

LUND UNIVERSITY

Transposable Elements in the Healthy and Diseased Human Brain

Garza, Raquel

2024

Document Version:
Publisher's PDF, also known as Version of record

Link to publication

Citation for published version (APA):
Garza, R. (2024). *Transposable Elements in the Healthy and Diseased Human Brain*. [Doctoral Thesis (compilation), Department of Experimental Medical Science]. Lund University, Faculty of Medicine.

Total number of authors:
1

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Transposable Elements in the Healthy and Diseased Human Brain

Raquel Garza



LUND UNIVERSITY

DOCTORAL DISSERTATION

Doctoral dissertation for the degree of Doctor of Philosophy (PhD)
at the Faculty of Medicine at Lund University
to be publicly defended on 19th of January 2024
at 09.00 in Segerfalksalen, Department of Experimental Medical Science,
Sölvegatan 17, 22362, Lund Sweden.

Faculty opponent

Dr Miguel R. Branco

Blizard Institute, Queen Mary University of London,
London, UK.

Organization
LUND UNIVERSITY

Molecular Neurogenetics Laboratory
Department of Experimental Medical Science,
Faculty of Medicine, Sweden

Document name
DOCTORAL DISSERTATION

Date of issue
2024-01-19

Sponsoring organization

Author(s)
Raquel Garza

Title and subtitle
Elements in the Healthy and Diseased Human Brain

Abstract Transposable elements (TEs) are mobile genetic sequences that comprise around 50% of our genomic DNA. Their mutagenic potential and gene regulatory effect have shaped the evolution of transcriptional networks involved in development, pluripotency, and inflammation. More so, TEs are a rich source of genetic variation, which makes them an intriguing research avenue to investigate human-specific traits, including their impact on human brain evolution and their relevance in disease. However, studies focused on TEs face technological challenges due to their repetitive nature, which require special bioinformatic considerations. The work presented in this thesis focuses on the bioinformatic methods for a TE-centric analysis of next- and third-generation sequencing technologies. Using a multi-omics approach, we demonstrate that TEs introduce a layer of transcriptome complexity to the human brain. We found that the regulation of TE transcription during brain development is essential for the establishment of long-term transcriptional repression carried to adulthood (Paper I and IV). More so, our results show that the epigenetic regulation of TE transcription is dynamically regulated throughout life (Paper II), upon the beginning of neuroinflammation (Paper III), and in a disease-driving polymorphic TE insertion (Paper IV). Overall, our findings highlight the importance of TEs as regulatory agents and their dynamic activity during development, adult life, and disease in the human brain.

Key words: Transposable elements, Neuroscience, Neuroinflammation, Evolution, Epigenetic regulation, Bioinformatics

Classification system and/or index termes (if any):

Supplementary bibliographical information:

Language
English

ISSN and key title:
1652-8220

ISBN
978-91-8021-506-0

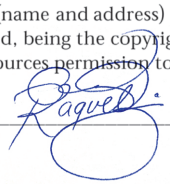
Recipient's notes

Number of pages
179

Price

Security classification

Distribution by (name and address)
I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature  Date 2023-12-04

Transposable Elements in the Healthy and Diseased Human Brain

Raquel Garza

2024

Molecular Neurogenetics Laboratory
Department of Experimental Medical Science,
Faculty of Medicine, Sweden



LUND
UNIVERSITY

Coverphoto by Raquel Garza.

In a landscape format, the watercolour painting shows the first few hundred base pairs of the consensus sequence of some of the human-specific TEs the studies in this thesis focused on – L1HS, SVA F, and LTR5Hs. Digitally overlayed are co-workers', friends' and family's drawings of the cells used for this investigation.

Copyright Raquel Garza Gómez and the respective publishers

Paper 1 © The EMBO Journal
Paper 2 © Science Advances
Paper 3 © Cell Reports
Paper 4 © by the Authors (Manuscript unpublished)

Faculty of Medicine
Department of Experimental Medical Sciences

ISSN 1652-8220
ISBN 978-91-8021-506-0
Lund University, Faculty of Medicine, Doctoral Dissertation Series 2024:13
Printed by Exakta printing AB, Malmö, Sweden
Lund 2024

