Supplemental Material

Transposable elements and their KZFP controllers are drivers of transcriptional innovation in the developing human brain

Christopher J. Playfoot[1], Julien Duc[1], Shaoline Sheppard[1], Sagane Dind[1], Alexandre Coudray[1], Evarist Planet[1] and Didier Trono[1, 2]

[1]School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

[2]Corresponding author. Didier.Trono@epfl.ch

Contents

Supplemental Figures 2

Figure S1. Summary of brain developmental datasets analysed 2

Figure S2. KZFP expression trajectory is different in the CB compared to the DFC 3

Figure S3. TE subfamilies and unique loci exhibit spatiotemporal expression patterns in the cerebellum 5

Figure S4. Non-KZFP TF:TE subfamily expression and binding relationships in neurogenesis 7

Figure S5. TcGTs are cell type specific 9

Figure S6. TcGTs are spatially expressed in broad or specific brain regions and are bound by KZFPs 10

Figure S7. The TcGT detection criteria of one spliced read in over 20% of samples represents a sensitive detection approach 12

Figure S8. TcGTs are expressed in SH-SY-5Y neuroblastoma cells, are primarily brain specific and L2:DDRGK1 is a predicted chimeric protein 14

Figure S9. The L2:DDRGK1 TcGT is conserved in primates and has the same behaviour in macaque as in humans 16

Figure S10. The signal peptide is lost in N-truncated TcGTs 18

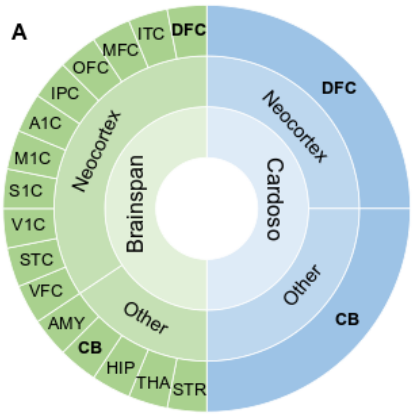
Supplemental Table Descriptions 19

Supplemental Methods 20

Supplemental Acknowledgements 25

Supplemental References 26

Supplemental Figures



B

Brainspan Region	Abv.	4-7pcw	8-9pcw	10-12pcw	13-15pcw	16-18pcw	19-24pcw	25-38pcw	Birth-5month	6-18 month	19month-5yr	6-11 yr	12-19 yr	20-60+ yr	Sum
		1	2A	2B	3A	3B	4	5	6	7	8	9	10	11	
Amygdala	AMY	0	2	4	4	1	2	1	3	1	1	2	3	7	31
Cerebellar cortex	CB	0	0	3	2	1	3	2	2	2	3	2	4	7	31
Dorsolateral prefrontal cortex	DFC	0	2	4	4	2	4	1	2	2	2	2	3	6	34
Hippocampus	HIP	0	2	4	4	2	4	1	3	1	1	2	2	7	33
Inferior temporal cortex	ITC	0	1	4	4	0	3	1	3	2	2	2	4	7	33
Medial prefrontal cortex	MFC	0	2	3	4	2	4	1	2	2	2	2	3	6	33
Mediodorsal nucleus of the thalamus	THA	0	0	4	2	2	4	1	3	2	2	1	2	7	30
Orbital prefrontal cortex	OFC	0	2	4	3	1	3	1	2	2	1	2	3	7	31
Posterior inferior parietal cortex	IPC	0	0	4	4	2	3	1	2	3	2	2	4	7	34
Primary auditory (A1) cortex	A1C	0	0	4	4	2	2	1	3	1	2	2	2	7	30
Primary motor (M1) cortex	M1C	0	0	4	3	0	2	1	3	1	2	1	3	7	27
Primary somatosensory (S1) cortex	S1C	0	0	4	3	0	2	1	3	3	2	1	4	7	30
Primary visual (V1) cortex	V1C	0	0	4	4	2	4	1	2	2	1	2	4	6	32
Striatum	STR	0	0	4	4	2	3	1	2	1	2	1	1	7	28
Superior temporal cortex	STC	0	1	3	3	2	4	1	3	2	3	2	3	7	34
Ventrolateral prefrontal cortex	VFC	0	1	4	4	2	4	1	2	2	3	2	4	7	36
															507

Cardoso Region	Abv.	1	2A	2B	3A	3B	4	5	6	7	8	9	10	11	Sum
Dorsolateral prefrontal cortex	DFC	8	7	6	3	3	5	0	3	2	3	2	4	9	55
Cerebellar cortex	CB	12	6	9	3	3	0	0	5	3	2	3	4	9	59
															114

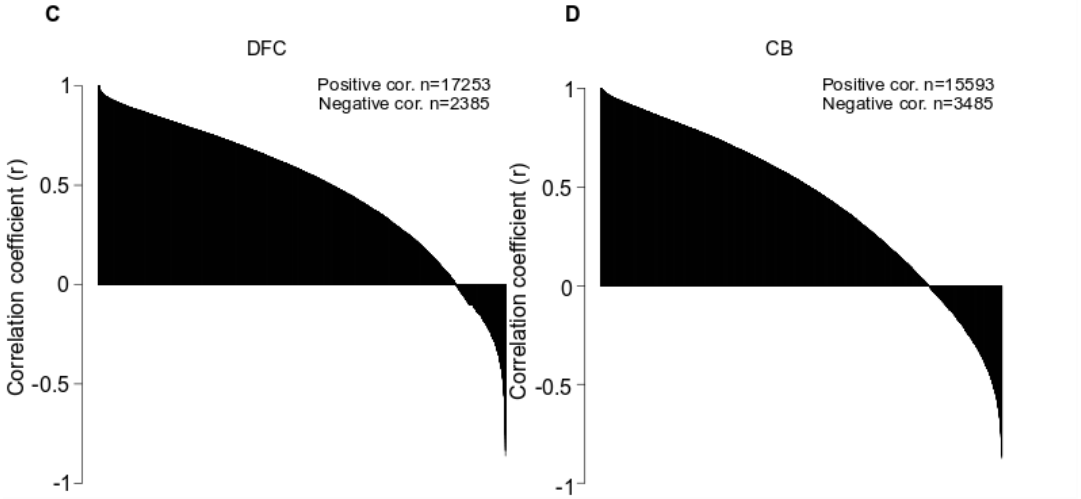


Figure S1. Summary of brain developmental datasets analysed. (A) Donut plot of brain regions incorporated in both datasets. Region abbreviations are shown in B. (B) Number of samples analysed in both datasets, highlighting numbers of samples per region and stage. Stages were as defined by the Brainspan consortia. (C) Barplots depicting the Pearson correlation coefficient between Brainspan and Cardoso datasets for all genes passing expression criteria (methods) for the DFC and the (D) CB.

