A retroviral origin of vertebrate myelin

Tanay Ghosh[1,2]*, Rafael G. Almeida[3], Chao Zhao[1,2], Ginez Gonzalez M[1,2], Katherine Stott[4], Ian Adams[5], David H Rowitch[1,6], Nick Goldman[7], David A. Lyons[3], Robin JM Franklin[1,2]*

Correspondence to:
tg369@cam.ac.uk
rjf1000@cam.ac.uk

This PDF file includes:

Materials and Methods
References (28-61)
Figs. S1 to S6
Tables S1

Materials and Methods

Retrotransposable element annotation from Affymetrix microarray

Retrotransposon probes in Affymetrix Rat Genome 230 2.0 were identified and annotated as described previously (28). Affymetrix Rat Genome 230 2.0 contains 31,000 probesets. Probe sequence were used for BLAT search in the rat genome (rn4, Nov. 2004, version 3.4) and multimapping probes were identified. Genomic co-ordinates of these probes were then compared with Repeatmasked region (rn4 release rat genome) by querying repeat masker database (29) and annotation (class, family and element) were tabulated.

Probe level analyses of Affymetrix microarray data

Raw data with accession number GSE11218 and GSE5940 (using Affymetrix expression GeneChip Rat Genome 230 2.0 and 230A, respectively) were extracted from Gene expression omnibus (GEO). Raw data from both the platform comprise 9 OPC and 8 OL samples. Probe level information was extracted using R library rat2302probe (for Rat genome 230 2.0) and rae230aprobe (for Rat genome 230A). Unique probe sequence in 230 2.0 and 230A platform were retained and the intersect probes (90507 in number) between two arrays were used for analysis. Background subtracted raw data were quantile normalized. Probes for which intensity were greater than sample median level in at least half the arrays for one experimental condition in a dataset were considered present (28); absent probes were removed. Further differentially expressed probes between OL and OPC were identified by using R Bioconductor package limma (30). P-values were adjusted to correct for multiple testing using FDR (Benjamini-Hochberg). Probes for which the difference between groups were $\geq$ 2 fold and the adjusted p-value < 0.05 were considered differentially expressed.

Weighted gene co-expression network analysis (WGCNA)

WGCNA was performed by using functions (31) available as a R package (1.20 library) installer. Because each gene is represented by multiple probes in microarray, aggregate (32) function was used to collapse probe with median expression. Pearson correlation matrix (preserving the sign of the correlation coefficient) of pairwise comparison of gene expression levels was generated and transformed to an adjacency matrix by using soft threshold power=16 for fitting with approximate scale free topology; thereafter a topological overlap matrix (TOM) was generated. Topological overlap (TO) based dissimilarity (1-TO) measure was used for generating average linkage hierarchical clustering and cutreeDynamic function was used to identify modules. Module eigengene (ME) (i.e. 1st principal component obtained after singular value decomposition) is a summary of the gene expression levels of all genes in a module. Binary numeric variables called 'oligodendrocytes' was generated to define OPC (=0) and OL (=1). Correlation co-efficients (Pearson) between MEs with oligodendrocytes and p-values (Student asymptotic) of corresponding correlations were obtained by using WGCNA functions. Bonferroni threshold level of significance (0.01/14 = 0.00071) was set based on the number of modules (=14).

Overrepresentation analysis

DAVID (the Database for Annotation, Visualization and Integrated Discovery) (33) functional annotation tool was used for analyzing GO (Gene ontology) terms overrepresented in WGCNA modules. Benjamini-Hochberg adjusted p value <0.05 was considered significant.

TF binding site

Conserved SOX10 binding motif mapping in RNLTR12-int consensus sequence were analyzed using JASPAR V 5.0_ALPHA (34) and the PWM (Position Weighted Matrices) scores were calculated. The p values corresponding to the PWM scores were obtained by using TFM-PVALUE algorithm (35).

Coding potential determination

Coding and non-coding potential was evaluated using CNIT (Coding-Non-Coding Identifying Tool) program (36).

RNA-seq data analysis

RNAseq raw data from OPC (GSM3967160) were pre-processed through sortMeRNA (37) to filter out reads from ribosomal RNA, adapter was trimmed, and quality paired reads were extracted using Trimmomatics (38). RNLTR12-int consensus sequence (from Repbase-GIRI) was indexed and reads alignment to RNLTR12-int was performed by STAR_2.5.2b (39). Subsequently sorted bam file was generated, and depth was calculated using samtools (40).

Identification of RNLTR12-int like sequence (Retromyelin)

We employed the remote homology search algorithm nhmmer (41), probabilistic inference method based on a hidden Markov models, and used RNLTR12-int consensus sequence (Repbase-GIRI) as an input to search *Retromyelin* in other genomes. E-value threshold for inclusion was <0.01. Co-ordinate of the top hits were then queried to the RepeatMasker (29) to extract repeat annotation.

Also, RepeatMasker (29) was run (search engine: rmblast/hmmer/cross_match and selecting model organism as a DNA source) independently using the sequence of the top hits as an input to identify as well as to verify repeat type and family. Identified repeats which are just a simple repeat (e.g. $(TC)_n$) are not considered as *Retromyelin* (**Table S1**).

Taxonomic common tree display

Using taxonomic ID of the subset of organisms as an input and by employing common tree tool from NCBI, a hierarchical representation of the relationships among the taxa and their lineages was generated. Phylip tree file was visualized by iTOL (42).

Phylogenetic analysis

Top hit sequence of the identified *Retromyelin* in different animals were used for phylogenetic analysis (**Table S1**). MAFFT version 7 (a fast Fourier transform based multiple sequence alignment program) (43) with an iterative refinement strategy, E-INS-I, is employed to align *Retromyelin*. This alignment was curated by trimAI algorithm (44) and then used as an input for phylogenetic analysis using PhyML 3.1 (45), a maximum-likelihood based algorithm. Generalised time reversible (GTR) substitution model with empirical equilibrium frequency and four gamma-distributed rate categories were used. NNI (Nearest Neighbor Interchange) algorithm was used for tree topology search (46, 47). Approximate Bayes (aBays) statistical test was used for evaluation of branch support (45). Newick-format tree files were visualized in iTOL (42).

Rate heterogeneity analysis was performed using BASEML (in paml version 4.8a) (48). We used the REV+G (GTR+G) model with the assumption of the variation of evolutionary rates over sites

following a gamma distribution. The parameters in the rate matrix (REV) were as described (49, 50). The shape parameter (α) of the gamma distribution of rates over sites were estimated. H0 (no rate heterogeneity i.e., $\alpha \to \infty$) was compared with H1 (gamma distributed rate heterogeneity) using a likelihood ratio test and the 2δ test statistics as described (23). Following rejection of H0, site wise empirical Bayesian (posterior mean) relative rate estimates were computed as described (51).

EST mapping

Input nucleotide sequences were individually queried to expressed sequence tags (ESTs) database by using NCBI blastn suit. ESTs for which $0 \leq$ E-value $\leq$ 6e-26 and $119 \leq$ alignment score ≤ 2298, were mapped to the RNLTR12-int like sequences from the following animals: consensus sequence of Human (Dfam ID: DF0000205.5), mouse (Dfam ID: DF0001916.1), rat (RNLTR12-int from Repbase), zebrafish (Dfam ID: DF0002922.2). Consensus sequence was not available for elephant sharks, so we mapped to the genomic location obtained after nhmmer i.e. KI636671.1: 5058-5765 (Callorhinchus_milii-6.1.3).

