transcribed, often in antisense direction, as part of other transcripts including those coding for protein. The function of TEs expressed in the human brain is unknown, but an interesting possibility is that they act as templates for RNA-binding proteins hereby contributing to post-transcriptional regulation.

In this study, we found, using Argonaute – RNA Immunoprecipitation (AGO-RIP) on adult human brain tissue, that many small RNAs bound by AGO2 are derived from TEs, with enrichment for LINE-2 (L2) elements. These L2-derived microRNAs (miRNAs) show strong sequence complementarity to L2-elements found in the 3'UTR of protein coding genes. Transcripts containing L2-elements in the 3'UTR are incorporated into the RNA induced Silencing complex (RISC) in the human brain and are regulated by L2-derived miRNAs. Together our results demonstrate a TE-based post-transcriptional network that influences the expression of protein-coding genes in the human brain.

Results

Identification of transposable element-derived miRNAs expressed in the human brain

To identify small RNAs that participate in gene silencing in the human brain we performed Argonaute2-RNA interacting Immunoprecipitation followed by small RNA sequencing (AGO2-RIP-seq) on surgical biopsies obtained from either human cortex (n = 3) or from human glioblastoma (n = 6) (Fig 1A). The use of AGO2-RIP-seq on fresh human brain tissue circumvents several challenges associated with detecting functional small RNAs derived from transposable elements (TEs) since it reduces background noise generated by degradation products as well as avoids problems arising with the use of cell lines, where a loss of DNA methylation could activate aberrant TE expression.

We found that AGO2-bound small RNAs displayed a high enrichment for 20-24 nucleotide (nt) long reads, the typical size of microRNAs (miRNAs), while input samples, which include all small RNAs in the tissue, displayed an expected broad size profile of RNAs including e.g. many RNAs in the size range of 30-36 nt (Fig 1B). We found similar results in both cortex and glioblastoma, although there was a trend for an increased number of RNAs in the size range of 30-36 nt in glioblastoma samples (Fig 1B & Suppl. Fig 1A).

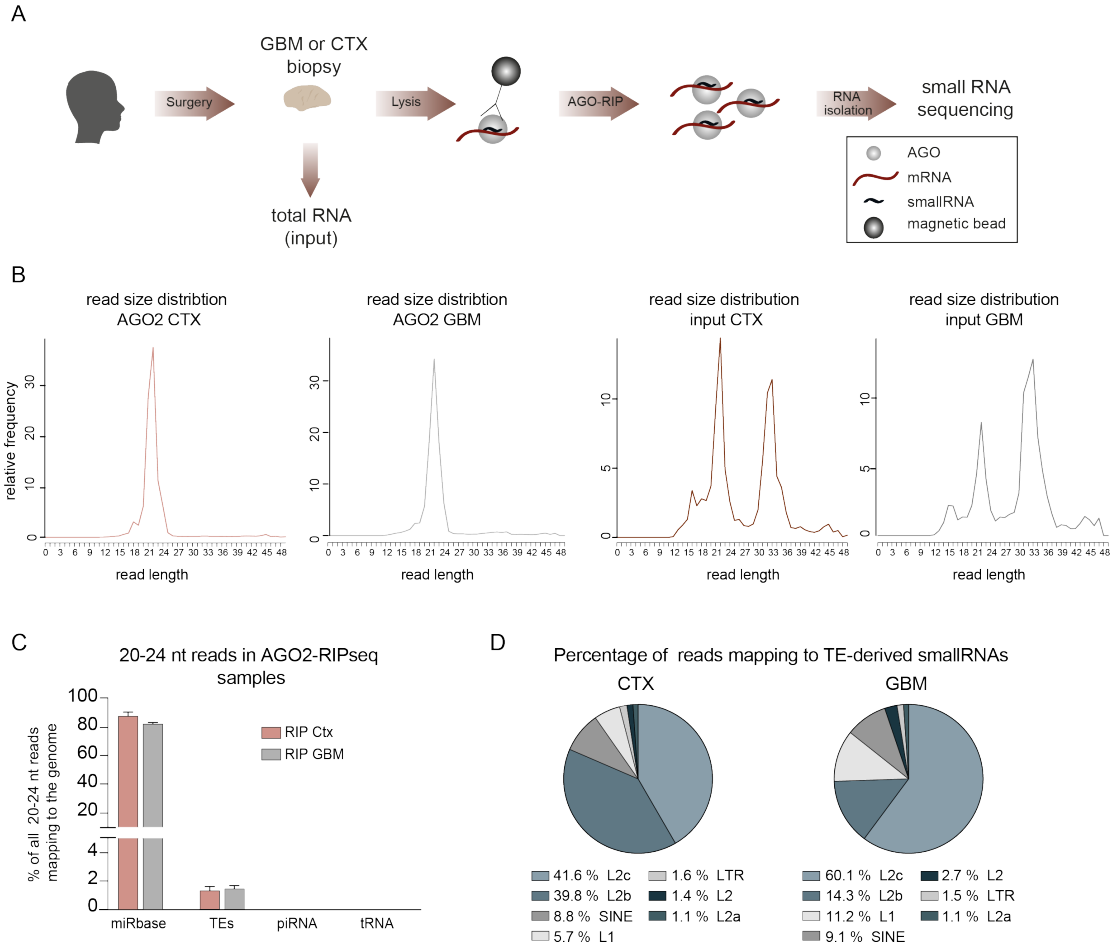

Figure 1: AGO2-associated small RNAs in the human brain.
A) Schematics of AGO-RIP-seq on human glioblastoma or cortex samples followed by small RNA sequencing. B) Read size distribution of cortex and glioblastoma RIP and input samples. C) Bar graph showing the percentage of 20-24 nucleotide long reads in the human genome mapping to miRNAs, transposons, piRNAs and tRNAs. Data is represented as mean ± SEM (RIP Ctx n = 3; RIP GBM n = 6). D) Pie charts showing the distribution of reads mapping to transposable elements. AGO – Argonaute, CTX – cortex, GBM – glioblastoma, TE – transposable elments.

We next investigated the genomic origin of small RNAs expressed in the human brain, and found, as expected, that most AGO2-bound RNAs were classical miRNAs (Suppl. Fig 1A). We found very limited evidence for binding of transfer RNAs (tRNAs) and piwi-RNAs (piRNA) to AGO2, although tRNA-fragments, of mostly 30-36 nt in size, were abundant in input samples (Suppl. Fig 1A).