To women in STEM

Table of contents

Abstract 9
English Lay Summary 11
Resumen popular en español 13
Svensk Populärvetenskaplig sammanfattning 15
Original papers included in the thesis 17
Original papers not included in the thesis 18
Research review articles not included in the thesis 19
Abbreviations 21
Introduction 23
- Preface 23
- Friends or foes 24
- Human TEs 25
  - **LINE-1** 25
  - **SVA** 26
  - **HERV** 27
  - **Bioinformatic considerations of TEs** 28
  - **Technology status and limitations** 28
  - **Mappability of TEs** 29
  - **TE analyses from single cell RNA sequencing** 30
- Epigenetic regulation of TEs 31
  - **Transcriptional silencing of TEs via H3K9me3 deposition** 31
  - **Transcriptional silencing of TEs via DNA methylation** 32
  - **Other silencing or inactivating mechanisms** 33
- Potential consequences of TE transcription 34
  - **Polymorphic TE insertions as a trigger for disease development** 34
  - **TEs as immunogenic sequences** 35
  - **TEs in relation to a neuroinflammatory state** 36

Aims of the thesis 37
Materials and Methods 39
- Sequencing strategies 39
  - **RNA-based** 39
  - **DNA-based** 40
- Bioinformatic analyses 40
  - **Short-read bulk RNAseq** 40

Short-read single nuclei RNAseq 44
Long read bulk RNAseq 45
CUT&RUN 45
Long read DNA sequencing 45
Other TE-related analyses 47
Computer systems and workflow management 49
Summary of results 51
Accurate TE subfamily quantification from single-cell RNA sequencing technologies using pseudo-bulk (papers I, II, III) 51
TRIM28-mediated silencing of TEs is required for the establishment of a stable silencing during early development (papers I, IV) 54
TEs contribute to the human transcriptome complexity (papers II, IV) 56
Aberrant expression of ERVs can lead to an immune response (papers I, III) 59
Conclusions and future perspectives 65
References 67
Acknowledgements 75
Appendix 79
Paper I 79
Paper II 99
Paper III 121
Paper IV 151

Abstract

Transposable elements (TEs) are mobile genetic sequences that comprise around 50% of our genomic DNA. Their mutagenic potential and gene regulatory effect have shaped the evolution of transcriptional networks involved in development, pluripotency, and inflammation. More so, TEs are a rich source of genetic variation, which makes them an intriguing research avenue to investigate human-specific traits, including their impact on human brain evolution and their relevance in disease. However, studies focused on TEs face technological challenges due to their repetitive nature, which require special bioinformatic considerations. The work presented in this thesis focuses on the bioinformatic methods for a TE-centric analysis of next- and third-generation sequencing technologies. Using a multi-omics approach, we demonstrate that TEs introduce a layer of transcriptome complexity to the human brain. We found that the regulation of TE transcription during brain development is essential for the establishment of long-term transcriptional repression carried to adulthood (Paper I and IV). More so, our results show that the epigenetic regulation of TE transcription is dynamically regulated throughout life (Paper II), upon the beginning of neuroinflammation (Paper III), and in a disease-driving polymorphic TE insertion (Paper IV). Overall, our findings highlight the importance of TEs as regulatory agents and their dynamic activity during development, adult life, and disease in the human brain.

English Lay Summary

Thanks to technological advances in sequencing technologies during the past couple of decades, we have increasing knowledge about our genome's content. We now know that about half of it is made of "transposable elements" (TEs). TEs are fragments of our genome that can make a copy of themselves and insert it at a different place. Through millions of years of evolution, this has resulted in the accumulation of millions of TEs within our genome – a process that is still ongoing and may cause variation between individuals.

Computational analyses over our genome can be challenging due to TEs' repetitive nature and their varying content among individuals in the population. Thus, this thesis focused on the development of computational workflows to study TEs using a variety of current sequencing technologies. I summarise some of the results below.

Not only can TEs create new copies of themselves, but they can also affect the functioning of our genes. For example, a new TE copy can end up in the middle of a gene, changing its original sequence to one that will not function in a normal way. Depending on this gene's function and the TE placement in it, the change (or mutation) can be dangerous to the person carrying it. Therefore, our genome has evolved mechanisms to prevent TEs from creating new insertions. Nonetheless, these mechanisms sometimes fail, and TE insertions may lead to disease, an example of which is XDP – a neurodegenerative disorder we studied in this thesis. XDP patients carry a TE insertion at a crucial gene called TAF1. TAF1 malfunction leads to a chain of events that produces motor and neurological symptoms in the patients.

Not everything about TEs has been disadvantageous or dangerous – otherwise evolution would have already weeded them out. In fact, previous studies have shown that TEs have helped the evolution of our genome by forcingly introducing changes. Most TEs in our genome are motionless with seemingly no effect in the genome's functioning. However, some TEs regulate genes, a small minority still create new insertions, and some have even been "recycled" to gain new, beneficial roles.