Rats

Sprague–Dawley rats were received from C. River (Margate, UK). Rats were maintained in individually vented cages at temperature 22^{o}C ± 1^{o}C and humidity 60% ± 5%, in a 12 hr light:dark cycle; food and water were supplied ad libitum in a standard facility for rodents at the University of Cambridge, UK. All animal studies were conducted under the Animals (Scientific Procedures) Act 1986 Amendment Regulations 2012 following ethical review by the University of Cambridge Animal Welfare and Ethical Review Body (AWERB).

Isolation of OPC and OL

Postnatal day 7 (P7) rats were sacrificed using an overdose of pentobarbital and whole brain was dissected immediately in ice cold hibernate A low fluorescence (HALF) medium (52, 53). Subsequently OPC was isolated as described earlier (52). Per 10 million cell suspension, we used 2.5 µg mouse-anti-rat-A2B5-IgM antibody (Millipore; MAB312) and 20 µl of rat-anti-mouse-IgM antibody magnetic beads (Miltenyi, 130-047-302) for OPC isolation by magnetic-activated cell sorting (MACS). OLs were isolation from A2B5 negative fraction by using 2.5 µg of goat-anti-mouse-MOG-Biotinylated (RD systems, BAF2439) and 20µl mouse-anti-biotin magnetic micro beads (Miltenyi, 130-105-637).

Isolated OPC/OL were immediately placed in cell lysis buffer from mirVana (Ambion) RNA isolation kit. Otherwise, OPCs were seeded in 24-well-plate or 6-well-plate with poly-d-lysine (PDL) coated wells.

Culture and differentiation of OPC

OPCs were maintained in OPC medium (52, 53) [supplemented with 30 ng ml^{-1} basic fibroblast growth factor (bFGF) (Peprotech, 100-18B) and 30 ng ml^{-1} platelet-derived growth factor (PDGF) (Peprotech, 100-13A)] and 60µg ml^{-1} N-Acetyl cysteine (Sigma, A7250) in an incubator at 37°C and 5% CO_2 and 5% O_2. Medium was replaced in alternative days. OPC medium (100 ml) composition: A total volume of 100 ml of Dulbecco's modified Eagle's medium (DMEM)/F12 (Thermo Fisher, 11039-021) is supplemented with 1mM sodium pyruvate (Thermo Fisher, 11360-070), 5 mg apo-transferrin (Sigma, T2036), 1 ml SATO stock solution [described earlier (52)], 10µg ml^{-1} human recombinant insulin (GIBCO, 12585014).

For differentiation of OPC to OL: OPC medium was supplemented with 40 ng ml^{-1} 3,3′,5-Triiodo-L-thyronine (T3) (Sigma; T6397) (52, 53). Medium was replaced on alternate days for 5 days.

Transfection

Lipofectamine™ RNAiMAX transfection reagent (ThermoFisher scientific) was used and followed the manufacturer protocol. 48 hours before transfection MACS isolated OPC (approx. 25,000) were seeded per well of a 24-well-plate (PDL coated) and cells were maintained in proliferation media. Before transfection media was replaced by differentiating media. 1 ul Lipofectamine™ RNAiMAX and 30pmol siRNA was used for transfection. siRNA against RNLTR12-int (Dharmacon) and control siRNA (siGENOME Non-Targeting siRNA pool #2) were procured from Dharmacon. 48 hours after transfection cells were harvested for RNA isolation. 5 days after transfection cells were fixed for immunofluorescence analyses.

siRNA sequence against RNLTR12-int: 5' AAGUGAGGGCCUUCUAUGCUU 3'

shmiR cloning and AAV PHP.eB packaging

SOX10 Multiple Species Conserved enhancer element 5 conjugated with cfos basal promoter (https://www.addgene.org/115783/, 54) sequence was used for oligodendrocyte lineage specific expression of EGFP-shmiR with a bGHp(A) poly A signal (**Fig. S5**). This SOX10 driven entire sequence (**Fig. S5**) was cloned between the inverted terminal repeat of adeno associated virus 2 (AAV2ITR) vector. Cloning and large scale AAV PHP.eB packaging of the construct and purification was performed from AMS Biotechnology (Europe) Limited, UK. The viral tire was 10^{13} $GCml^{-1}$. The following sequences were embedded into the miR-30 backbone of the shmiR vector:

shmiR-RNLTR12-int:

5'ATAGAAGGCCCTCACTTTTA[GTGAAGCCACAGATG]TAAAAGTGAGGGCCTTCTATGC 3’

shmiR-Scrambled:

5’ TGGACGCTCGATTGATCATA[GTGAAGCCACAGATG]TATGATCAATCGAGCGTCCAAT 3’

The loop sequence is indicated by parenthesis.

In vitro and *in vivo* infection

For *in vitro* study, OPC was isolated from rat pups using MACS (above) and approximately 30,000 cells were seeded to each well of a 24-well-plate, 48 hours before infection. Media was replaced to differentiating media (above) before infection. Approximately 5×10^{10} GC of virus was used per well. 4 days after infection total RNA was isolated using the mirVana kit (Ambion), DNase I (ThermoFisher scientific) treated before reverse transcription.

For *in vivo* experiment, 1µl of virus suspension (approximately 1×10^{10} GC) was injected stereotaxically into the cerebral cortex at 0.1mm anterior, 1.2mm lateral and 1.2mm deep with reference to Bregma under isoflurane anaesthesia, in neonatal rats at P1.

Immunofluorescence staining

Cells were fixed with 4% PFA (10 min, room temperature), washed with PBS and proceeded for immunofluorescence staining as described (53). For *in vivo* experiment, animals (at P14) were perfused using 4% PFA in PBS, brains were postfixed overnight with 4% PFA, cryoprotected with

20% sucrose, and embedded in OCT-medium (TissueTek). Immunofluorescence was performed on 12 µm cryostat section and closely followed the protocol as described earlier (55).

The following antibodies were used in this study:

Primary antibody	**Class**	**Host species**	**Dilution**	**Source**	**Catalogue number**
Anti-MBP	IgG	Rat	1 to 500 or 1 to 100	Serotec	MCA4095
Anti-OLIG2	IgG	Rabbit	1 to 1000 or 1 to 100	Millipore	AB9610
Anti-GFAP	IgY	Chicken	1 to 1000	Abcam	ab4674
Anti-O4	IgM	Mouse	1 to 1000	R&D Systems	MAB1326
Anti-CC1	IgG2b	Mouse	1 to 100	Merck/milipore	OP80
Anti-GFP	IgY	Chicken	1 to 100	Abcam	ab13970
Alexa Fluor 488 Anti-Mouse	IgM	Goat	1 to 1000	Invitrogen	A21042
Cy™3 AffiniPure Anti-Rat	IgG (H+L)	Donkey	1 to 1000 or 1 to 500	Jackson Immuno Research	712-165-150
Alexa Fluor 647 Anti-Rabbit	IgG	Donkey	1 to 1000 or 1 to 500	Invitrogen	A31573
Alexa Fluor 488 Anti-Chicken	IgG (H+L)	Goat	1 to 1000 or 1 to 500	Invitrogen	A11039
Alexa Fluor-568-Anti-mouse	IgG2b	Goat	1 to 500	Invitrogen	A21144

Images were acquired with an SP5 (Leica TCS) or SP8 (Leica TCS) confocal microscope. All acquisition settings were kept constant for the whole experiment. Images were processed using the ImageJ version 2.1.0 software (56). For *in vivo* immunoassayed sections, integrated density of MBP immunostaining was measured using ImageJ 2.1.0. Cell counting was performed using Cell Counter tool in ImageJ 2.1.0.

qRT-PCR analysis

Total RNA was isolated using the mirVana kit (Ambion). Total RNA was DNase I (ThermoFisher scientific) treated before reverse transcription. cDNA was generated by using random hexamer (Invitrogen) and SuperScript™ IV Reverse Transcriptase (Invitrogen™) following manufacturer protocol. Real time PCR was performed using SYBR green master mix (Applied Biosystems, ABI) in QuantStudio 7 Flex (Applied Biosystems, ABI) machine. The real time PCR program was 50°C/ 2 min, 95°C / 10 min., 40 cycles 95°C/ 15 s, 60°C/ 1min. Dissociation protocol was run.