We next focused our analyses on AGO2-bound small RNAs of 20-24 nt in size. We detected high expression of classic brain-enriched miRNAs, such as miR-128 and miR-124, in cortex samples (Akerblom et al. 2012; Tan et al. 2013). We also discovered altered expression of e.g. miR-21 and miR-10b in the glioblastoma samples, which is in line with previous studies (Suppl. Fig. 1B) (Karsy et al. 2012). Interestingly, we found a substantial fraction (around 2-3 % of reads) of AGO2-bound RNAs that mapped to TEs, including LINE, LTR and SINE elements (Fig 1C). To investigate if these results were unique for AGO2 or also applied to other Argonaute family members we performed the same RIP-seq approach using an antibody against AGO1, which revealed that AGO2 and AGO1 loading is very similar, not only when it comes to miRNAs, but also other small RNAs including TE-derived sequences (Suppl. Fig 1C-E).

L2-derived miRNAs are abundantly expressed in the human brain

We next set stringent criteria to identify miRNA-like small RNAs that are derived from TEs (see methods for details). We analysed the genomic position of individual TEs and detected several TE-derived small RNAs that are likely to participate in gene silencing. The majority of the small RNAs we identified derived from L2 elements, with enrichment for L2c in both cortex and glioblastoma samples. Interestingly, we saw a much higher proportion of reads mapping to L2b in cortex than in glioblastoma samples, which was explained by the high expression of a single L2b-derived small RNA in the cortical samples (Fig 1D & Table 1).

TE	coordinates	direction	miRNA	paired	mean expression (reads in RIP)
L2b	chr 4: 8005199 - 8005343 (+)	AS	miR-95-3p	L2c	2626.71
L2c	chr 8: 140732529 - 140732650 (-)	S	miR-151a-3p/5p	L2c	2771.44
L2c	chr 8: 140732622 - 140732734 (+)	AS	miR-151a-5p	L2c	4.75
L2c	chr 3: 188688700 - 188688813 (-)	AS	miR-28-5p	L2c	652.61
L2c	chr 3: 188688783 - 188688877 (+)	S	miR-28-3p/-5p	L2c	313.10
L2c	chr 11: 79402022 - 79402117 (+)	AS	miR-708-3p/-5p	L2c	96.41
L2c	chr 11: 79401986 - 79402051 (-)	S	miR-708-3p	L2c	142.75
L2c	chr 22: 45200672 - 45200995 (-)	S	miR-1249-3p	-	179.36
L2c	chr X: 74287157 - 74287324 (-)	S	miR-374a-3p/ miR-545-5p	-	41.12
L2c	chr X: 74218419 - 74218583 (-)	S	miR-374b-3p/ miR-374c-5p/ miR-421 5p	-	5.89
L2b	chr14: 100869090 - 100869150 (-)	S	miR-493-3p	-	40.53
L2b	chr 5: 176367379 - 176367999 (+)	S	miR-1271-5p	L2a	16.44
L2c	chr 22: 35335672 - 35335822 (-)	AS	miR-3909-3p/ -5p	-	31.77
L2a	chr 5: 15934793 - 15935362 (-)	AS	miR-887-3p/5p	-	20.74

Table 1: L2-derived miRNAs in the human genome.

When comparing the L2-derived small RNAs with miRbase annotations, we found that they have previously been identified as miRNAs including e.g. L2c-derived miR-151 and L2b-derived miR-95 (Table 1) (Smalheiser and Torvik 2005; Piriyapongsa et al. 2007). Many L2-derived miRNAs that we identified in the human brain originate from two L2 elements that are overlapping in the genome but oriented into opposite direction, thereby providing a source of hairpin structures (Fig 2A, B) (Smalheiser and Torvik 2005). However, we also found cases where one single L2 element gives rise to two different miRNAs e.g. mir-545 and mir-374 (Fig 2B). We found similar cases of L2-derived miRNAs in AGO1-RIPseq samples from human brain tissue (Suppl. Fig 2A) and in AGO2-RIPseq samples from pure neuronal cultures derived from human embryonic stem cell (hESCs) (Suppl. Fig 2B). We moreover conducted AGO2-RIPseq on mouse striatum to investigate if L2-miRNAs are conserved among

mammals. We found several conserved L2-derived miRNAs, including e.g. miR-151, but also examples of L2-derived miRNAs present in the human but not mouse genome, e.g. miR-95. Additionally, we identified miRNAs with the same mature sequence in the human and mouse genome, however, originating from different unique retrotransposition events such as miR-28 (Suppl. Fig 2C). This shows that L2-derived miRNAs are bound by both AGO1 and AGO2 and expressed in the brain of different mammalian species, suggesting a conserved functional role for these non-coding RNAs.