Given their prevalence and the changes they have introduced throughout evolution, TEs are perfect research candidates to understand, for example, how the human brain evolved. Part of this thesis investigates the role of TEs in the evolution of the human brain, and if they have a current role in the regulation of genes in the brain.

Previous research has shown that TE activity correlates with an inflammatory response in many human diseases. We know from studies in other tissues that, given certain conditions or disease contexts, the controlling mechanisms of TEs can fail and our immune system react as if in the presence of a virus. However, given that TEs have been part of our genome throughout evolution, the immune system has also been shaped around TEs. Thus, it may also be that an immune reaction (regardless of the trigger), can result into increased TE activity.

Not much is known about this process in the brain, however. Specifically – can it happen? Under which contexts? What comes first, the TE activity or the immune reaction? Can they boost each other? In this thesis, we found that mice which undergo brain development with an induced increase of TE activity develop an inflammatory response that persists until adulthood. Thus, we know that brain development is a crucial timepoint for the control of TE activity.

To understand if TEs play a role in an inflammatory response, we also investigated if the level of TE activity changes at the beginning of an inflammatory response caused by traumatic brain injury (TBI) in humans. We collected tissue from TBI patients and, with our latest sequencing technologies, studied the inflammatory response and how it differed between the different cells. We found crucial cell types that play a role during inflammation, and we observed that the response caused by the injury is sufficient to trigger an increase in TE activity in certain brain cells.

In summary, this thesis contains studies that aim to understand the role of TEs in the evolution of the human brain and how they can cause disease.

Resumen popular en español

Gracias a avances en las últimas décadas en el tema de tecnologías de secuenciación, conocemos cada vez más el contenido de nuestro genoma. Hemos descubierto que aproximadamente la mitad de nuestro genoma está compuesto de "elementos transponibles" (ET). Los ET son fragmentos de nuestro genoma que pueden crear copias de sí mismos e insertarlas en diferentes lugares. Dado el paso de millones de años de evolución, esto ha resultado en la acumulación de millones de ET dentro de nuestro genoma. Esta expansión sigue en proceso, y sabemos que puede causar variación genética entre individuos.

Los análisis computacionales de nuestro genoma pueden complicarse debido a la naturaleza repetitiva de los ET y su contenido variable entre individuos. Por esta razón, esta tesis se centra en el desarrollo de flujos computacionales para estudiar ET utilizando una variedad de tecnologías de secuenciación. A continuación, resumo algunos de nuestros resultados.

Los ET no solo pueden crear nuevas copias de sí mismos, también pueden afectar el funcionamiento de nuestros genes. Por ejemplo, una nueva copia de un ET puede terminar en medio de un gen – lo cual puede cambiar su secuencia a una que no funcionará de manera normal. Dependiendo de la función de este gen y de la ubicación del ET dentro de él, este cambio (o mutación) puede ser peligroso para la persona que lo porta. Es por esto que nuestro genoma ha desarrollado mecanismos para evitar que los ET creen nuevas inserciones. Sin embargo, estos mecanismos a veces fallan y las inserciones de ET pueden desencadenar el desarrollo de enfermedades. Un ejemplo de ello es XDP (por sus siglas en inglés), una enfermedad neurodegenerativa que estudiamos en esta tesis. Los pacientes con XDP portan la inserción de un ET en un gen llamado TAF1. El malfuncionamiento de TAF1 inicia una reacción en cadena que produce síntomas motores y neurológicos en los pacientes.

No todos los cambios que los ET han introducido a nuestro genoma han sido perjudiciales o peligrosos. De haber sido así, la evolución ya los habría eliminado. Al contrario, estudios han demostrado que los ET han ayudado a la evolución de nuestro genoma. La mayoría de los ET en nuestro genoma son inmóviles y no parecen tener ningún efecto en el funcionamiento del genoma. Sin embargo, algunos siguen creando copias, otros pueden regular genes, y otros han sido "reciclados" y ahora tienen funciones nuevas.

Debido a su prevalencia e influencia a lo largo de la evolución, los ET son buenos sujetos de estudio para investigar, por ejemplo, cómo evolucionó el cerebro humano. Parte de esta tesis se centra en el rol de los ET durante la evolución del cerebro humano, así como su rol actual en la regulación de genes en el cerebro.

Estudios anteriores han demostrado que la actividad de algunos ET se correlaciona con una respuesta inflamatoria en muchas enfermedades humanas. Sabemos por estudios en otros tejidos que, dadas ciertas condiciones o enfermedades, los mecanismos para el control de los ET pueden

fallar y resultar en una respuesta inmunológica como si se estuviera en presencia de un virus. Sin embargo, dado que los ET han sido parte de nuestro genoma a lo largo de la evolución, el sistema inmunológico también ha evolucionado alrededor de ET. Por lo tanto, puede ser que una reacción inmune (independientemente de la razón por la cual inició) pueda resultar en el aumento de actividad de ET.