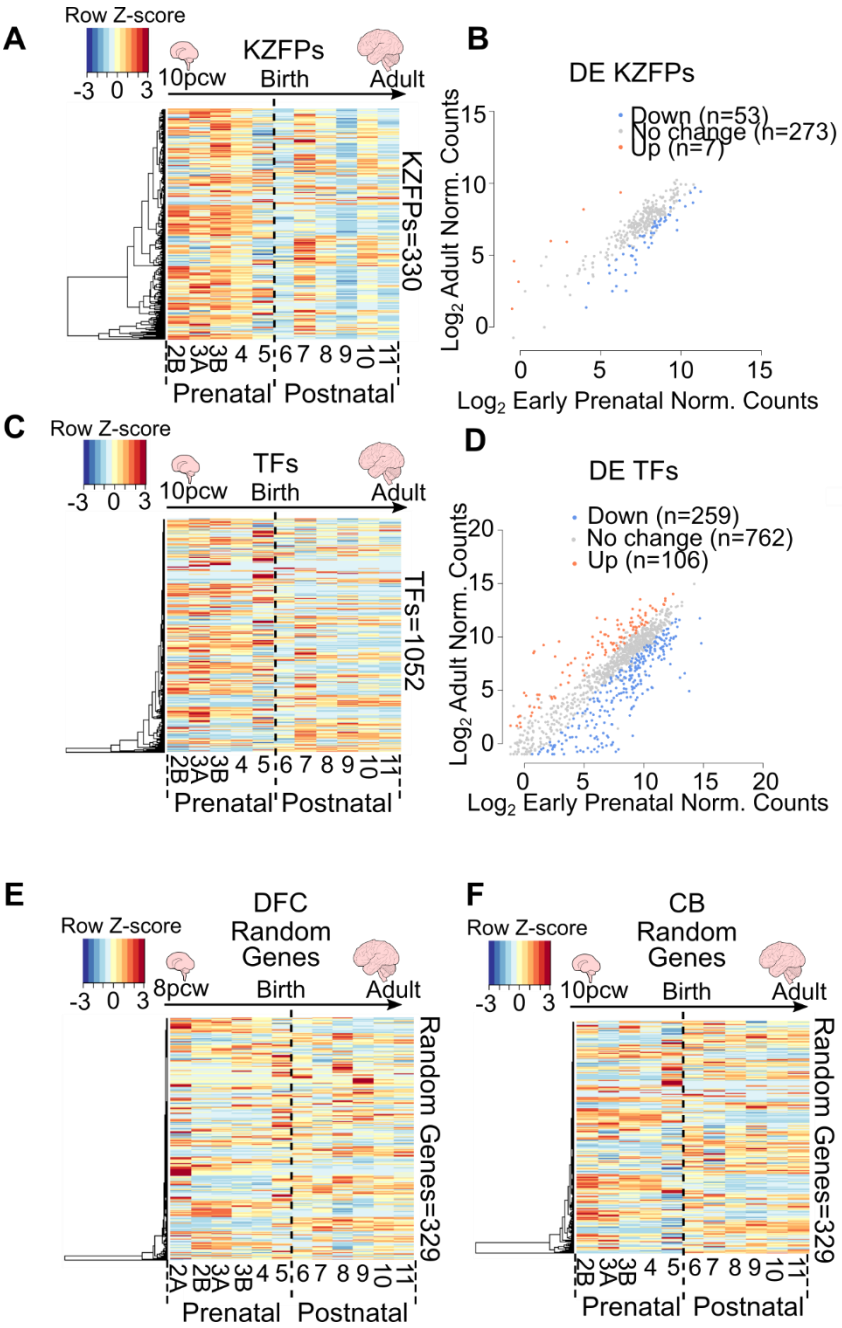


Figure S2. KZFP expression trajectory is different in the CB compared to the DFC. (A) Heatplots of KZFP expression across human neurogenesis in the cerebellum. Scale represents the row Z-score. Stage 2A was omitted due to lack of samples for CB (see Fig. S1B). See also Supplemental Table 2. (B) Dot plot of differential expression analysis of KZFPs in the CB comparing adult (stage 11) to early prenatal stages (stage 2A to 3B) of neurogenesis. Only KZFPs behaving the same in both datasets are shown. Up (orange) represents KZFPs significantly upregulated in adult versus early prenatal (Fold change ≥ 2, FDR ≤ 0.05). Down (blue) represents KZFPs significantly downregulated in adult (Fold change ≤ -2, FDR ≤ 0.05). See also Supplemental Table 3. (C) Heatplots of TF expression across human neurogenesis in the CB. Scale represents the row Z-score. Stage 2A was omitted due to lack of samples for CB (see Fig. S1B). See also Supplemental Table 2. (D) Dot plot of differential expression analysis of TFs in the CB comparing adult (stage 11) to early prenatal stages (stage 2A to 3B) of neurogenesis. Only TFs behaving the same in both datasets are shown. Up (orange) represents TFs significantly upregulated in adult

versus early prenatal (Fold change ≥ 2, FDR ≤ 0.05). Down (blue) represents TFs significantly downregulated in adult (Fold change ≤ -2, FDR ≤ 0.05). See also Supplemental Table 3. All plots show expression data from Brainspan. (E) Heatplots of non-TF, non-KZFP random gene expression across human neurogenesis in the DFC and (F) CB. Scale represents the row Z-score. Stage 2A was omitted due to lack of samples for CB (see Fig. S1B). See also Supplemental Table 2.

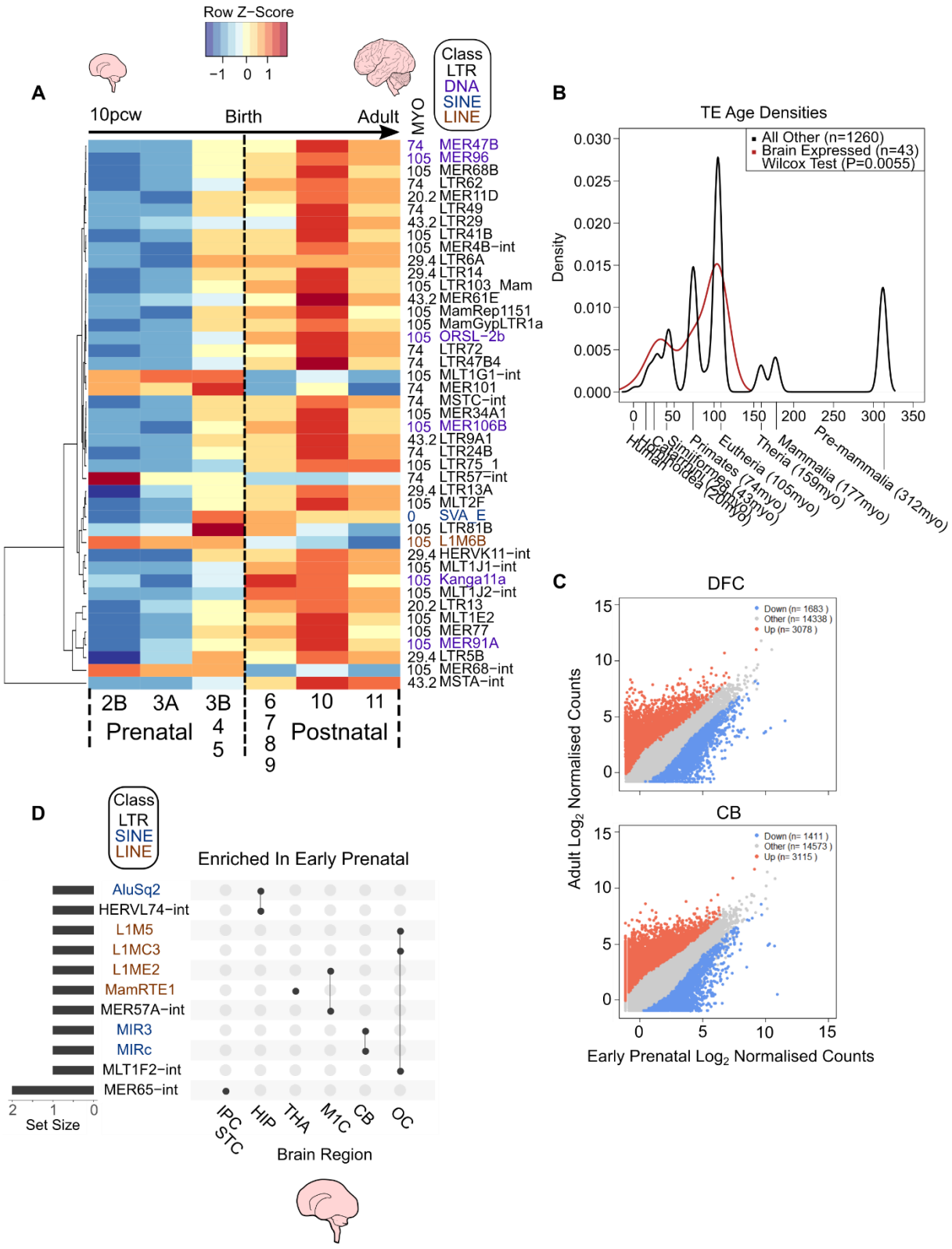
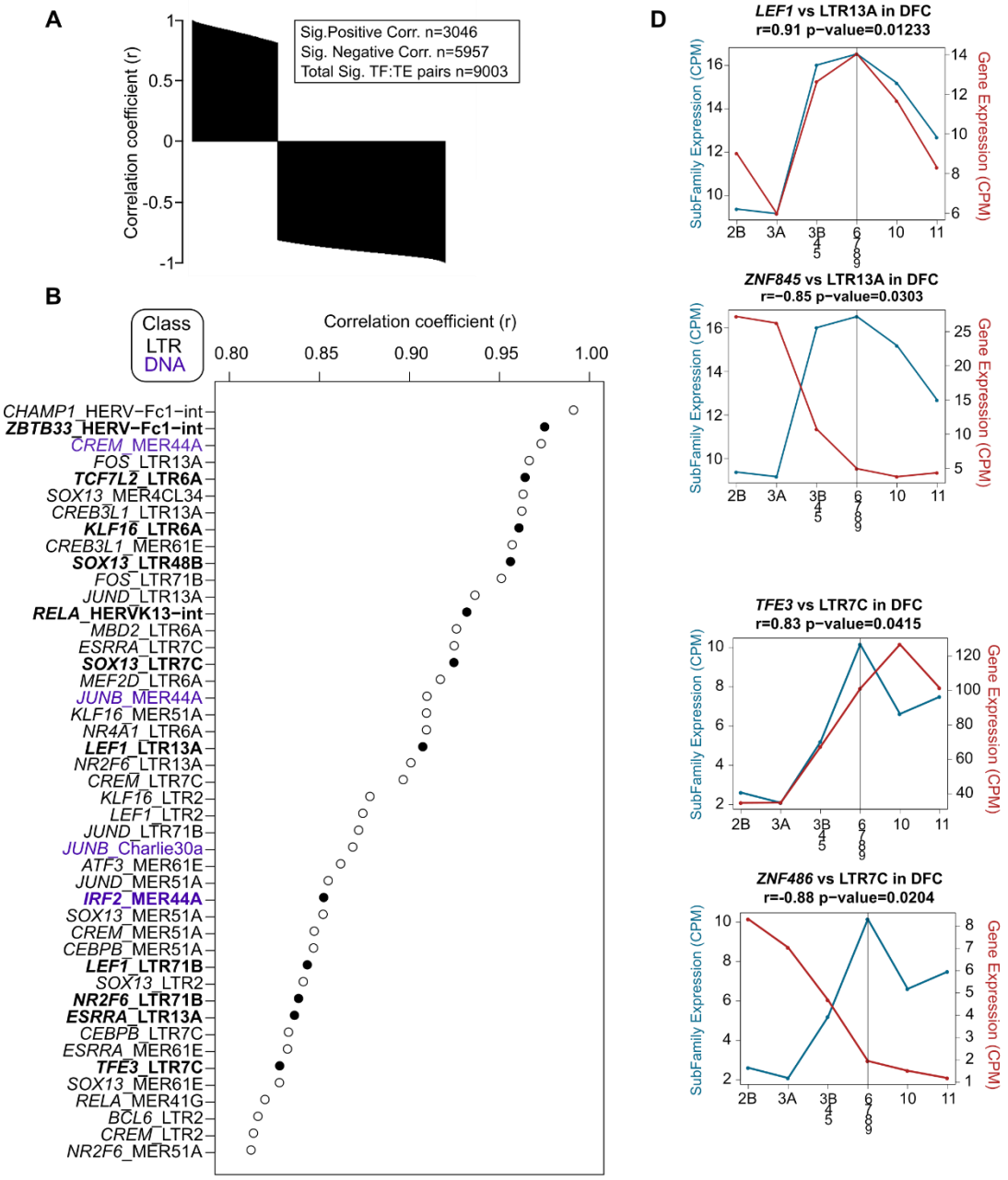