Rat specific primer sequences used in qPCR:

Primer	**Sequence (5' to 3')**	**Amplicon (bp)**
RNLTR12-int	F: CCACTGAGGAAAGACGGAAT	125
	R: ACCTGGAGCACTCCCTACCT	
Actb	F: CACCATGTACCCAGGCATTG	111
	R: CACACAGAGTACTTGCGCTC	
Mbp	F: ATCGGCTCACAAGGGATTCA	108
	R: CGTCTTGCCATGGGAGATCC	

Tspan2	F: TTAAGCTCCAGCTCATTGGAA	101
	R: TGAGTTCCGTATTGCACAGC	
Mag	F: CCTGGCAGAGAATGCCTATG	115
	R: GACTGTCTCCCCCTCTACCG	
PB1D9	F: CAAGAGGCAGAGGCAAACAC	81
	R:TTTGGTTTTTAGATGTATAAGGTCTCA	
Pdgfra	F: AAGATCTGTGACTTCGGGCT	106
	R: AAATGCTCTCAGGTGCCATC	
CNP1	F: GTGCTGCACTGTACAACCAA	111
	R: GGACAGTTTGAAGGCCTTGC	

RNLTR12-int sequence was located in multiple chromosomes which can be identified using Affymetrix probe set (1379497_at). We did BLAT (from UCSC) and designed the primers from the top hit (chr7:82148213-82148189, assembly: rn5), which was one among many that showed 100% identity. Since RNLTR12-int is intron-less, we excluded the possibility of trace genomic DNA (gDNA) in our RNA samples by treating with DNase I and designing intron flanking primers targeting cytoplasmic beta actin (Actb) gene to detect gDNA contamination. We observed only cDNA specific expected size amplicon band (approx. 125 bp) in our OPC and OL samples (**Fig. S2A**). Further melting curve analysis by quantitative PCR (qPCR) yielded only single peak, suggesting a single amplicon (**Fig. S2A**), confirming that there was no gDNA contamination.

Actb was used as an endogenous gene for normalization control. We used Actb as a normalization control because in our microarray data analysis there was no expression difference of this gene between OPCs and OLs was observed (**Fig. S2 B**). Relative quantification of qPCR data was analyzed by Paffl method (57). PCR efficiency of the target gene and the endogenous control was determined from the slope of the respective standard curve.

RNA immunoprecipitation (RIP)

Crosslinking RNA immunoprecipitation was performed by following standard Abcam protocol. 5-6 Sprague–Dawley rat brains (P7) were used. Meninges were removed and cellular dissociation (52) was performed up to 33% Percoll step to remove debris and fat. Approximately 4-6 X 10^7 cells were obtained, re-suspended in Hank's balanced salt solution (HBSS) and immediately cross-linked with formaldehyde (with a final concentration 1%) for 10 min at room temperature. Crosslinking was stopped with glycine (final concentration 0.125M) and subsequently washed with PBS. Nuclei was isolated in nuclei isolation buffer and nuclear pellet was re-suspended in RIP buffer. Sonication (30 s on, 30 s off, 10 min) was performed using Bioruptor (Diagenode). Pre-cleaning of lysate was carried out using blocked Protein G sepharose (ab193259, Abcam). Blocking of Protein G sepharose was carried out by using 500ng/ul yeast tRNA and 1mg/ml RNase-free BSA. A total of 45% (450 ul) of pre-cleaned lysate was used for anti-SOX10 (sc-365692, Santa Cruz) or anti-immunoglobulin G1 (IgG1) (sc-3877, Santa Cruz) samples, and 10% was used for input sample. A total of 4 ug antibody (SOX10 or immunoglobulin G1) was used for immunoprecipitation, and blocked Protein G sepharose was used for the pull-down step after immunoprecipitation. Eluted solution was de-crosslinked and proteinase K treated, DNAase

treatment (TurboDNAse) was performed and RNA was isolated. RNA samples were analyzed by qPCR.

Chromatin-immuno-precipitation (ChIP)

Four P7 rat brains were pooled each time for OPC isolation and approximately 2×10^5 cells were seeded per well of a six-well-plate (PDL coated). Transfection were performed as mentioned above. 4 days after transfection, cells were cross linked with 16% HCHO added per well onto the media (final concentration 1%) and kept for 10 min and subsequently followed the standard ChIP protocol (58). After fixation and quenching by glycine (final concentration: 0.125 M), cells were dissociated from PDL-coated wells using TrypLE Express (Thermo Fisher; 12604013).

Sonication (30 s on, 30 s off, 30 min) (59) was performed using a bioruptor (Diagenode) and the fragment size 200–500 bp was verified. Protein G Sepharose beads were blocked using 1 mg/ml BSA (stock 10 mg/ml). As described in RIP method (above), lysate was pre-cleaned and immunoprecipitation using antibody (anti SOX10 or IgG1) and blocked Protein G sepharose was used for the pull-down step after immunoprecipitation. Washing and preparation of DNA was performed as described (58). Precipitated DNA samples were analysed by PCR and qPCR. The following primers were used for PCR after ChIP:

Mbp promoter:

FP: 5’ CATTGTTGTTGCAGGGGAGG 3’

RP: 5’ GCTCGTCGGACTCTGAGG 3’

Cloning, *in vitro* transcription (IVT) and labelling of RNA

Consensus sequence of RNLTR12-int (obtained from Repbase-Giri: https://www.girinst.org/repbase/) was cloned in pcDNA3.1 (+) using the restriction site NheI/XbaI.

Purified linearized plasmid (from above) was used as a template and transcribed *in vitro* using HiScribe™ T7 ARCA mRNA Kit (with tailing) (New England Biolabs) following manufacturer protocol (E2060) and purified using Direct-zol^{TM} RNA Miniprep kit (Zymo Research).

For surface plasmon resonance control, the following RNA was synthesized from Dharmacon:

5' CCUGAUUUUUAAGGAAUAUCGCAAGAAUGCCGCGAAUGAAAAA 3'

RNA was 3' labelled with biotin using Pierce™ RNA 3' End Biotinylation Kit (ThermoFisher) and purified using Monarch® RNA Cleanup Kit (New England Biolabs).

Surface plasmon resonance (SPR) experiment

The SPR experiment was performed on a Biacore T200 (GE Healthcare/Cytiva) using a Series S Sensor Chip SA (Cytiva) following the manufacturer's instructions. Tris Buffered Saline (TBS) pH7.4, supplemented with 0.002% Tween20, 100 mM glycerol and 5% glycine, was used as the running buffer throughout. Immobilization was achieved by flowing approximately 5 ugml^{-1} biotinylated RNA for 300 seconds at a flow rate of 2 ulmin^{-1}. Recombinant Human SOX10 protein (Euprotein) at a concentration of 2 uM was then flowed as analyte for 60 seconds at a flow rate of 30 ulmin^{-1}. Data were analysed using the inbuilt BIAEval software (Cytiva).