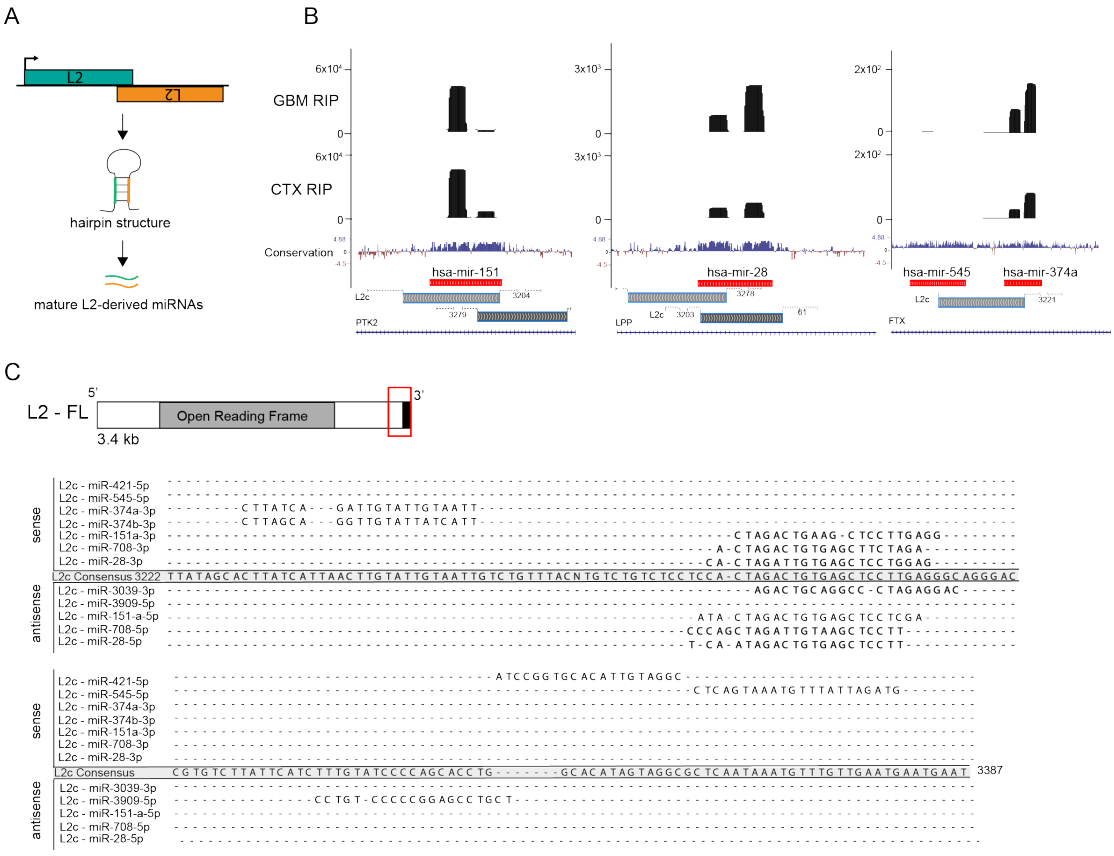

Figure 2: miRNAs derive from L2 and share similar sequences.
A) Schematics of the generation of miRNAs from L2 elements. B) UCSC genome browser tracks showing examples of L2-derived miRNAs. C) Schematics of a full-length (FL) L2 element (red box indicates the 3' end) and alignment of L2c-derived miRNAs to the 3' end (position 3222 – 3387) of the L2 consensus sequence.

To investigate the similarity between different L2-derived miRNAs, we mapped all the L2c-derived small RNAs to the L2 consensus sequence (RepeatMasker). This analysis showed that most L2c-derived miRNAs are generated from the exact same position within the consensus sequence of the 3' end of L2 elements, even though these elements are scattered throughout the genome. Thus, L2c-derived miRNAs are therefore very similar in sequence and are likely to share targets (Fig 2C).

L2-derived fragments are found in the 3'UTR of protein coding genes

L2-derived AGO2-associated miRNAs have a large number of potential "self-targets" since these elements extensively colonized the genome of our ancestors around 100-300 million years ago, resulting in almost 500,000 L2-derived elements including both fragments and more complete sequences (Vladimir V. Kapitonov 2006). Thus, L2-derived miRNAs could guide the RISC to 3'UTRs containing these elements transcribed in the opposite direction of their element of origin, herby providing a possibility for TEs to post-transcriptionally shape the expression of numerous protein-coding genes. To investigate this possibility, we analysed the location of L2 elements in the human genome. We found that only a small fraction, 2847 out of 471,716, of human L2-elements are found in 3'UTRs (Fig 3A). However, when analysing RNA sequencing data from human neural progenitor cells (hNPCs) we found that more than 40 % of L2 reads are mapping to L2 elements located in the 3'UTRs of genes. This percentage was even higher for L2b (48 %) and L2c (51 %), demonstrating that L2 elements located in 3'UTR of genes are preferentially expressed in human neural cells (Fig 3B).

We next analysed the structural conservation of L2c elements in the human genome and found that while the 3' end of the L2c-consensus sequence has primarily been maintained, most of the upstream sequences of L2c were lost during evolution. This enrichment of the 3'-part was even more apparent when we specifically looked at L2c elements located in the 3'UTR of genes (Fig 3C). Interestingly, the conserved 3'-part of the L2c elements was identical to the L2c-region from where the miRNAs are generated (Fig 3C).

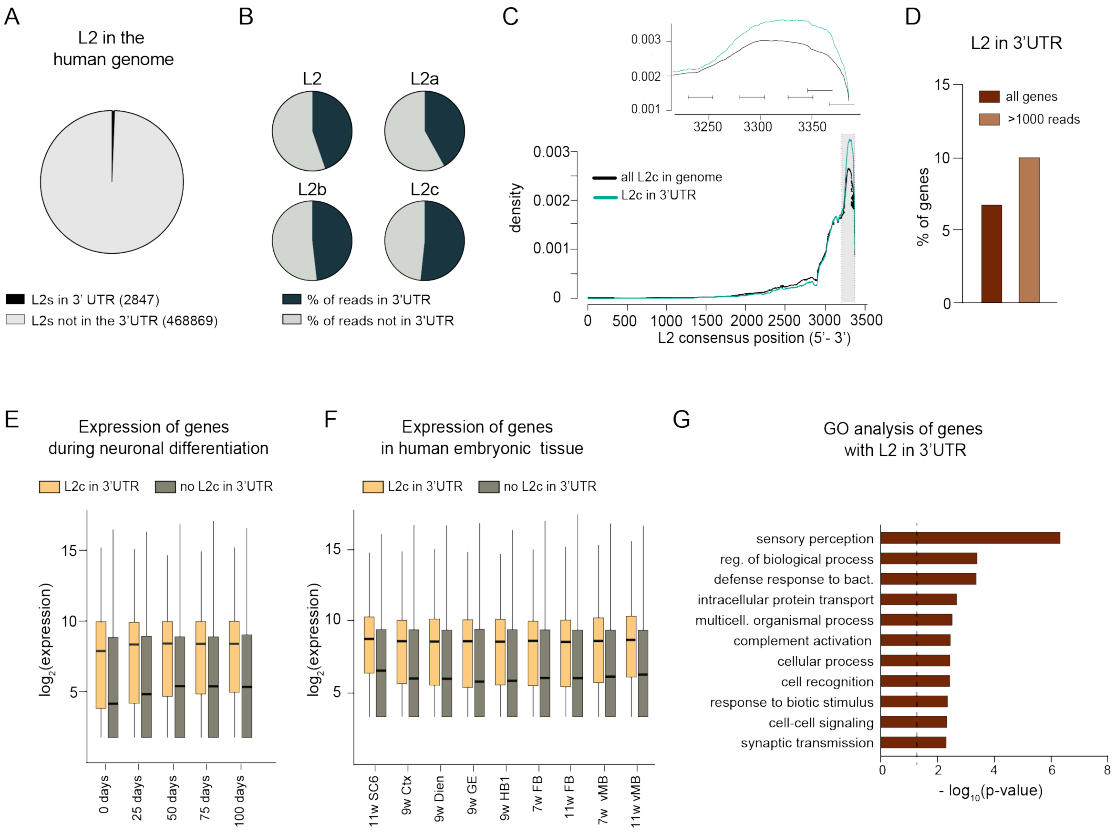

Figure 3: Genes carrying L2 in their 3'UTR are expressed in hNPCs, during neuronal differentiation and human embryonic tissue.
A) Pie chart showing the proportion of L2 within 3'UTRs out of all L2 elements in the human genome. B) Pie charts showing the percentage of reads mapping to L2 / L2a / L2b / L2c that are located either within (dark blue) or outside (grey) of 3'UTRs of genes. C) Graph showing the density of the L2c consensus sequence of all L2c in the genome (black line) and of L2c element in the 3'UTR (blue line). D) Bar graph showing the percentage of genes with L2 in their 3'UTR among all expressed genes and genes with reads above 1000 in hNPCs. E) Box plots showing the expression of