Figure S3. TE subfamilies and unique loci exhibit spatiotemporal expression patterns in the cerebellum. (A) Heatplot of TE subfamilies with concordant expression behaviours between both datasets (Pearson correlation coefficient ≥ 0.7) across human neurogenesis in the cerebellum. See also Supplemental Table 4. The mean expression values for stages 3B, 4 and 5, and also stages 6, 7, 8 and 9 were combined and averaged to reduce inherent variability due to low numbers of samples for some stages (see Supplemental Fig. S1B). Scale represents the row Z-score. TE subfamily age in million years

old (MYO) and class is shown to the right of the plot. (B) Density plot depicting estimated age of TEs in A (*P≤0.05*, Wilcoxon test). Evolutionary stages and corresponding ages are shown beneath the plot. (C) Dot plots of differential expression analysis of unique TE loci in the DFC and CB comparing adult (stage 11) to early prenatal stages (stage 2A to 3B) of neurogenesis. Only TEs behaving the same in both datasets are shown. Up (orange) represents TEs significantly upregulated in adult versus early prenatal (Fold change ≥ 2, FDR ≤ 0.05). Down (blue) represents TEs significantly downregulated in adult (Fold change ≤ -2, FDR ≤ 0.05). See also Supplemental Table 6 & 7. (D) UpSet plot showing the significantly enriched differentially expressed subfamilies for early pre-natal versus adult stages per region from unique mapping analyses. Set size represents the number of regions the specific TE was significantly differentially enriched in. Joined points represent combinations of significantly differentially expressed TE subfamilies. All plots show data from Brainspan. See also Supplemental Table 6 & 7.



C

Motif ID	TE Subfam ID	Start	Stop	Strand	Score	p-value	q-value	Matched Sequence
LEF1	LTR13A	607	621	-	18.9677	2.86E-07	0.000533	AAAGATCAAAGACTT
IRF2	MER44A	17	34	-	12.1927	5.66E-06	0.00339	GGAAAGCGAAACCGCGGA
TCF7L2	LTR6A	151	164	+	14.7636	6.18E-06	0.00617	GCACATCAAAGAGT
SOX13	LTR48B	364	374	-	13.0565	1.24E-05	0.0122	AGACAATGGCT
KLF16	LTR6A	104	114	+	12.1034	2.96E-05	0.0158	TACCCACCCCC
NR2F6	LTR71B	262	275	+	11	0.000023	0.0261	TAGGTCATAAGTCA
LEF1	LTR71B	194	208	+	11	2.62E-05	0.0301	TAAGATGAAAGTTTA
TFE3	LTR7C	463	470	+	11.6854	4.87E-05	0.0389	CGCGTGAC
SOX13	LTR7C	165	175	-	11.7661	5.42E-05	0.0401	TCACAATGGTG
ZBTB33	HERV-Fc1	2246	2260	+	13.5455	1.21E-05	0.0985	CCCTCGCGCGGCTCC
ESRRA	LTR13A	254	266	-	10.4146	8.47E-05	0.16	CTTCTAGGTCACT
RELA	HERVK13	2425	2434	+	12.5306	2.89E-05	0.421	GGGGCATTCC

Figure S4. Non-KZFP TF:TE subfamily expression and binding relationships in neurogenesis. (A) Barplots depicting the Pearson correlation coefficient (r>0.7, *P-value ≤ 0.05*) between non-KZFP TF genes and all TE subfamilies behaving the same in Brainspan and Cardoso datasets (r>0.2) for the DFC. The mean expression values for stages 3B, 4 and 5, and also stages 6, 7, 8 and 9 were averaged to

perform the correlation test. (B) Plot showing only the TF:TE subfamily pairs with both a positive significant expression correlation (r ≥ 0.8, *P-value ≤ 0.05*) and significantly enriched binding of the TF to the TE subfamily using ENCODE TF ChIP-seq data and a custom ChIP-seq binding enrichment script (Fold change between expected and observed binding >3, adjusted *P-value ≤ 1e-4*). TF:TE pairs shown in bold also had a detectable TF binding motif within the consensus target TE subfamily sequence from Dfam as shown in C. (C) Output table from FIMO (Grant et al., 2011), showing TF:TE subfamily pairs from B with detected binding motifs within consensus TE subfamily sequences from Dfam. (D) Line plots showing expression in counts per million (CPM) of selected TF:TE target pairs from B, or KZFP:TE target pairs with their Pearson correlation coefficient and significance shown. Grey line indicates birth at stage 6.