Zebrafish experiment

Zebrafish were maintained in the University of Edinburgh BVS Aquatics facility under project license PP5258250. To induce mutations in *MyRetro* loci in the zebrafish genome, we designed a guide crRNA targeting the sequence TAATGAAGCAATCAAACAAG*TGG* (PAM sequence italicized), which is conserved (**Fig. 4D**) in the top 10 hit loci most similar to the RNLTR12-int. Ribonucleoprotein complex (RNP) were prepared by annealing 20µM crRNA with 20µM tracrRNA (IDT DNA) at 95°C, and incubating 1.6µM of annealed gRNA with 2µM Engen Cas9 enzyme (New England Biolabs) at 37°C for 10min. 1nL RNP or 1nL of Cas9-only control solution were microinjected into fertilized one-cell stage eggs of the transgenic myelin reporter line Tg(mbp:EGFP-CAAX)ue2 (60). Embryos were kept at 28.5°C in 10mM HEPES-buffered E3 Embryo medium. At 5dpf, larvae were anesthetized in 0.16 mg/mL tricaine (ethyl 3-aminobenzoate methanesulfonate salt, Sigma-Aldrich) and immobilised in 1.5% low melting-point agarose on their sides. Animals were imaged using a Zeiss LSM880 confocal with Airyscan in Fast mode, a Zeiss Plan-Apochromat 20x/0.8 NA objective, and a 488nm laser. The urogenital opening was used as a landmark to consistently image the same anterior-posterior position in every larva. An optimally sectioned z-stack comprising the whole depth of the spinal cord (as determined by the EGFP-CAAX signal) was acquired, maintaining z-step, frame size, laser power and gain settings throughout to enable fluorescence intensity comparisons between animals. For image processing, maximum intensity projections were created for all samples and the integrated densities were obtained using Fiji/ImageJ.

Statistical analysis

R (version 4.0.2) (http://www.R-project.org) or GraphPad Prism (version 8.4.3) was used for statistical analysis. Shapiro-Wilk test (61) was performed to test normality. Welch's t test was performed whenever heteroscedasticity existed in the data set.

References

28. J Reichmann et al., PLoS Comput Biol. 8, e1002486 (2012).

29. AFA Smit, R Hubley and P Green, RepeatMasker Open-4.0. 2013-2015 http://www.repeatmasker.org.

30. Ritchie ME et al., Nucleic Acids Research 43, e47 (2015).

31. Langfelder P, Horvath S, BMC Bioinformatics 29, 559 (2008).

32. Becker RA et al., The New S Language. Wadsworth & Brooks/Cole (1988).

33. W da Huang et al., Nat Protoc. 4, 44-57 (2009).

34. A Mathelier et al., Nucleic Acids Res. 42, D142–7 (2014).

35. H Touzet, JS Varré, Algorithms Mol Biol. 11,15 (2007).

36. JC Guo et al., Nucleic Acids Res. 47, W516-W522 (2019).

37. E Kopylova et al., Bioinformatics 28, 3211-3217 (2012).

38. AM Bolger et al., Bioinformatics 30, 2114–2120 (2014).

39. A Doblin et al., Bioinformatics 29, 15-21 (2013).

40. Li Heng et al., Bioinformatics 25, 2078-9 (2009).

41. TJ Wheeler, SR Eddy, Bioinformatics 29, 2487-9 (2013).

42. I Letunic et al., Nucleic Acids Res. 47, W256-W259 (2019).

43. K Katoh and DM Standley, Mol Biol Evol. 30, 772-80 (2013).

44. S Capella-Gutiérrez et al., Bioinformatics 25,1972-3 (2009).

45. I Guindon et al., Syst Biol. 59, 307-21 (2010).

46. DF Robinson, J Comb Theory Ser B 11, 105–119 (1971).

47. GW Moore et al., J Theor Biol. 38, 423-57 (1973).

48. Z Yang, Comput Appl Biosci. 13, 555-556 (1997).

49. Z Yang, J Mol Evol. 39, 105-111 (1994).

50. Z Yang, J Mol Evol. 39, 306-314 (1994).

51. Yang Z, Wang T. Biometrics. 51:552-61 (1995).

52. M Segel et al., Nature 573, 130–134 (2019).

53. B Neumann et al., Cell Stem Cell 25, 473-485 (2019).

54. SU Pol et al., Exp Neurol., 247, 694-702 (2013).

55. KS Rawji et al., J Neurosci. 38, 1973-1988 (2018).

56. J Schindelin et al., Nat Methods. 9:676-82 (2012).

57. MW Pfaffl, Nucleic Acids Res. 29, e45 (2001).

58. AS Weinmann, PJ Farnham, Methods 26, 37-47 (2002).

59. T Ghosh et al., Cell Rep. 7, 1779-88 (2014).

60. RG Almeida et al., Development 138, 4443–4450 (2011).

61. JP Royston, Appl. Stat. 181, 176–180 (1982).

Fig S1

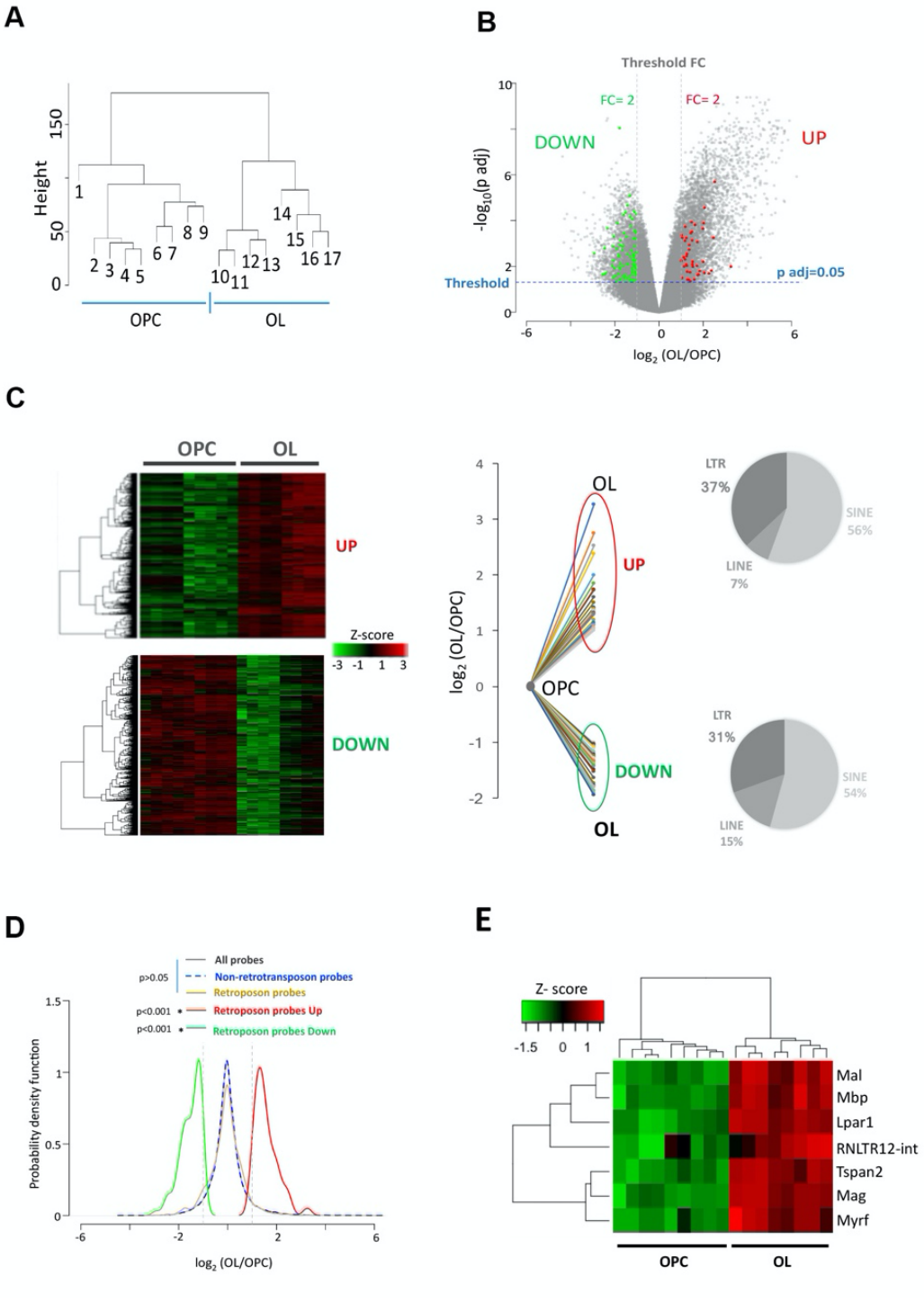
A
Height
150
50
0
1
2 3 4 5
6 7
8 9
10 11
12 13
14
15
16 17
OPC
OL
B
Threshold FC
FC= 2
FC= 2
DOWN
UP
-log10(p adj)
Threshold
p adj=0.05
log2 (OL/OPC)
C
OPC
OL
UP
DOWN
Z-score
-3 -1 1 3
log2 (OL/OPC)
OL
UP
OPC
DOWN
OL
LTR
37%
SINE
56%
LINE
7%
LTR
31%
SINE
54%
LINE
15%
D
All probes
p>0.05
Non-retrotransposon probes
Retroposon probes
p<0.001 *
Retroposon probes Up
p<0.001 *
Retroposon probes Down
Probability density function
log2 (OL/OPC)
E
Z- score
-1.5 0 1
Mal
Mbp
Lpar1
RNLTR12-int
Tspan2
Mag
Myrf
OPC
OL