genes (log_2 transformed reads – Variance stabilising transformation) with or without L2c in their 3'UTR in human embryonic tissue. F) Box plots showing the expression of genes (log_2 transformed reads – Variance stabilising transformation) with or without L2c in their 3'UTR during human neuronal differentiation G) GO-analyses of genes carrying L2 in their 3'UTR. The red line indicates a p-value of 0.05.
SC – spinal cord; Dien – diencephalon: GE – ganglionic eminences; HB – hindbrain: FB – forebrain; vMB – ventral midbrain.

We next investigated the expression of genes carrying L2-elements in the 3'UTR in hNPCs. We identified 2042 such expressed genes, which is about 7 % of all genes expressed in hNPCs (Fig 3D). The fraction of genes carrying L2 in their 3'UTR was even higher, around 10 %, when looking at highly expressed genes (above 1000 reads) (Fig 3D). We also found that genes carrying L2c in their 3'UTR are abundantly expressed during neuronal differentiation, as well as in fetal human brain tissue (Fig 3 E-F). This analysis suggests that L2-derived miRNAs have the potential to target numerous L2-derived sequences in the 3'UTR of genes, transcribed in the opposite direction, and hereby regulate hundreds of protein-coding transcripts in human neural cells. Gene ontology analysis of expressed L2-containing genes revealed enrichment for terms such as sensory perception, protein transport, cell signalling and synaptic transmission (Fig 3G).

L2-derived fragments in the 3'UTR of protein coding genes are L2-miRNA-targets

To provide functional evidence that L2- containing 3'UTRs are regulated by miRNAs in the human brain, we conducted AGO2-RIP on both cortex (n = 2) and glioblastoma (n = 5) tissue followed by total RNA sequencing to identify miRNA target genes (Fig 4A). As expected, most reads in the RIP samples were mapping to RefSeq annotations (Suppl. Fig 3), while in input samples ribosomal RNA (rRNA) was highly abundant, the amount of rRNA was, as expected, strongly decreased in the RIP fraction. We

found an enrichment of reads mapping to 3'UTRs of genes in the RIP samples compared to input fractions, which is line with the high prevalence of miRNA target sites in this part of a transcript. Long-non coding RNAs (lncRNAs) were also slightly enriched in the RIP samples (Suppl. Fig 3), which is in line with previous studies, showing that some lncRNAs are bound by AGO2 (Weinmann et al. 2009; Imig et al. 2015). Strikingly we also found that reads mapping to repeats including LINEs, SINEs and LTRs, including many L2-transcripts, were enriched in RIP-samples.

We next analysed reads mapping to L2 elements in detail. We found that a substantial fraction of L2 reads were originating from the 3'UTR of genes and this fraction was higher in RIP samples compared to input samples, showing that transcripts with L2 in their 3'UTR are bound by AGO2 (Fig 4B). Moreover, genes carrying L2 in their 3'UTR were enriched in RIP compared to input samples and many were highly expressed (Fig 4C).

We found that approximately 2 % of all transcripts detected in RIP samples carry L2 in their 3'UTR, both in cortex and glioblastoma samples (Fig 4D). To investigate the abundance of L2-carrying genes bound to AGO2, we analysed genes that belong to the 100 most abundant transcripts in the RIP samples and which were also more than 4-fold enriched in RIP compared to the input fractions (hereafter referred to as AGO-bound genes). Strikingly, we found that among the AGO-bound genes, L2 carrying transcripts were highly enriched and made up 15 % of the transcripts. Although the overall numbers of L2-carrying transcripts were very similar in glioblastoma and cortex samples, we saw profound differences when analysing L2b and L2c-carrying transcripts separately (Fig 4E&F). L2b-carrying transcripts were more abundant in

cortex samples compared to glioblastoma tissue, where no L2b-carrying genes were AGO-bound, which is in line with the small RNA data, where the L2b-derived miR-95 was highly expressed in samples from cortex, but not from glioblastoma (Fig 4E). However, genes carrying L2c in their 3'UTR were enriched in the RIP samples from both sample types (Fig. 4F).

Together this analysis shows that genes carrying L2 in their 3'UTR are highly abundant in RIP fractions of glioblastoma and cortex samples suggesting that they are miRNA target genes. When we looked for potential miRNA target sites within the L2-elements in the 3'UTRs of highly abundant AGO-bound genes, we found several potential non-canonical target sites with high complementary to the L2-derived miRNAs: miR-28, miR-95, miR-151a or miR-708 (Figure 4G). To validate the functionality of these non-canonical target sites, we performed luciferase assays and confirmed that all tested target sites are regulated by L2-derived miRNAs (Figure 4H). Taken together, these data demonstrate that L2-derived miRNAs regulate numerous target genes carrying L2-derived sequences in their 3'UTR resulting in a TE-based post-transcriptional gene regulatory network (Figure 4I).