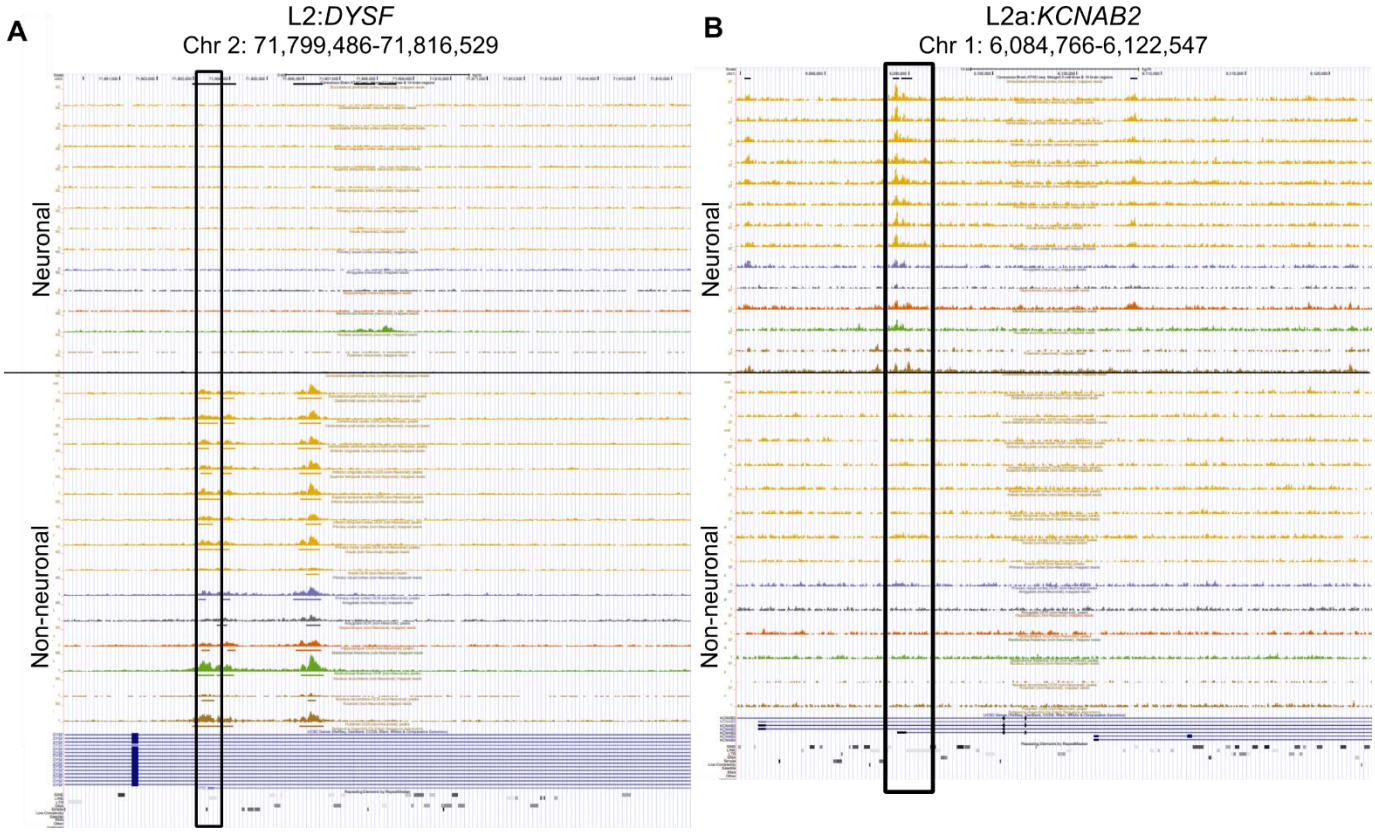


Figure S5. TcGTs are cell type specific. UCSC genome browser shots of TcGT TE TSS loci (black box) with consensus ATAC-seq peaks from isolated neuronal and non-neuronal cells from different regions of the adult human brain from BOCA (Fullard et al., 2018) for (A) the non-neuronal associated TcGT L2:*DYSF* and (B) the neuronal associated TcGT L2a:*KCNAB2*. Track colours correspond to the following regions: Yellow = different neocortex regions, purple = primary visual cortex, black = amygdala, red = hippocampus, green = mediodorsal thalamus, brown= nucleus accumbens and putamen.

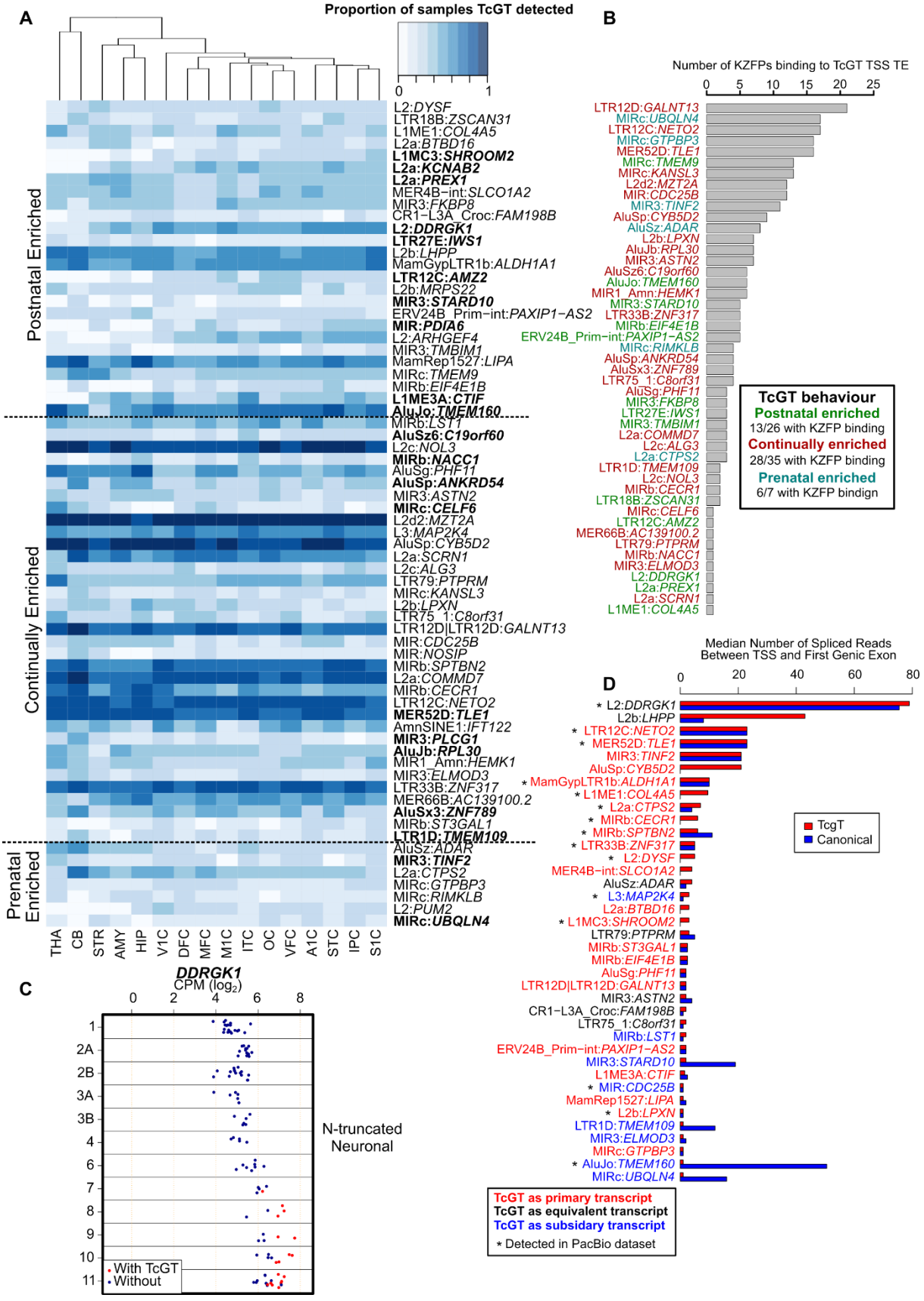


Figure S6. TcGTs are spatially expressed in broad or specific brain regions and are bound by KZFPs. (A) Heatplot showing the proportion of samples per brain region the 68 TcGTs were detected in the

Brainspan dataset regardless of developmental stage it was detected. TcGTs are in the same order as Fig. 4A. (B) Barplot showing the number of KZFPs binding to the TE derived TSS of each TcGT. TcGT behaviour is indicated for postnatal (green), continual (red), prenatal (aqua) detected TcGTs. (C) Dot plot showing the gene expression level per stage for *DDRGK1* for samples where the TcGT was detected (red) and where it was not (blue) from Cardoso dataset as comparison to Fig. 4A. (D) Barplot indicating the median number of spliced reads between the TE (red bars) or canonical promoter (blue bars) and the first genic exon. Colored text indicates the manually determined contribution of the TcGT to gene expression. * indicates the TcGT was detected in the PacBio dataset from Fig. 5. A caveat with this analysis is that in some cases the canonical 'WT' transcript is not the longest, resulting in an over-representation of the contribution of the TcGT because the actual WT isoform expressed is a shorter transcript. Another caveat is where the TcGT isoform was the longest and was annotated in Ensembl, thus resulting in a ratio of 1:1; an under-representation of the contribution of the TcGT.