Fig. S1. Genome wide expression of retrotransposons in OPC and OL. (**A**) Dendrogram representation of OPC and OL samples, showing OPC and OL samples were formed two separate clusters. Average linkage hierarchical clustering (using Euclidean distance matrix) was performed with respect to the intensity levels of probes in Affymetrix GeneChip. (**B**) Probe level analysis of differential expression. Differentially expressed retrotransposons probes (red and green dots) and probes for protein coding genes (grey dots). Probes for which the expression difference between OL and OPC is 2-fold up or down and the adjusted p value<0.05 were considered differentially expressed. FC: Fold change. Padj: adjusted p value (Benjamini Hochberg FDR). (**C**) Left: Heatmap plot and dendrogram representation of all differentially expressed probes in OL and OPC samples. Right: Differential gene expression (log_2 value) of retrotransposon probes were plotted. UP: induced probes, DOWN: repressed probes. Proportion of different retrotransposon type in induced and repressed probes were plotted in a pie chart. (**D**) Probability density plots showing the behavior of retrotransposon probes, non-retrotransposon probes, all probes and differentially expressed retrotransposon probes. Vertical dotted lines: 2-fold change of expression. * $p<0.05$, Wilcoxon rank sum test. (**E**) Expression levels of RNLTR12-int and myelination hubs were represented as heat map and a dendrogram derived after two-way hierarchical clustering.

Fig S2

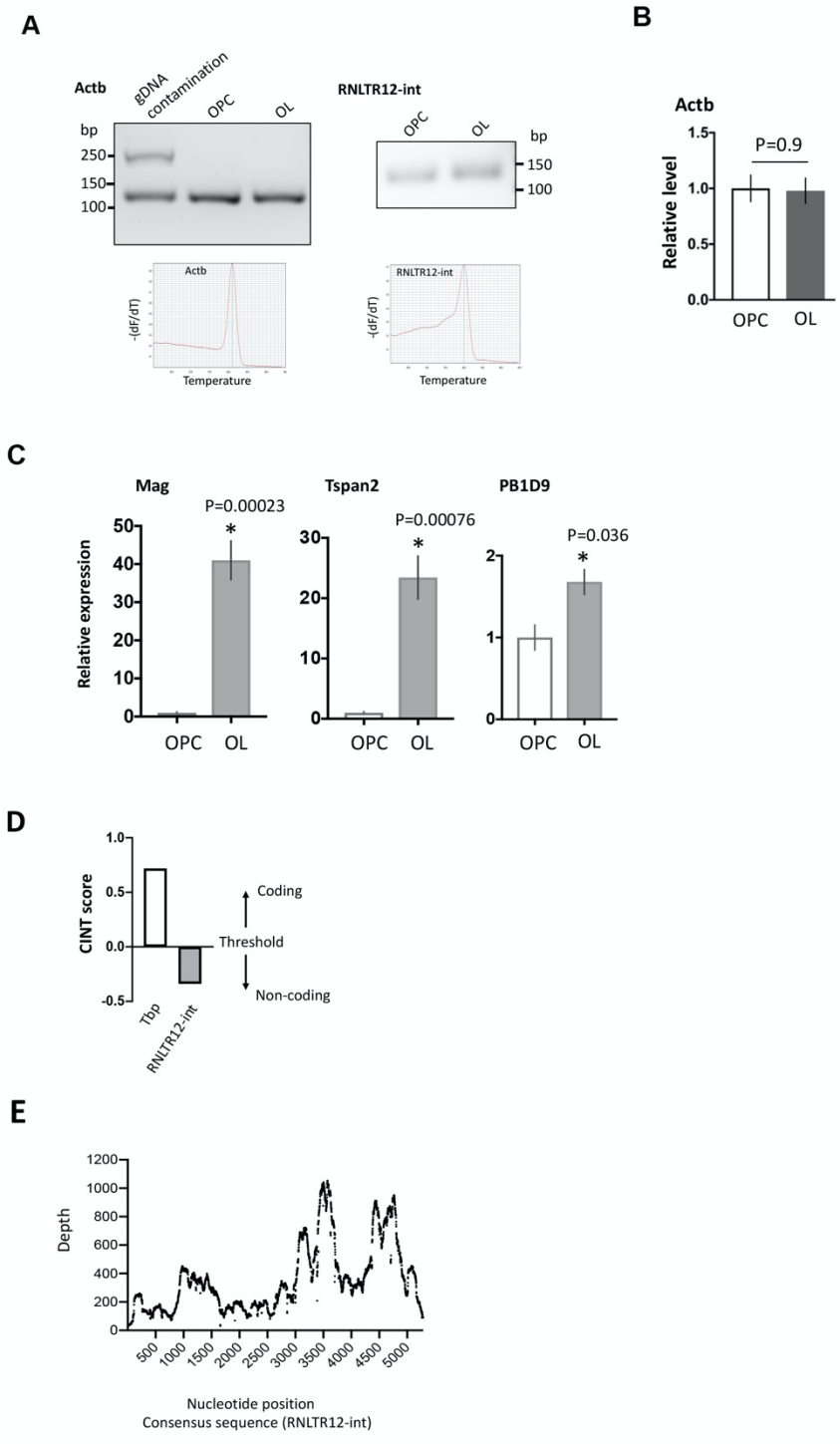
A
Actb
gDNA contamination
OPC
OL
bp
250
150
100
RNLTR12-int
OPC
OL
bp
150
100
Actb
-(dF/dT)
Temperature
RNLTR12-int
-(dF/dT)
Temperature
B
Actb
P=0.9
Relative level
1.5
1.0
0.5
0.0
OPC
OL
C
Mag
P=0.00023
Tspan2
P=0.00076
PB1D9
P=0.036
Relative expression
50
40
30
20
10
0
OPC
OL
30
20
10
0
OPC
OL
2
1
0
OPC
OL
D
CINT score
1.0
0.5
0.0
-0.5
Coding
Threshold
Non-coding
Tbp
RNLTR12-int
E
Depth
1200
1000
800
600
400
200
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
Nucleotide position
Consensus sequence (RNLTR12-int)