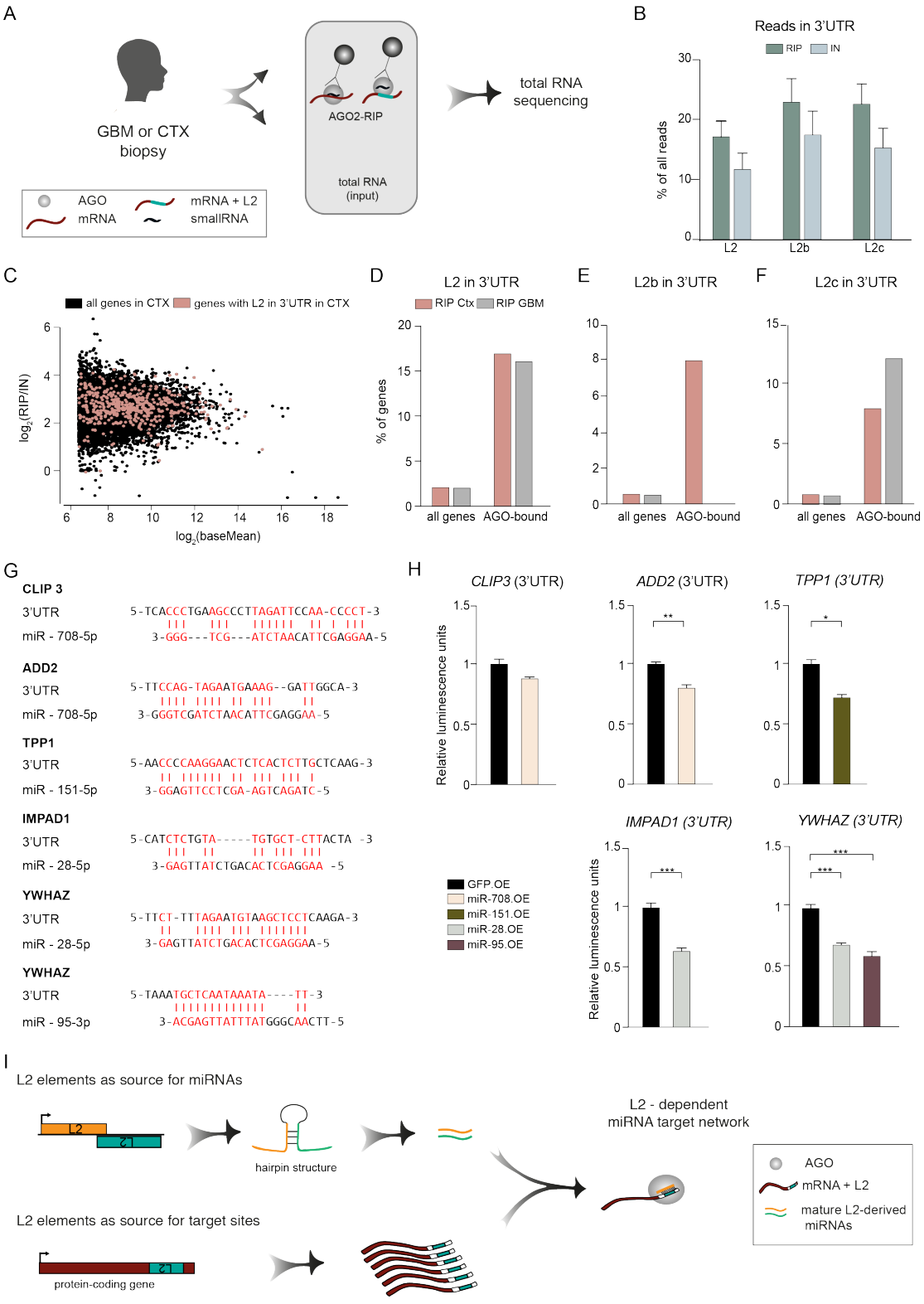

Figure 4: Genes with L2 in 3'UTR are bound by AGO2 in the human brain.
A) Schematics of AGO2-RIP followed by total RNA sequencing on glioblastoma and cortex biopsies. B) Bar graph showing the percentage of L2/ L2b/ L2c reads in the 3'UTR of genes in RIP and input sample (RIP, n=7; IN, n=7). Data is presented as mean ± SEM C) Dot plot of all genes in cortex samples of the human brain. Genes with L2 in the 3'UTR are marked in red. The $\log_2$ transformed mean expression (x-

axis) is plotted against the $\log_2$ transformed fold change of RIP (n=2) versus input samples (n=2) (y-axis). D-F) Bar plots showing the percentage of genes with L2 / L2b / L2c in the 3'UTR of all genes, and of the top 100 highest expressed genes that are more than 4-fold enriched in RIP compared to input samples. G) Potential target sites of L2-derived miRNAs in the L2 sequence in 3'UTRs of genes. Complementary bases are marked in red and with a vertical line. H) Luciferase assay of L2-derived target sites. Data is shown as mean ± SEM. I) Schematics of the TE-based post-transcriptional gene regulatory network.

Discussion

An increasing number of studies have demonstrated a role for TEs in fine-tuning gene networks in different tissues by influencing gene expression from their integration sites, by acting for instance as enhancers or repressors (Jern and Coffin 2008; Elbarbary et al. 2016). In this study, we demonstrate that L2 elements serve as an important source for several miRNAs that are bound by AGO proteins in the human brain with the ability to target hundreds of transcripts carrying L2-fragments in their 3'UTR. Our study therefore provides a model for how TEs control post-transcriptional networks in the human brain.

L2-elements invaded the genome of our ancestors some 100-300 million years ago, well before mammalian radiation (Vladimir V. Kapitonov 2006). Throughout evolution, these TE sequences likely underwent a high evolutionary pressure, leading to the degradation of non-functional sequences, hence, only leaving fragments of sequences remaining in the genome. Today, these sequences form a post-transcriptional network composed of miRNAs and miRNA-target templates present in the 3'UTR of thousands of protein-coding genes. It is tempting to speculate that the majority, if not all, miRNAs have originally emerged from TEs, since TEs are scattered throughout the genome in high numbers and therefore have the potential to

give rise to a large number of miRNAs and miRNA target sites. However, since most miRNAs are evolutionary old (more than a billion years) it is impossible to determine their origin with certainty. Several TEs have previously been implicated in both the generation of miRNAs as well as miRNA-targets (Smalheiser and Torvik 2005). (Piriyapongsa et al. 2007; Roberts et al. 2013; Boudreau et al. 2014; Roberts et al. 2014). However, most of these descriptions are based on computational predictions and the experimental evidence for TE-derived miRNA–target interaction was up until now limited. Key to the findings in this study is the use of AGO-RIP on human tissue coupled to next-generation sequencing. Using this approach we could demonstrate the functionality of L2-derived miRNAs, such as miR-28, miR-151, miR-95 and miR-708 and show that hundreds of transcripts with L2 in their 3'UTR are AGO-bound miRNA targets.

TEs have been suggested to influence speciation and to contribute to primate evolution. Although, we found that ancient L2-derived miRNAs are the most abundant TE-derived miRNAs in the human brain, we also found evidence for miRNAs derived from MIR, L1, L3, LTR and SINE elements, including some younger TEs. We also found evidence for many different TEs being bound to AGO when looking at the target fraction. Thus, further investigations into other TE-derived miRNAs and TE derived targets will be very interesting.