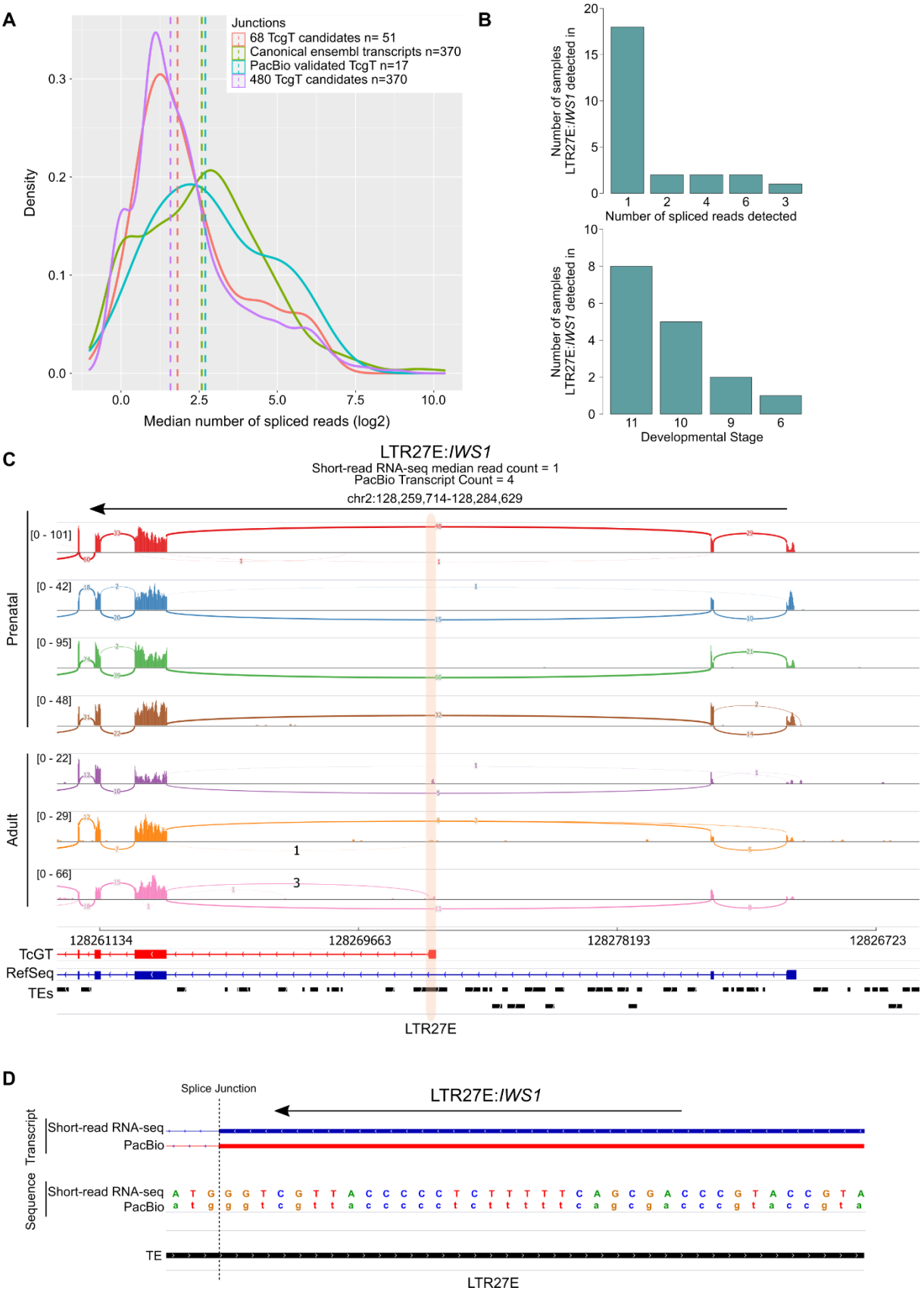


Figure S7. The TcGT detection criteria of one spliced read in over 20% of samples represents a sensitive detection approach. (A) Density plot showing the distribution of the number of spliced reads between the TSS and first genic exon for TcGTs and canonical Ensembl transcripts. The dashed vertical

line represents the median number of spliced reads per category. Canonical ensembl transcripts with zero reads splicing were omitted, along with their corresponding TcGT. This new number is represented by the “n=” in the legend. (B) Bar plot showing the number of samples with a certain number of spliced reads between LTR27E and a genic exon of *IWS1* (LTR27E:*IWS1*) and the developmental stage the LTR27E:*IWS1* TcGT was detected in the Cardoso dataset. (C) Sashimi plot of representative prenatal and adult samples from the short-read RNAseq of DFC. The TcGT is in the antisense orientation. Orange bar highlights the LTR27E TE. (D) Zoom in showing the DNA sequence and splicing junction of LTR27E:*IWS1* highlighting a window of 40bp. The GTF of the consensus transcript generated from the short-read RNA-seq analysis is shown (blue), alongside the PacBio determined transcript (red) from Jeffries et al., 2020.

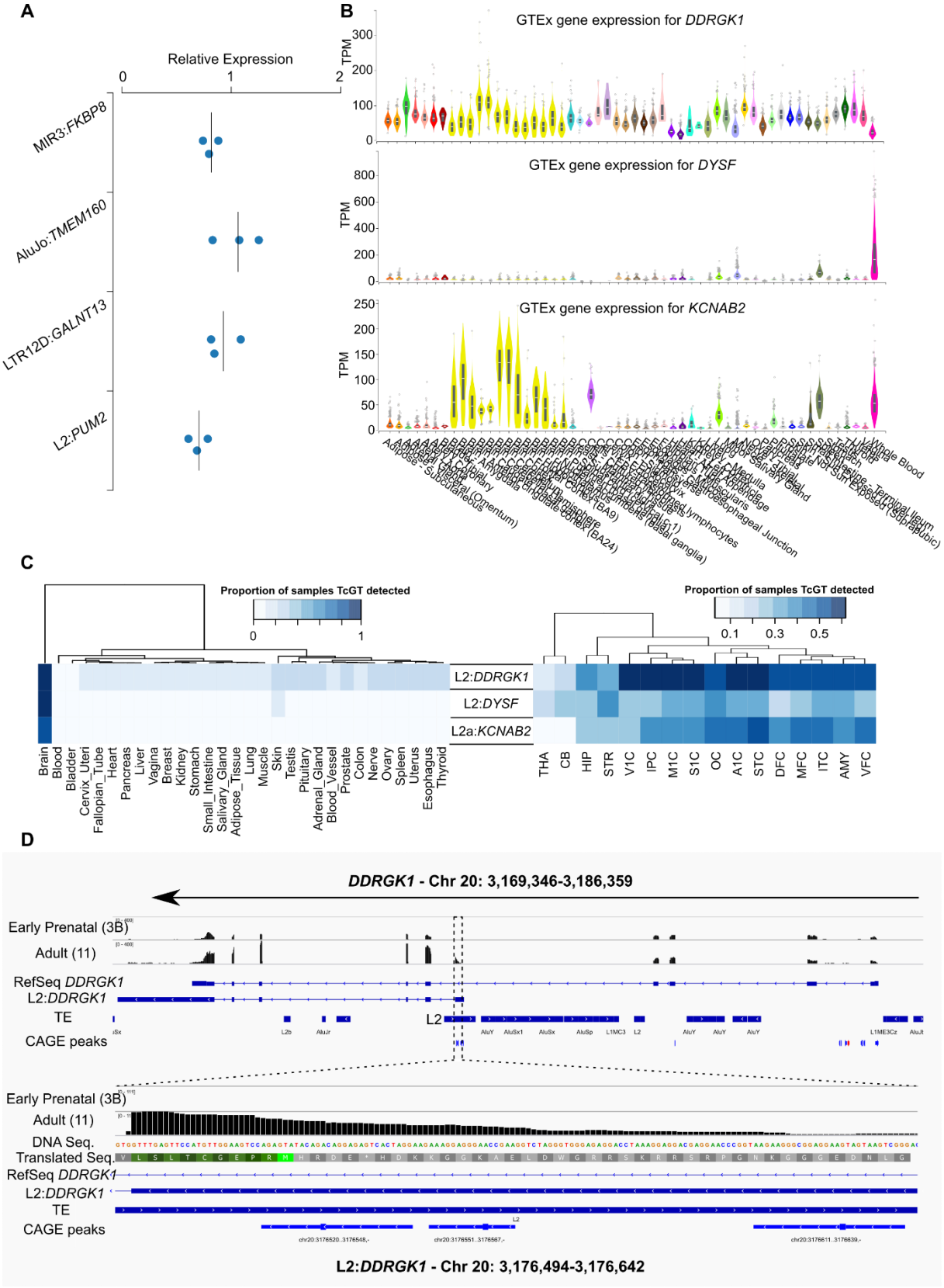


Figure S8. TcGTs are expressed in SH-SY-5Y neuroblastoma cells, are primarily brain specific and L2:*DDRGK1* is a predicted chimeric protein. (A) qRT-PCR expression plots of the indicated TcGTs relative to *ACTB* using primers designed within the TcGT TE TSS and the first exon of the TcGT associated gene. (n=3 independent cultures of SH-SY-5Y cells). (B) GTEx gene expression plots in

transcripts per million (TPM) for *DDRGK1, DYSF* and *KCNAB2.* (C) Heatplot of the proportion of samples the TcGT was detected in tissues from GTEx (left) and brain regions from Brainspan, regardless of stage of detection. (D) Integrated genome viewer (IGV) image of the L2:*DDRGK1* TcGT locus showing representative RNA-seq read pile-ups from early prenatal (stage 3B) and adult (stage 11). Zoom in highlights the DNA sequence and amino acid sequence with the L2 derived start codon highlighted in light green with subsequent peptides in dark green. CAGE peaks are also shown. The gene and TcGT transcript is in the antisense strand orientation (right to left) whereas the L2 element is in the sense strand orientation (left to right).