Fig. S2. Validation of RNLTR12-int expression in MACS isolated OPC (A2B5+) and OL (Mog+) (**A**) Top: (left) RT-PCR analysis of cytoplasmic beta-actin (Actb) using intron flanking primers. gDNA contamination in RNA resulted an upper band. Only single lower band obtained in OPC and OL samples, confirms no gDNA contamination. (right) RT-PCR analysis of RNLTR12-int expression in OPC and OL. PCR product was run on a 2% agarose gel. Bottom: qPCR melting curve analysis resulted only single peak for Actb (left) and RNLTR12-int (right). (**B**) Actb expressions remained unchanged in OPC and OL as determined by Affymetrix GeneChip. Normalized intensity of all Actb specific probes (17 probes) were used and plotted relative to OPC. N=9 (OPC), 8 (OL), $p>0.05$, Student's t-test (unpaired, two-tailed). mean$\pm$SEM. (**C**) Expression levels of Mag, Tspan2 and PB1D9 is elevated in OL as compared to OPC determined by RT-qPCR. N=3-4, *$p<0.05$, Student's t-test (unpaired, two-tailed). mean$\pm$SEM (**D**) RNLTR12-int encoded transcript is likely to be a non-coding RNA as predicted by CNIT algorithm (36). Score of a known protein coding gene, TATA-box binding protein (Tbp, Transcript ID: ENSRNOT00000002038.4) is shown. (**E**) RNA sequencing reads were aligned to RNLTR12-int consensus sequence and the depth was plotted.

Fig S3

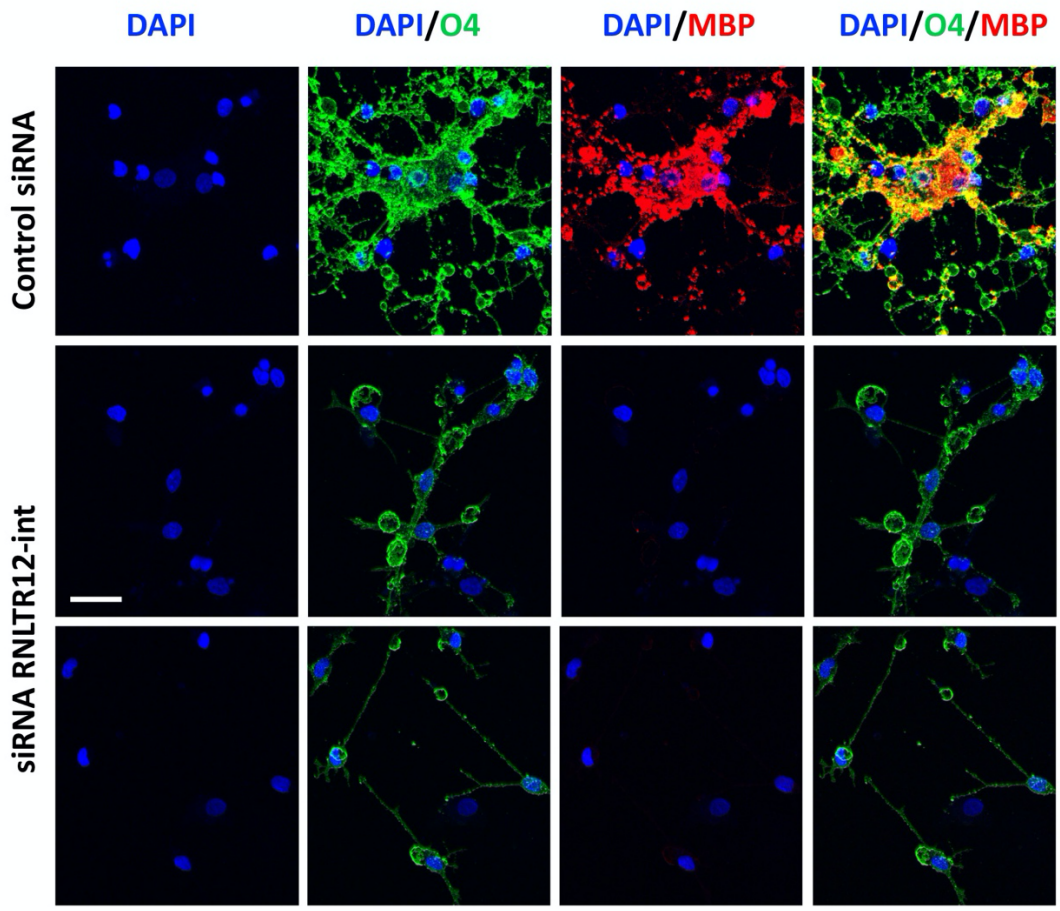

Fig S3. Absence of complex oligodendrocyte morphology due to inhibition of RNLTR12-int. Immunofluorescence analysis O4 immunostaining 5 days after transfection of siRNA. Scale bar: 26 µm. siRNLTR12-int: siRNA against RNLTR12-int, control siRNA: siGENOME non-targeting siRNA pool (Dharmacon). Representative image of 3 independent experiments.

Fig S4

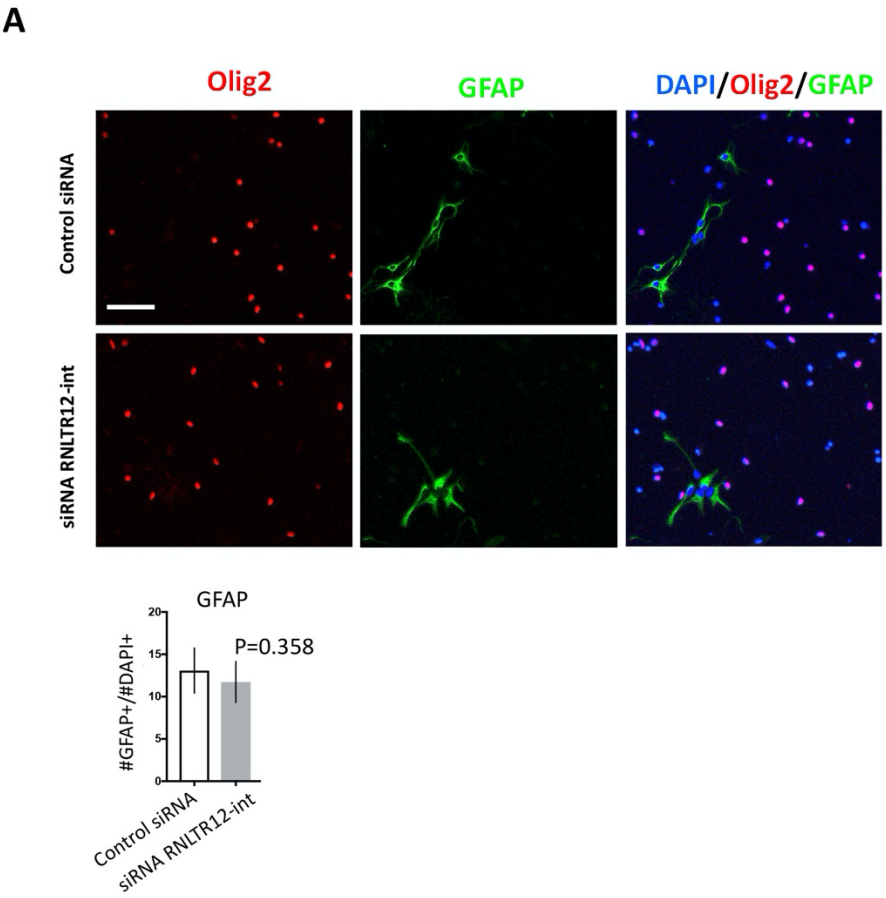
A
Olig2
GFAP
DAPI/Olig2/GFAP
Control siRNA
siRNA RNLTR12-int
GFAP
#GFAP+/#DAPI+
P=0.358
Control siRNA
siRNA RNLTR12-int
B
Pdgfra
Relative expression
p=0.135
Control siRNA
siRNA RNLTR12-int
Cnp
p=0.032
*
Control siRNA
siRNA RNLTR12-int