It is commonly accepted that miRNAs regulate genes via the seed sequence, which spans from nucleotides 2-7 in the 5' end of the miRNA (Bartel 2009). Additionally, several alternatives to canonical seed sequences have been proposed (Shin et al. 2010; Kim et al. 2016). For instance, miR-151 and miR-28 have previously been found to

use centred seed pairing to regulate genes (Shin et al. 2010). Since TE-derived miRNAs show high sequence similarities with TE sequences in the 3'UTR of genes, those miRNAs most likely can use extensive base-paring for target recognition and regulation. Those sequence similarities might also explain the wide-spread distribution of TEs in the genome, since the high number of possible interactions of TE-derived miRNAs and TE-sequences in genes most likely led to conserved interactions during evolution. This suggests that non-canonical miRNA target sites might be more broadly used than previously thought.

Glioblastoma is the most aggressive and most common type of brain tumour in adults and is characterised by vast cellular and genetic heterogeneity. Nevertheless, we found, at large, surprisingly comparable miRNA regulation of transcripts in GBM compared to normal neocortical tissue, with a few notable differences including previously identified glioma-associated miRNAs such as miR-21 and miR-10b. In addition we found that the L2b-derived miR-95 was highly expressed in cortex, but not in GBM, which is in line with a previous report (Skalsky and Cullen 2011). Together with our finding that L2b-carrying transcripts are not bound to AGO in glioblastoma tissue compared to cortex samples, this suggests that miR-95 could play an important role in the regulation of transcripts related to tumour progression or tumour defence in GBM. Future studies analysing the function of miR-95 are therefore warranted.

In summary, this work demonstrates the existence of a TE-based post-transcriptional regulatory network that shapes the expression of hundreds of TE-carrying transcripts

and hence provides an additional mechanism for TEs to influence crucial gene networks in the developing and adult human brain.

Methods

Human Tissue

Fresh human adult neocortical and GBM tissues were obtained during resective surgery in patients suffering from GBM or pharmacologically intractable epilepsy, respectively. The tissue was snap frozen immediately following removal. The use of human brain tissue was approved by the local Ethical Committee in Lund (212/2007 for epilepsy and H15 642/2008 for GBM) in accordance with the declaration of Helsinki. Prior to each surgery written informed consent were obtained from all subjects.

Mouse brain tissue

All animal-related procedures were approved and conducted in accordance with the committee for use of laboratory animals at Lund University. For AGO2-RIPseq the striata of mice were quickly dissected and immediately homogenised and lysed in ice-cold lysis buffer.

Human ESC culturing and differentiation into forebrain-like cells

Human ESC H9 (WA09, passage 31-45) (Thomson et al. 1998) was expanded and maintained on γ-irradiated mouse embryonic fibroblasts (MEFs) in DMEM/F12, 20%

KSR, 0.05 mM 2-mercaptoethanol, 0.5% pen/strep, 0.5% glutamate and 10 ng/ml FGF-2 (R&D Systems). The cells were passaged once weekly with EDTA (0.5 mM).

Differentiation was initiated by detaching the hESC colonies with EDTA and grown as free-floating aggregates in DMEM/F12:Neurobasal (1:1) supplemented with N2 (1:100), B27 (without vitamin A) (1:50) for 4 days and Y-27632 (10 mM, Tocris Bioscience) for the initial 2 days. The formed EBs were thereafter plated in DMEM/F12:Neurobasal (1:1), N2 (1:200) and B27 (without vitamin A) (1:100) onto a surface coated with polyornithine (PO), fibronectin (FN) and laminin (lam).

From d0 to d9, SB431542 (10 mM, Tocris Bioscience) and noggin (200 ng/ml, R&D) were present in the medium for neuralisation. At day 11, the attached cell clusters were dissociated to a single-cell suspension with Accutase, and were replated in 20 µl droplets of 10,000-15,000 cells/µl onto dry PO/FN/lam-coated plates in Neurobasal, B27 (without vitamin A) (1:50), brain-derived neurotrophic factor (BDNF) (20 ng/ml), glial-derived neurotrophic factor (GDNF) (10 ng/ml) and ascorbic acid (200 mM).

AGO-RIPseq

Fresh or snap-frozen tissue was homogenised in ice-cold lysis buffer (10 mM HEPES (pH = 7.3), 100 mM KCl, 0.5 % NP40, 5 mM $MgCl_2$, 0.5 mM dithiothreitol, protease inhibitors, recombinant RNase inhibitors, 1 mM PMSF) using TissueLyser LT (30 Hz, 4 minutes).

Homogenates were centrifuged twice for 15 minutes at 16,200 x g, 4 °C to clear the lysate. 1/10 of the sample was then saved as input (total RNA) control. The remaining lysate was incubated with anti-AGO2 or anti-AGO1-coated Dynabeads® Protein G beads (Life Technologies) at 4 °C for 24 hours with end-over-end rotation (AGO2

antibody: anti-Ago2-3148 for human brain tissue (Grey et al. 2010); AGO1 antibody: Active Motif, 61071 and Sigma-Aldrich 2E12-1C9 for mouse tissue).

After incubation, the beads were collected on a Dynamagnet (1 minute, on ice) and gently resuspended in low-salt NT2 buffer (50 mM Tris-HCL (pH = 7.5), 1 mM $MgCl_2$, 150 mM NaCl, 0.5 % NP40, 0.5 mM dithiothreitol, 1 mM PMSF, protease inhibitors and recombinant RNAse inhibitors). The beads were transferred into a new collection tube and washed once with low-salt NT2 buffer, followed by two washes with high-salt NT2 buffer (50 mM Tris-HCl (pH = 7.5), 1 mM MgCl2, 600 mM NaCl, 0.5 % NP40, 0.5 mM DTT, protease inhibitors, 1 mM PMSF and recombinant RNAse inhibitors). After the last washing step, the RNA fraction was resuspended in QIAzol buffer and RNA was isolated from RIP and input samples according to the miRNeasy micro kit (Qiagen).

cDNA library preparation and sequencing of human AGO-RIP samples

All RNA sequencing data have been submitted to NCBI Gene Expression Omnibus database and assigned the GEO series accession number GSE106810. cDNA library preparation was conducted using the NEB small RNA library prep kit for small RNA sequencing and the NuGen Ovation RNAseq V2 kit, followed by the Ovation® Ultralow V2 Library or Ovation® Rapid Library Systems, for total RNA sequencing. Illumina high-throughput sequencing (HiSeq2500 SR 1x50 run and HiSeq3000 1x50) was applied to the samples (total number of reads for small RNA sequencing: 344508344; total number of reads for total RNA sequencing: 527413532) at the UCLA Microarray Core Facility.

Identification of miRNA target sites

The L2 sequence in the 3'UTR of a gene was aligned to the sequence of a L2-derived miRNA using the EMBOSS Needle nucleotide pairwise alignment tool (Rice et al. 2000) to find sequence complementarity.