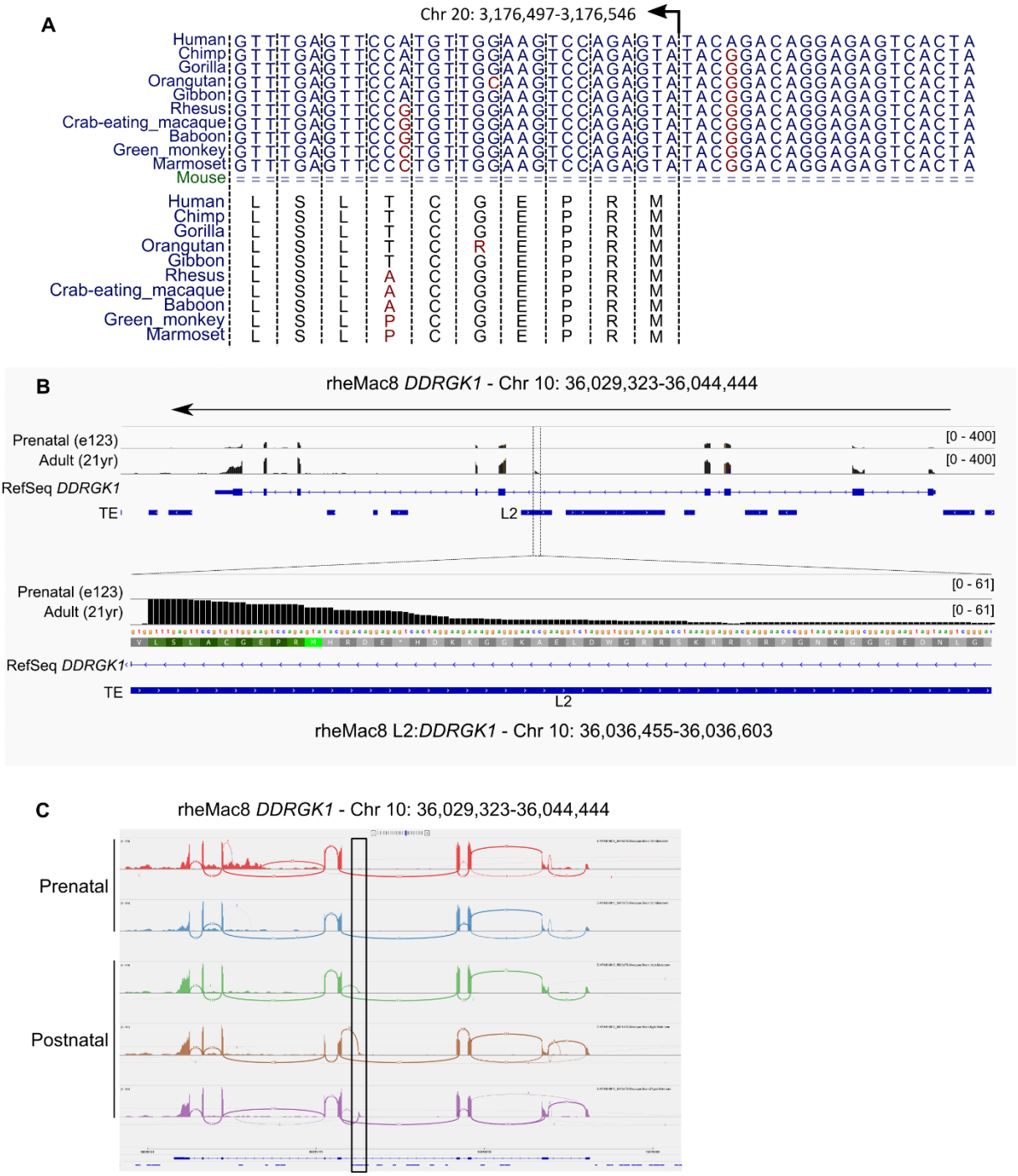


Figure S9. The L2:*DDRGK1* TcGT is conserved in primates and has the same behaviour in macaque as in humans. (A) MULTIZ DNA sequence alignments of the L2 element coding for the chimeric *DDRGK1* TcGT isoform. The methionine start codon is indicated by the square arrow. Some of the L2 5' UTR was omitted for clarity but contained no overt sequence differences. The *in silico* translated product is shown below the DNA sequence, with dashed lines denoting codons. Differences in DNA or protein sequences are highlighted in red. (B) Integrated genome viewer (IGV) image of the rhesus macaque L2:*DDRGK1* TcGT locus showing representative RNA-seq read pile-ups from prenatal (embryonic day 123) and adult (21 years). Zoom in highlights the DNA sequence and translated amino acid sequence with the L2 derived start codon highlighted in light green with subsequent peptides in dark green. The gene and TcGT transcript is in the antisense strand orientation (right to left) whereas the L2 element is in the sense strand orientation (left to right). (C) Sashimi plot indicating splicing events at the

rheMac8 *DDRGK1* locus for prenatal and postnatal samples. Black box indicates the identical L2 element splicing into the same *DDRGK1* genic exon as in humans.