Fig. S4. Inhibition of RNLTR12-int did not initiate astrocytic fate. (**A**) Top: Immunofluorescence analysis of GFAP immunostaining 5 days after transfection of siRNA. Scale bar: 60μm. Bottom: Percentage of GFAP+ cell in total DAPI+ cells were plotted. N=3 independent experiments (each time 3 replicates), mean±SEM, $p>0.05$, Two-way ANOVA. siRNLTR12-int: siRNA against RNLTR12-int, control siRNA: siGENOME non-targeting siRNA pool (Dharmacon). (**B**) Effect of the inhibition of RNLTR12- int on RNA level expression of Pdgfa and Cnp were determined by RT-qPCR. Data were normalised to Actb. N=3 independent experiments, mean±SEM, * $p<0.05$, Student's t test (unpaired, two tailed).

Fig S5

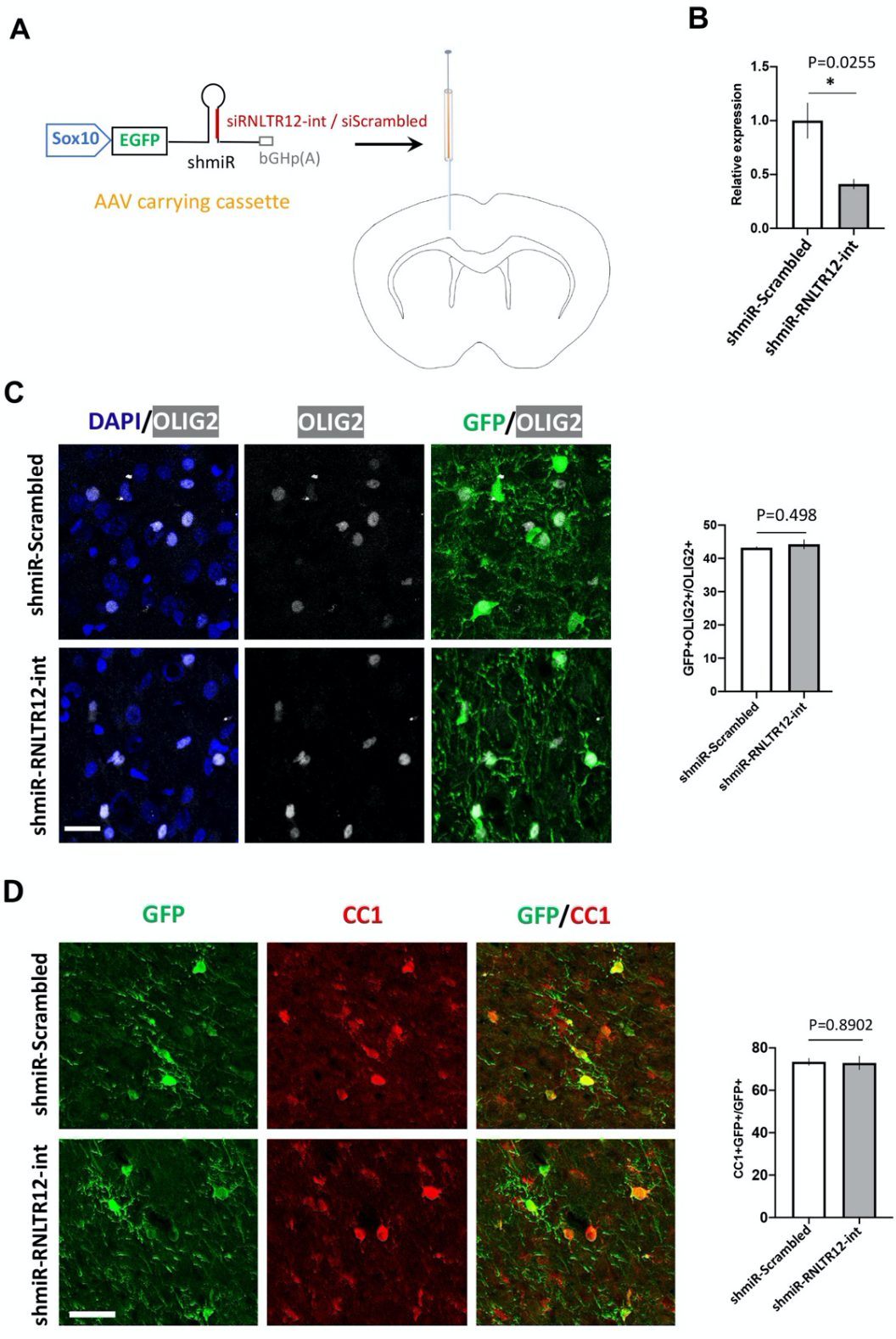
A
Sox10
EGFP
shmiR
siRNLTR12-int / siScrambled
bGHp(A)
AAV carrying cassette
B
P=0.0255
*
Relative expression
1.5
1.0
0.5
0.0
shmiR-Scrambled
shmiR-RNLTR12-int
C
DAPI/OLIG2
OLIG2
GFP/OLIG2
shmiR-Scrambled
shmiR-RNLTR12-int
P=0.498
GFP+OLIG2+/OLIG2+
50
40
30
20
10
0
shmiR-Scrambled
shmiR-RNLTR12-int
D
GFP
CC1
GFP/CC1
shmiR-Scrambled
shmiR-RNLTR12-int
P=0.8902
CC1+GFP+/GFP+
80
60
40
20
0
shmiR-Scrambled
shmiR-RNLTR12-int

Fig. S5. (**A**) SOX10 driven EGFP (Emerald GFP) construct that carry shmiR where shRNA is embedded into a microRNA (miR-30) cassette. bGHp(A): Bovine growth hormone poly A signal. This construct is SOX10 driven and expresses EGFP. This also produces shRNA that function through RNAi pathway. Shown is the rat brain area where AAV (carrying SOX10-EGFP-shmiR-bGHp(A) were injected. (**B**) AAV carrying the above construct were infected into the cultured OPC and allowed to differentiate for 4 days, then RNA was isolated. RT-qPCR analysis of RNLTR12-int revealed its reduced expression in shmiR-RNLTR12-int infected cells. N=3 independent experiments, mean$\pm$SEM, * $p<0.05$, Student's t test (unpaired, two tailed). (**C-D**) AAV carrying the above construct were injected into newborn rat brain (at P1). Brains were harvested at P14 for immunofluorescence. (**C**) Left: immunofluorescence analysis of OLIG2 and GFP immunostaining. Right: quantification of OLIG2+GFP+ cells and plotted as a percentage to OLIG2+ cells. N=3 rats, mean$\pm$SEM, $p>0.05$, Student's t test (unpaired, two tailed). (**D**) Left: immunofluorescence analysis of CC1 and GFP immunostaining. Right: quantification of CC1+GFP+ cells among GFP+ cells and represented as a percentage. N=3 rats, mean$\pm$SEM, $p>0.05$, Student's t test (unpaired, two tailed).