Analyses of small RNA sequencing data from AGO-RIP

For the analyses of the genomic distribution of reads, the data was aligned to the human genome (hg38) using StarAligner 2.5.0a (Dobin et al. 2013) allowing multimapping and two mismatches per 22bp (--outFilterMismatchNoverLmax 0.05). Reads were quantified with the SubRead package FeatureCounts (Liao et al. 2014) (minimal overlap 19 nt) using annotations obtained from miRbase (Kozomara and Griffiths-Jones 2014), UCSC genome browser RepeatMasker track (GRCh38) and piRNAbank (Sai Lakshmi and Agrawal 2008).

For the analyses of small RNAs derived from individual transposable elements, the data was uniquely aligned to the human genome (hg38) using StarAligner 2.5.0a (0 mismatches allowed (--outFilterMultimapMax 1) and reads were quantified strand-specifically (-s1) using the SubRead package FeatureCounts (Liao et al. 2014).

For the identification of high-confidence smallRNAs, 3p and 5p strands and typical Dicer cleavage patterns had to be present. We also assessed conservation of the smallRNAs by using the 100 vertebrates Basewise Conservation data by phyloP from the PHAST package.

For the alignment of the L2c small RNAs to the L2c consensus, the raw alignment data used by RepeatMasker to generate the annotation of genomic L2c were assessed though the detailed visualisation of RepeatMasker annotations in the UCSC genome browser. For each L2c-small RNA sequence assessed, the RepeatMasker raw alignment was used to map the genomic position of the small RNA to the L2c consensus.

Analyses of total RNA sequencing data from AGO-RIP on human tissue

For the expression analyses of AGO2-bound genes, reads were aligned to the human genome (hg38) using StarAligner 2.5.0a allowing for multimappers and default settings. For the quantification of reads in the 3'UTR of genes, annotations were obtained from Ensembl (Kersey et al. 2016). For the identification of L2 elements within 3'UTRs, reads were uniquely aligned, allowing for 2 mismatches in 50bp (--outFilterMismatchNoverLmax 0.04). The obtained files were intersected with L2 annotations, obtained from RepeatMasker with the requirement of 80 % overlap (-f 0.8). The reads were then intersected with 3'UTR annotations (Ensembl) with a requirement of 1 bp overlap. The intersection was conducted using BEDTools (2.26.0).

Culturing and RNA extraction from hNPCs

Human neural epithelial-like stem cell lines, Sai2 (embryo-derived) (Tailor et al. 2013) were cultured in DMEM/F12 (Thermo Fisher Scientific) supplemented with Glutamine (2 mM, Sigma), Penicillin/Streptomycin (1x, Gibco), N2 supplement (1x, Thermo Fisher Scientific), B27 (0.05x, Invitrogen), EGF and FGF2 (both 10 ng/ml, Thermo Fisher Scientific). hNES were grown on NuncTM T25 or T75 flasks pre-coated with Poly L-Ornithine (15 µg/ml, Sigma) and Laminin (2 µg/ml, Sigma). Cells were passaged every 2-3 days using TryplLETM Express enzyme (Life Technologies) and Defined Trypsin Inhibitor (Life Technologies) and plated at a density of $6x10^4$ per cm^2. RNA from cells was extracted using the RNeasy mini kit from Qiagen.

cDNA library preparation and sequencing of human NPC

cDNA library preparation was conducted using the TrueSeq mRNA stranded kit (with poly-A enrichment) from Illumina according to the supplier's recommendations. Paired-end 126 bp sequencing was conducted with Hiseq3000 at SciLife Lab, Uppsala, Sweden.

Analyses of RNA sequencing data from human NPC

The paired-end 126 bp reads were aligned to the human reference genome (hg38) using STAR (v2.5.0a)(Dobin et al. 2013). For mapping of L2, the alignment allowed 3 mismatches per read (--outFilterMismatchNoverLmax 0.03) and discarding all multimapping reads (--outFilterMultimapMax 1). For mapping of mRNAs, default parameters were used. Gene abundances were counted using the SubRead package FeatureCounts with strand-specific counting (-s 2) of primary alignments (--primary) in paired-end mode (-p) using NCBI annotations. To quantify reads mapping to L2 and 3'UTR, the uniquely aligned read coordinates were intersected with RepeatMasker L2 coordinates, using the intersect module of BEDTools (v2.26.0) requiring that at least 80 % of the read overlap with an L2 element (-f 0.8). These read coordinates were further intersected if more than 1 bp overlapped with 3'UTR annotation (Ensembl) to get the number of L2 reads derived from 3'UTR sequences. For obtaining a list of transcripts with L2 in the 3'UTR, the Ensembl 3'UTR annotation was overlapped with L2 RepeatMasker annotation using BEDTools intersection requiring at least 1 bp overlap. To get the coverage of bases derived from each base position in L2c consensus, the relative position of L2c elements aligning to L2c consensus were extracted from the RepeatMasker output file (http://hgdownloadtest.cse.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.out.gz). The coverage module of BEDTools (v2.24.0) was used to calculate coverage of each position in consensus.

Gene ontology analyses was conducted using Gene Ontology Consortium (Ashburner et al. 2000; Gene Ontology 2015). As background, a list of all expressed genes in hNPCs was used.

Analyses of RNA sequencing data from human embryonic tissue and differentiated hNPCs

The RNA-seq data from human embryonic tissue have been reported previously (GSE84259) (Brattas et al. 2017). Reads were mapped to the human genome assembly (GRCH38) using STAR (v2.5.0a). Multiple mapping reads were discarded, and a maximum of 3 mismatches were allowed. mRNA and TE expression was quantified using the subread package FeatureCounts (Liao et al., 2014). Gene coordinates for mRNA and retroelements were obtained from NCBI and the UCSC genome browser RepeatMasker track (GRCh38). Read counts were normalised to the total number of reads mapping to the genome. Downstream analyses were performed using DESeq2 (Love et al. 2014), in-house R and unix scripts.

Luciferase reporter assay

A 400 bp sequence incorporating the L2-derived miRNA binding sites in the 3'UTR of CLIP3, Impad1, YWHAZ, TPP1 and ADD2 was cloned into the dual luciferase reporter vector pSICHECK-2 (Promega). The luciferase reporter constructs were co-transfected with either a GFP overexpression vector, a miR-10a non-targeting over-expression construct or the respective L2-derived miRNA overexpression vector into three independent replicates of 293T cells using Turbofect (Fermentas). 48 hours after transfection, cells were assayed for luminescence using a dual-luciferase assay (Promega). One-way ANOVA followed by a Tukey's multiple comparison post hoc

test were performed in order to test for statistical significance. Data is presented as mean ± SEM.