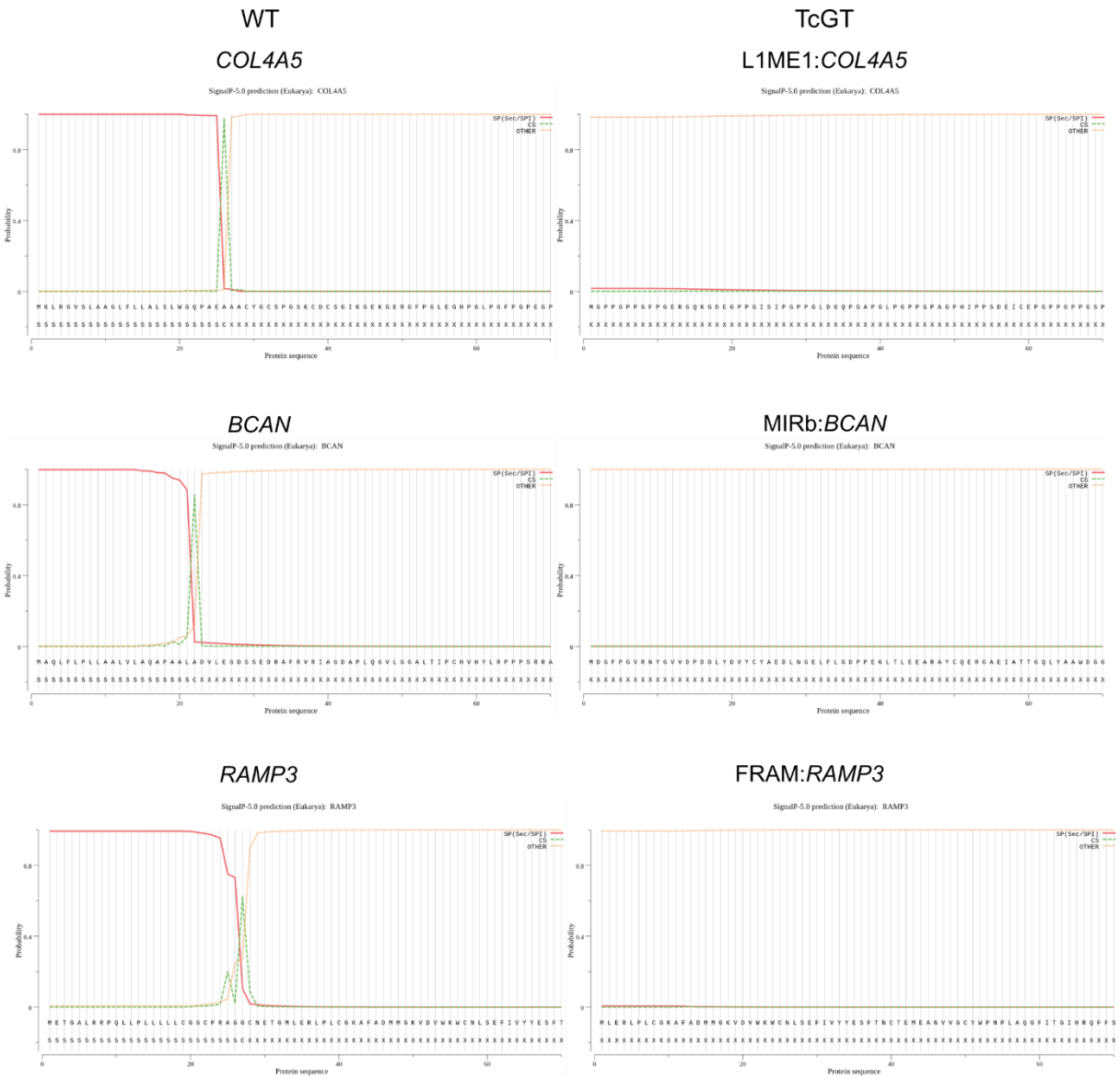


Figure S10. The signal peptide is lost in N-truncated TcGTs. Plots generated from SignalP 5.0 (Almagro Armenteros et al. 2019) showing computationally determined N-terminal signal peptide sequence and cleavage sites for canonical gene transcripts and for consensus N-truncated TcGT derived transcripts. Red line denotes the predicted signal peptide, dashed green line the predicted cleavage site and orange represents non-signal peptide sequence.

Supplemental Table Descriptions

Supplemental Table S1 – Gene expression values in counts per million (CPM) from Brainspan (BS) for all brain regions excluding DFC and CB (see Supplemental Table S2).

Supplemental Table S2 – Gene expression values in counts per million (CPM) for DFC and CB from both Brainspan (BS) and Cardoso (CM) datasets with the Pearson correlation coefficient for the expression of each gene between datasets.

Supplemental Table S3 – Differential gene expression analysis between adult (stage 11) and early prenatal stages (stage 2A to 3B) for DFC and CB in both Brainspan (BS) and Cardoso (CM) datasets.

Supplemental Table S4 –TE subfamily expression values in counts per million (CPM) for DFC and CB from both Brainspan (BS) and Cardoso (CM) datasets.

Supplemental Table S5 – TE subfamily expression values in counts per million (CPM) from Brainspan (BS) for all brain regions excluding DFC and CB (see Supplemental Table S4).

Supplemental Table S6 – Unique TE integrant differential expression analysis between adult (stage 11) and early prenatal (stages 2A to 3B) from both Brainspan (BS) and Cardoso (CM) datasets for DFC and CB. As CM data was stranded, differential expression was calculated for reads mapped in each orientation.

Supplemental Table S7 - Unique TE integrant differential expression analysis between adult (stage 11) and early prenatal (stages 2A to 3B) from Brainspan (BS) for all regions excluding DFC and CB (see Supplemental Table S7).

Supplemental Table S8 – TcGT detection analyses from Brainspan (BS) and Cardoso (CM) datasets. The table shows the detected TcGT hg19 genome co-ordinates for the associated TE and gene. Numbers and percentages per stage represent the number of samples the TcGT was detected in, with at least one spliced read between the TE and a genic exon, regardless of region. Log fold change of detection rates in prenatal versus postnatal stages are provided. Overlap of the TE derived TcGT start site (+/- 200bp) in Ensembl, CAGE and ATAC-seq datasets are indicated. TcGT detection in regions regardless of stage of detection is also provided. *In silico* translation of TcGTs provides the predicted ORF length, nucleotide sequence, amino acid sequence and predicted effect on protein product relative to the canonical protein. TcGT presence in GTEx is also indicated, along with the presence of a signal peptide.

Supplemental Table S9 – Oligonucleotide information

Supplemental Methods

RNA-seq analysis

For genes and TE integrant analysis, only uniquely mapped reads were used for counting on genes and TEs with the command 'featureCounts -t exon -g gene_id -Q 10'. For the Brainspan dataset, samples with less than 10 million unique mapped reads on genes were discarded from the analysis. TEs that did not have at least one sample with 50 reads or overlapped an exon were discarded from the mapping TE integrant analysis. For estimating TE subfamilies expression level, multi-mapping reads were summarized using the command featureCounts -M --fraction -t exon -g gene_id -Q 0 then, for each subfamily, counts on all TE members were added up. As the Cardoso-Moreira et al., 2019 RNA-seq was stranded data, reads on both strands were combined for TEs to facilitate comparison to the non-stranded Brainspan dataset. Normalization for sequencing depth was done for both genes and TEs using the TMM method as implemented in the limma package of Bioconductor (Gentleman et al. 2004) and using the counts on genes as library size. Differential gene expression analysis was performed using voom (Law et al. 2014) as implemented in the limma package of Bioconductor (Gentleman et al. 2004). A gene (or TE) was considered to be differentially expressed when the fold change between groups was greater than two and the p-value was smaller than 0.05. A moderated *t*-test (as implemented in the limma package of R) was used to test significance. P-values were corrected for multiple testing using the Benjamini-Hochberg's method (Benjamini and Hochberg 1995). Temporal expression correlation analyses of individual genes, TE integrants or subfamilies were performed between Brainspan and Cardoso datasets using the 'Pearson' method. For inter-regional correlations within the Brainspan dataset, only expressed genes or TEs common to all regions were considered. BAM files and sashimi plots were visualised using the Integrative Genomics Viewer (Katz et al. 2015; Robinson et al. 2011).

TE and KZFP age estimation

TE subfamily ages were downloaded from Dfam (Hubley et al. 2016). To compare KZFP ages we developed a score we called Complete Alignment of Zinc Finger (CAZF) (Thorball et al. 2020) which rely on the alignments of zinc finger domains, using only the four amino acid presumably touching DNA. Briefly, alignment scores made with BLOSUM80 matrix were used, normalised by the 'perfect' alignment score (alignment against itself) and by the length of the alignment. To compute an age for KZFPs, we relied on inter-species clusters of KZFPs made with CAZF score. KZFPs with CAZF>0.5 were clustered together, using a bottom-up approach. The divergence time between human and the farthest species present in the cluster was used as the age of individual KZFPs in the cluster. MULTIZ alignments for L2:*DDRGK1* locus were extracted from the UCSC Genome Browser.

Protein product prediction

DNA sequences were retrieved for each TcGTs consensus and protein products were derived from the longest ORF in the three reading frames using Biopython (Cock et al. 2009). The resulting translation products were aligned against the protein sequence of the most similar cognate gene isoforms (exons intersect between TcGTs and each gene isoform) and classified into several categories. Proteins with no alignment for any isoform were classified as out-of-frame, therefore not clear or not aligned. In-frame peptides were further classified according to their N-terminal modifications: Normal, TcGT ORF peptides align perfectly with cognate ORF peptides; N-add, TcGT ORF peptides encode novel in-frame N-terminal amino acids followed by the full-length cognate protein sequence; N-truncated, TcGT ORF peptides lack parts of the cognate N-terminal protein sequence and might contain novel in-frame N-terminal amino acids. TcGTs that we could not clearly classify were grouped in the 'other' category, such as TcGTs including C-terminal modifications. If the classification was ambiguous for different protein isoforms, the normal category was always privileged.

CRISPRa

gRNAs were designed with CRISPOR (Concordet and Haeussler 2018) using input DNA sequence 50 to 300bp upstream of the TE resident CAGE peak and the most 5' location of RNA-seq reads mapping to the TcGT TE TSS loci. Multiple gRNAs were selected for each TcGT to control for gRNA specific effects and increase experimental robustness. UCSC BLAT (Kent 2002) analysis of gRNAs confirmed that each was uniquely mapping to their expected target locus. gRNA oligonucleotides were synthesised (Microsynth) with the recommended overhangs (Supplemental Table 9) for integration into the gRNA cloning vector (Mali et al. 2013). gRNA oligonucleotides were annealed and extended using Phusion High Fidelity DNA polymerase master mix (NEB) with thermal cycling conditions of 98^{o}C two minutes (1x), 98^{o}C 10 seconds + 72^{o}C 20 seconds (3x) and 72^{o}C for five minutes. 10µg of SP-dCas9-VPR was digested with Af1II (NEB) in CutSmart buffer for two hours at 37^{o}C, followed by gel electrophoresis and purification of the correct sized band of linearised plasmid with E.Z.N.A Gel Extraction Kit (Omega Bio-tek). The resulting linearised plasmid and double stranded oligonucleotides were ligated using Gibson Assembly Master Mix (NEB) as per manufacturer's recommendations. The resulting gRNA containing plasmid was transformed into HB101 chemically competent *E.coli,* with colonies containing the transformed plasmid selected on agar plates containing kanamycin, followed by colony picking for growth in kanamycin agar broth followed by GeneJET Plasmid Miniprep (Thermo Fisher Scientific). gRNA plasmids were Sanger sequenced to detect the correct insertion of specific gRNA sequences. 300,000 HEK293T cells were seeded per well of a six well plate. 24 hours later, co-transfection was performed with 1µg each of SP-dCas9-VPR and TcGT targeting gRNA containing gRNA cloning vector. SP-dCas9-VPR or empty gRNA cloning vector alone were transfected as non-targeting controls. Cells were harvested for RNA 48 hours post-transfection.