Fig S6

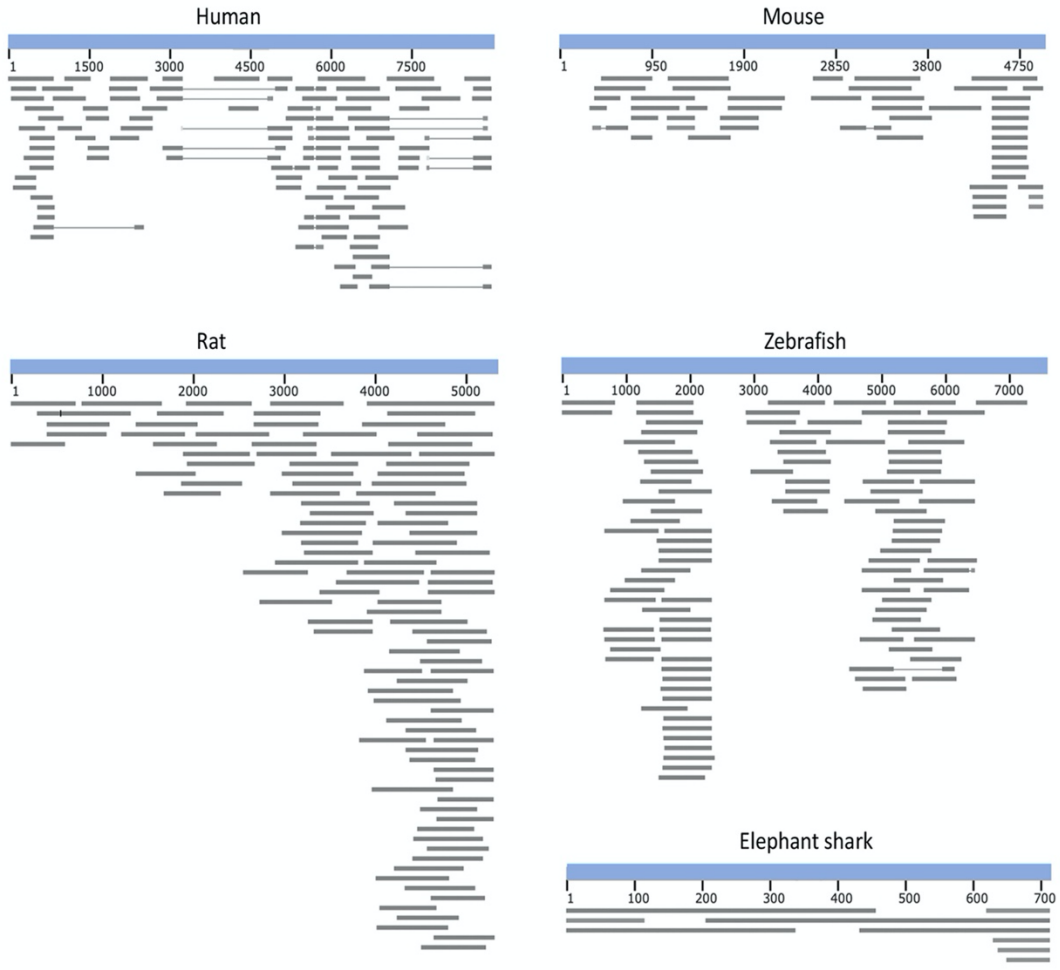

Fig. S6. Expressed sequence tags (ESTs) were aligned to the RetroMyelin sequence of human, mouse, rats, zebrafish and elephant sharks. BLASTN was used to search ESTs. For the displayed ESTs: 0<E-value<6e-26 and 119<alignment score<2298. For human, mouse, rats and zebrafish: consensus sequence was used. For elephant shark: KI636671.1: 5058-5765 (Callorhinchus_milii-6.1.3) (this was the best hit obtained after nhmmer to RNLTR12-int; E-value=1.7e-14).

Table S1: Identification of RNLTR12-int like sequences (*Retromyelin*).

Vertebrates	Genome assembly	Top hit	E-value	Matching repeat	Family
Human	hg38	chr16:35270573-35269845	4.6E-36	ERV1-int (HERVS71-int)	LTR/ERV1
Chimpanzee	panTro4	chr13:18277202-18274245	7.4E-101	ERV1-int (PTERV1c-int)	LTR/ERV1
Gorilla	gorGor3	CABD02423608:217-1465	3.6E-87	ERV1-int (HERVS71-int)	LTR/ERV1
Mouse	mm10	chr7:30112940-30114182	1.5e-142	ERV1-int (MERV1_I-int)	LTR/ERV1
Rat	rn5	chr15:24404804-24399534	0	ERV1-int (RNLTR12-int)	LTR/ERV1
Cow	bosTau7	chr29:42154010-42153024	1.8e-29	ERV1-int (BtERVF2_I-int)	LTR/ERV1
Horse	equCab2	chr1:27399148-27399594	2.9e-18	ERV1-int (ERV1-3-EC_I-int)	LTR/ERV1
Zebrafinch	taeGut1	chr27:4207784-4209022	1.1e-19	ERV1-int (TguERV2_I-int)	LTR/ERV1

Chicken	galgal4	chrW_JH375236_random:9151-8296	4.1E-14	ERV1-int (GGLTR7-int)	LTR/ERV1
Turtle	chrPic1	AHGY01435164:4348-5025	1.3e-15	ERV1-int (ERV1-1B_CPB-I)	LTR/ERV1
Xenopus	xenTro7	KB021654:67698951-67699361	4.4E-08	ERV1-int (ERV1-5-I_XT)	LTR/ERV1
Coelacanth	latCha1	JH128184:158800-157936	2.5e-17	ERV1-int	LTR/ERV1
Stickleback	gasAcu1	chrUn:56084321-56084993	1.7E-07	ERV1-int (Gypsy-30_GA-I)	LTR/ERV1
Takifugu	fr3	HE592075:741-39	4E-14	ERV1-int (FERV-R2_I-int)	LTR/ERV1
Seabass	seabass_V1.0	HG916851.1:75130269-75131193	4.6E-12	ERV1-int	LTR/ERV1
Spiny chromis	ASM210954v1	MVNR01000444.1:145194-144494	1.60E-08	ERV1-int	LTR/ERV1
Cichlid	NeoBri1.0	JH422275.1:16925254-16925968	2.20E-10	ERV1-int	LTR/ERV1
Medaka	Om_v0.7.RACA	NVQA01006789.1:101671-102370	5.70E-12	ERV1-int	LTR/ERV1

Turbot	ASM318616v1	chr4:26831079-26830372	4.50E-10	ERV1-int	LTR/ERV1
Atlantic cod	gadMor3.0	chr4:29532684-29533397	2.1e-12	ERV1-int	LTR/ERV1
Zebrafish	danRer7	chr15:43196012-43195219	1.9e-14	ERV1-int (ERV1-3-I_DR)	LTR/ERV1
Whale shark	GCF_001642345.1_ASM164234v2	NW_018046349.1:22177-21460	2.8e-19	ERV1-int	LTR/ERV1
Elephant shark	Callorhinchus_milii-6.1.3	KI636671.1:5058-5765	1.7E-14	ERV1-int	LTR/ERV1
Lamprey	Pmarinus_7.0/petMar2	Not found	NA	NA	NA
Lancelet	braflo2	Bf_V2_65:4822224-4821722	3.6E-07	(TC)n	Simple_repeat
Sea Urchin	strPur2	Scaffold69942:23181-22682	1.70E-09	(TC)n	Simple_repeat
Sea anemone	nemVec1	Not found	NA	NA	NA
C. elegans	ce10	Not found	NA	NA	NA
Drosophila	dm6	Not found	NA	NA	NA

This table list the top hit of the identified repeat type, after searching remote homology using nhmmer (40). Specific repeat annotation, wherever available, is written under the parathesis in column 5. NA: Not applicable. Not found: wherever no hit is being obtained above the inclusion threshold (E-value<0.01).