Acknowledgement

We would like to thank Stefan Thor, Volker Busskamp and Didier Trono as well as all members of the Jakobsson lab for useful comments on the manuscript. We also thank J. Johansson, M. Persson Vejgården, U. Jarl, A. Hammarberg, E. Ling, B. Mattson, S. da Rocha Baez and M. Sparrenius for technical assistance.

The work was supported by grants from the Swedish Research Council, the Swedish Foundation for Strategic Research, the Swedish Brain Foundation; the Swedish excellence project Basal Ganglia Disorders Linnaeus Consortium (Bagadilico), and the Swedish Government Initiative for Strategic Research Areas (MultiPark & StemTherapy).

Author contributions

R.P., P.L.B, M.J., K.P., J.B. and J.J. designed and performed research and analyzed data. R.P. and P.L.B performed Bioinformatics analysis. R.P. and J.J. wrote the paper and all authors reviewed the manuscript.

References

Akerblom M, Sachdeva R, Barde I, Verp S, Gentner B, Trono D, Jakobsson J. 2012. MicroRNA-124 is a subventricular zone neuronal fate determinant. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **32**(26): 8879-8889.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**(1): 25-29.

Bartel DP. 2009. MicroRNAs: target recognition and regulatory functions. *Cell* **136**(2): 215-233.

Boudreau RL, Jiang P, Gilmore BL, Spengler RM, Tirabassi R, Nelson JA, Ross CA, Xing Y, Davidson BL. 2014. Transcriptome-wide discovery of microRNA binding sites in human brain. *Neuron* **81**(2): 294-305.

Brattas PL, Jonsson ME, Fasching L, Nelander Wahlestedt J, Shahsavani M, Falk R, Falk A, Jern P, Parmar M, Jakobsson J. 2017. TRIM28 Controls a Gene Regulatory Network Based on Endogenous Retroviruses in Human Neural Progenitor Cells. *Cell reports* **18**(1): 1-11.

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. *Nature reviews Genetics* **18**(2): 71-86.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**(1): 15-21.

Elbarbary RA, Lucas BA, Maquat LE. 2016. Retrotransposons as regulators of gene expression. *Science* **351**(6274): aac7247.

Gene Ontology C. 2015. Gene Ontology Consortium: going forward. *Nucleic acids research* **43**(Database issue): D1049-1056.

Grey F, Tirabassi R, Meyers H, Wu G, McWeeney S, Hook L, Nelson JA. 2010. A viral microRNA down-regulates multiple cell cycle genes through mRNA 5'UTRs. *PLoS pathogens* **6**(6): e1000967.

Imig J, Brunschweiger A, Brummer A, Guennewig B, Mittal N, Kishore S, Tsikrika P, Gerber AP, Zavolan M, Hall J. 2015. miR-CLIP capture of a miRNA targetome uncovers a lincRNA H19-miR-106a interaction. *Nature chemical biology* **11**(2): 107-114.

Jern P, Coffin JM. 2008. Effects of retroviruses on host genome function. *Annual review of genetics* **42**: 709-732.

Karsy M, Arslan E, Moy F. 2012. Current Progress on Understanding MicroRNAs in Glioblastoma Multiforme. *Genes & cancer* **3**(1): 3-15.

Kersey PJ, Allen JE, Armean I, Boddu S, Bolt BJ, Carvalho-Silva D, Christensen M, Davis P, Falin LJ, Grabmueller C et al. 2016. Ensembl Genomes 2016: more genomes, more complexity. *Nucleic acids research* **44**(D1): D574-580.

Kim D, Sung YM, Park J, Kim S, Kim J, Park J, Ha H, Bae JY, Kim S, Baek D. 2016. General rules for functional microRNA targeting. *Nature genetics* **48**(12): 1517-1526.

Kozomara A, Griffiths-Jones S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research* **42**(Database issue): D68-73.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**(7): 923-930.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**(12): 550.

Piriyapongsa J, Marino-Ramirez L, Jordan IK. 2007. Origin and evolution of human microRNAs from transposable elements. *Genetics* **176**(2): 1323-1337.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics : TIG* **16**(6): 276-277.

Roberts JT, Cardin SE, Borchert GM. 2014. Burgeoning evidence indicates that microRNAs were initially formed from transposable element sequences. *Mobile genetic elements* **4**: e29255.

Roberts JT, Cooper EA, Favreau CJ, Howell JS, Lane LG, Mills JE, Newman DC, Perry TJ, Russell ME, Wallace BM et al. 2013. Continuing analysis of microRNA origins: Formation from transposable element insertions and noncoding RNA mutations. *Mobile genetic elements* **3**(6): e27755.

Sai Lakshmi S, Agrawal S. 2008. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic acids research* **36**(Database issue): D173-177.

Shin C, Nam JW, Farh KK, Chiang HR, Shkumatava A, Bartel DP. 2010. Expanding the microRNA targeting code: functional sites with centered pairing. *Molecular cell* **38**(6): 789-802.

Skalsky RL, Cullen BR. 2011. Reduced expression of brain-enriched microRNAs in glioblastomas permits targeted regulation of a cell death gene. *PloS one* **6**(9): e24248.

Smalheiser NR, Torvik VI. 2005. Mammalian microRNAs derived from genomic repeats. *Trends in genetics : TIG* **21**(6): 322-326.

Tailor J, Kittappa R, Leto K, Gates M, Borel M, Paulsen O, Spitzer S, Karadottir RT, Rossi F, Falk A et al. 2013. Stem cells expanded from the human embryonic hindbrain stably retain regional specification and high neurogenic potency. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **33**(30): 12407-12422.

Tan CL, Plotkin JL, Veno MT, von Schimmelmann M, Feinberg P, Mann S, Handler A, Kjems J, Surmeier DJ, O'Carroll D et al. 2013. MicroRNA-128 governs neuronal excitability and motor behavior in mice. *Science* **342**(6163): 1254-1258.

Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ, Marshall VS, Jones JM. 1998. Embryonic stem cell lines derived from human blastocysts. *Science* **282**(5391): 1145-1147.

Vladimir V. Kapitonov AP, Jerzy Jurka. 2006. Anthology of Human Repetitive DNA. *Reviews in Cell Biology and Molecular Medicine*.

Weinmann L, Hock J, Ivacevic T, Ohrt T, Mutze J, Schwille P, Kremmer E, Benes V, Urlaub H, Meister G. 2009. Importin 8 is a gene silencing factor that targets argonaute proteins to distinct mRNAs. *Cell* **136**(3): 496-507.