No se sabe mucho sobre estos procesos en el cerebro. Específicamente, ¿en qué medida puede suceder? ¿Bajo qué contextos? ¿Qué viene primero, la actividad de ET o la reacción inmune? ¿Será que estos efectos se propulsan mutuamente? En uno de los estudios incluidos en esta tesis, encontramos que cuando se induce un aumento en la actividad de ET durante el desarrollo cerebral en ratones, los animales sufren una respuesta inflamatoria que persiste hasta la edad adulta. Por lo tanto, sabemos que el desarrollo cerebral es un momento crucial y delicado para el control de la actividad de ET.

Para estudiar el rol de los ET en el proceso inflamatorio, investigamos si los niveles de actividad de ET cambian al comienzo de una respuesta inflamatoria causada por lesiones cerebrales traumáticas (TBI por sus siglas en inglés) en humanos. Recolectamos tejido de pacientes con TBI y, con nuestras últimas tecnologías de secuenciación, estudiamos la respuesta inflamatoria y cómo difiere entre diferentes células. Identificamos cambios cruciales en diferentes tipos de células, y observamos que la respuesta inmunológica provocada por la lesión desencadena un aumento en la actividad de ET en cierto tipo de células en el cerebro.

En resumen, esta tesis contiene estudios que tienen como objetivo comprender el rol de los ET durante la evolución del cerebro humano y cómo pueden causar enfermedades.

Svensk Populärvetenskaplig sammanfattning

Tack vare tekniska framsteg inom sekvensering under de senaste decennierna har vi fått ökad kunskap om vår arvsmassas innehåll. Nu vet vi att ungefär hälften av den består av ”transposabla element” (TE). TE är fragment av vår arvsmassa som kan göra kopior av sig själva och infoga dessa på en ny plats. Genom miljontals år av evolution har detta resulterat i ackumulering av miljoner TE i vår arvsmassa – en process som fortfarande pågår och kan ge upphov till genetiska variationer mellan individer.

Beräkningsanalyser av vår arvsmassa kan vara utmanande att genomföra på grund av TE:s repetitiva natur och deras varierande innehåll mellan individer i befolkningen. Därför fokuserar denna avhandling på utvecklingen av beräkningsanalyser för att studera TE med hjälp av olika aktuella sekvenseringsteknologier. Nedan följer en sammanfattning av resultaten.

TE kan inte bara skapa kopior av sig själva utan de kan också påverka våra geners funktion. Till exempel, om en ny TE-kopia hamnar mitt i en gen kan dess ursprungliga sekvens ändras till en som inte fungerar på ett normalt sätt. Beroende på genens funktion och hur det TE:et placeras kan förändringen (eller mutationen) vara farlig för personen som bär den. Därför har vår arvsmassa utvecklat mekanismer för att förhindra TE från att skapa denna typ av insättningar. Ibland misslyckas dessa mekanismer och insättningen av TE leder till sjukdom – exempelvis XDP – en neurodegenerativ sjukdom vi studerade i denna avhandling. XDP patienter bär på en TE insättning i en viktig gen som heter TAF1. Bristande funktionsförmåga av TAF1 kan ge upphov till ett händelseförlopp som resulterar i motoriska och neurologiska symptom hos patienterna.

TE är inte bara ofördelaktiga och farliga, då hade de redan blivit utsållade under evolutionen. Faktum är att tidigare studier visat att TE har hjälpt till med utvecklingen av vår arvsmassa under evolutionen genom att tvinga fram förändringar. De flesta TE i vår arvsmassa är orörliga och saknar till synes effekt på våra geners funktion. Vissa TE reglerar gener, en liten minoritet skapar fortfarande nya insättningar och vissa har till och med ”återvunnits” för att få nya fördelaktiga roller. Med tanke på deras omfattning och de förändringar de har introducerat under evolutionen är TE perfekta kandidater för att förstå hur den mänskliga hjärnan har utvecklats. En del av denna avhandling undersöker därför den roll TE har spelat under utvecklingen av den mänskliga hjärnan och om de spelar en aktuell roll i hur gener regleras i hjärnan.