RT-PCR and qRT-PCR

One primer was required to be present in the TE sequence where RNA-seq reads were detected downstream of a CAGE-peak, whilst the other was present in the first or second genic exon. BLAT (Kent

2002) of primer sequences against the human genome ensured only uniquely mapping primers were used. RNA was extracted from cells using the NucleoSpin RNA mini kit (Macherey-Nagel) with on-column deoxyribonuclease treatment. 1µg RNA was used in the cDNA synthesis reaction with the Maxima H minus cDNA synthesis master mix (Thermo Fisher Scientific) and RT-PCR was performed with Phusion High Fidelity DNA polymerase master mix (NEB) each with the manufacturer recommended PCR thermal cycles, on a 9800 Fast Thermal Cycler (Applied Bioscience). PCR products were visualised by 1.5% agarose gel electrophoresis stained with SYBR Safe DNA gel stain (Thermo Fisher Scientific) and imaged with a BioDoc-It imaging system (UVP). Bands of the correct size were excised, gel purified with E.Z.N.A Gel Extraction Kit (Omega Bio-tek) and Sanger sequenced using primers used for PCR. The correct PCR product was confirmed using BLAT (Kent 2002) of the Sanger sequencing results against the human genome (Supplemental Material). qRT-PCR was performed with PowerUp SYBR Green Master Mix on a QuantStudio 6 Flex Real-Time PCR system. The standard curve method was used to quantify expression normalised to *ACTB* with no amplification in the no reverse transcriptase control.

Cloning

L2:*DDRGK1* was PCR amplified with Phusion High Fidelity DNA polymerase master mix (NEB), using cDNA generated in the L2:*DDRGK1* CRISPRa experiment with gRNA 1. This ensured the *bona fide* L2 driven transcript was cloned. Cloning primers used are shown in Supplemental Table 9, with the forward primer containing a CACC Kozak sequence and the reverse primer omitting the stop codon. Thermal cycling conditions were 98^{o}C 30 seconds (1x), 98^{o}C 10 seconds + 60^{o}C 15 seconds + 72^{o}C 15 seconds (35x) and 72^{o}C for 10 minutes. A 466bp PCR fragment was extracted after agarose gel electrophoresis, purified with E.Z.N.A Gel Extraction Kit (Omega Bio-tek), transformed into chemically competent HB101 E.coli, colonies picked and mini-prepped. WT *DDRGK1* and L2:*DDRGK1* in the *pENTR* vectors were then shuttled into *pTRE-3HA* (Imbeault et al. 2017) with the Gateway LR Clonase II Enzyme mix (Thermo Fisher Scientific) as per manufacturer's instructions.

Western blot

20µl of each cellular fraction was used for SDS-PAGE in a NuPAGE 4-12% Bis-TRIS gel and MOPs running buffer (Thermo Fisher Scientific). For subcellular fraction marker proteins, the same amount of lysate was added from each sample but for the HA blot, pTRE-WT:DDRGK1-HA samples were diluted 1:50 due to high over-expression levels compared to pTRE-L2:DDRGK1-HA. Proteins were transferred to a nitrocellulose membrane using an iBLOT 2 dry blotting system (Thermo Fisher Scientific) and analysed by immunoblotting using CANX (Bethyl A303-696A, 1:2000), LMNB1 (Abcam ab16048, 1:1000), β TUBULIN (Sigma-Aldrich T4026, 1:1000), HA-HRP conjugated (Roche 12013819001, 1:2000). HRP-conjugated anti-mouse (GE Healthcare NA931V, 1:10000) and HRP-conjugated anti-rabbit (Santa Cruz sc-2004 1:5000) antibodies were used where appropriate and the blot was visualised using the Fusion SOLO S (Vilber).

Immunofluorescence

HEK293T cells were plated on glass coverslips and immunofluorescence was performed as previously described (Helleboid et al. 2019) 48 hours post-transfection and expression induction with 1µg/ml doxycycline for *pTRE-WT:DDRGK1-HA* or *pTRE-L2:DDRGK1-HA*. Once 70% confluent, cells were washed three times with PBS, fixed in ice-cold methanol for 20 minutes at -20^{o}C then washed three more times with PBS. Cells were blocked with 1% BSA/PBS for 30 minutes and then incubated with antibodies for HA.11 (BioLegend MMS-101P, 1:2000) and HSPA5 (Abcam ab21685, 1:1000) in 1% BSA/PBS for one hour. Three washes with PBS were performed, followed by incubation with anti-mouse and anti-rabbit Alexa 488 or 568 (Thermo Fisher Scientific 1:800) for one hour. DAPI (1:10000) was added in the last 10 minutes of incubation, samples washed three times with PBS and coverslips mounted on slides with ProLong Gold Antifade Mountant (Thermo Fisher Scientific). Images were acquired on a SP8 upright confocal microscope (Leica) and processed in ImageJ.

Supplemental Acknowledgements

Submission of Brainspan data (phs000406.v2.p1) to dbGaP was provided by Dr. Nenad Sestan. Collection of the data and analysis was supported by grants from the National Institutes of Health (MH089929, MH081896, and MH090047). Additional support was provided by the Kavli Foundation, a James S. McDonnell Foundation Scholar Award, NARSAD, and the Foster-Davis Foundation.

The Genotype-Tissue Expression (GTEx) Project (phs000424.v7.p2) was supported by the Common Fund of the Office of the Director of the National Institutes of Health (commonfund.nih.gov/GTEx). Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI\Leidos Biomedical Research, Inc. subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to The Broad Institute, Inc. Biorepository operations were funded through a Leidos Biomedical Research, Inc. subcontract to Van Andel Research Institute (10ST1035). Additional data repository and project management were provided by Leidos Biomedical Research, Inc. (HHSN261200800001E). The Brain Bank was supported supplements to University of Miami grant DA006227. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951, MH090937, MH101825, & MH101820), the University of North Carolina - Chapel Hill (MH090936), North Carolina State University (MH101819), Harvard University (MH090948), Stanford University (MH101782), Washington University (MH101810), and to the University of Pennsylvania (MH101822)

Supplemental References

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *J R Stat Soc* **57**: 289–300.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**: 1422–1423.

Concordet J-P, Haeussler M. 2018. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res* **46**: W242–W245.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**: R80.

Helleboid P, Heusel M, Duc J, Piot C, Thorball CW, Coluccio A, Pontis J, Imbeault M, Turelli P, Aebersold R, et al. 2019. The interactome of <scp>KRAB</scp> zinc finger proteins reveals the evolutionary history of their functional diversification. *EMBO J* **38**: 1–16.

Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AFA, Wheeler TJ. 2016. The Dfam database of repetitive DNA families. *Nucleic Acids Res* **44**: D81–D89.

Imbeault M, Helleboid P-Y, Trono D. 2017. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**: 550–554.

Katz Y, Wang ET, Silterra J, Schwartz S, Wong B, Thorvaldsdóttir H, Robinson JT, Mesirov JP, Airoldi EM, Burge CB. 2015. Quantitative visualization of alternative exon expression from RNA-seq data. *Bioinformatics* **31**: 2400–2402.

Kent WJ. 2002. BLAT---The BLAST-Like Alignment Tool. *Genome Res* **12**: 656–664.

Law CW, Chen Y, Shi W, Smyth GK. 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**: R29.

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. 2013. RNA-Guided Human Genome Engineering via Cas9. *Science* **339**: 823–826.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24–26.

Thorball CW, Planet E, de Tribolet-Hardy J, Coudray A, Fellay J, Turelli P, Trono D. 2020. Ongoing evolution of KRAB zinc finger protein-coding genes in modern humans. *bioRxiv*