Tidigare studier har visat ett samband mellan aktivitet hos TE och inflammatoriskt svar vid olika mänskliga sjukdomar. Från studier i andra vävnader vet vi, beroende på tillstånd och sjukdomssammanhang, att de TE-kontrollerande mekanismerna kan fallera och göra så att vårt immunförsvar regerar som vid en virusinfektion. Med tanke på att TE har varit en del av vår arvsmassa under evolutionen bör immunförsvaret också formats kring TE. Därför är det mycket möjligt att ett inflammatoriskt svar (oavsett hur det triggas) också kan leda till ökad TE-aktivitet. Inte mycket är känt kring denna process i hjärnan. Framför allt – kan det hända? Under vilka förutsättningar? Vad kommer

först, TE-aktivitet eller det inflammatoriska svaret? Kan de förhöja varandra? I denna avhandling fann vi att möss som har en inducerad förhöjning av TE-aktivitet under hjärnans utveckling kan utveckla ett inflammatoriskt svar som kvarstår i vuxen ålder. Således vet vi att hjärnas utveckling är en känslig tidpunkt för kontroll av TE-aktivitet.

För att förstå om TE spelar en roll vid ett inflammatoriskt svar undersökte vi också om nivån av TE-aktivitet förändras i början av ett inflammatoriskt svar orsakat av traumatisk hjärnskada (THS) hos människor. Vi samlade in hjärnvävnad från patienter med THS och studerade med hjälp av vår senaste sekvenseringsteknologi det inflammatoriska svaret och hur det skiljer sig åt mellan olika celler. Vi fann celltyper som spelar en viktig roll under inflammationen och såg att responsen orsakad av skadan är tillräcklig för att utlösa en ökad TE-aktivitet i vissa typer av hjärnceller.

Sammanfattningsvis innehåller denna avhandling studier som syftar till att förstå hur TE spelar en roll under evolutionen av den mänskliga hjärnan samt hur de kan orsaka sjukdom.

Original papers included in the thesis

Paper I

Jönsson, M., **Garza, R.,** Sharma, Y., Petri, R., Södersten, E., Johansson, J., Johansson, P., Atacho, D., Pircs, K., Madsen, S., Yudovich, D., Ramakrishnan, R., Holmberg, J., Larsson, J., Jern, P., Jakobsson, J. (2021).

Activation of endogenous retroviruses during brain development causes an inflammatory response.

The EMBO Journal, 40(9), e106423.

Paper II

Garza, R., Atacho, D., Adami, A., Gerdes, P., Vinod, M., Hsieh, P., Karlsson, O., Horvath, V., Johansson, P., Pandiloski, N., Matas, J., Quaegebeur, A., Kouli, A., Sharma, Y., Jönsson, M., Monni, E., Englund, E., Eichler, E., Hammell, M., Barker, R., Kokaia, Z., Douse, C., Jakobsson, J. (2023).

L1 retrotransposons drive human neuronal transcriptome complexity and functional diversification.

Science Advances, 9(44), eadh9543.

Paper III

Garza, R., Sharma, Y., Atacho, D., Thiruvalluvan, A., Hamdeh, S., Jönsson, M., Horvath, V., Adami, A., Ingelsson, M., Jern, P., Hammell, M., Englund, E., Kirkeby, A., Jakobsson, J., Marklund, N. (2023).

Single-cell transcriptomics of resected human traumatic brain injury tissues reveals acute activation of endogenous retroviruses in oligodendroglia.

Cell Reports, 113395.

Paper IV

Horvath, V., **Garza, R.,** Jönsson, M., Johansson, P., Adami, A., Karlsson, O., Castilla Vallmanya, L., Christoforidou, G., Gerdes, P., Pandiloski, N., Douse, C., Jakobsson, J. (2023).

Mini-heterochromatin domains constrain the cis-regulatory impact of SVA transposons in human brain development and disease.

BioRxiv.

Original papers not included in the thesis

Johansson, P., Brattås, P., Douse, C., Hsieh, P., Adami, A., Pontis, J., Grassi, D., **Garza, R.,** Sozzi, E., Cataldo, R., Jönsson, M., Atacho, D., Pircs, K., Eren, F., Sharma, Y., Johansson, J., Fiorenzano, A., Parmar, M., Fex, M., Trono, D., Eichler, E., Jakobsson, J. (2022).

A cis-acting structural variation at the ZNF558 locus controls a gene regulatory network in human brain development.

Cell Stem Cell, 29(1), 52–69.e8.

Pircs, K., Drouin-Ouellet, J., Horváth, V., Gil, J., Rezeli, M., **Garza, R.,** Grassi, D., Sharma, Y., St-Amour, I., Harris, K., Jönsson, M., Johansson, P., Vuono, R., Fazal, S., Stoker, T., Hersbach, B., Sharma, K., Lagerwall, J., Lagerström, S., Storm, P., Hébert, S., Marko-Varga, G., Parmar, M., Barker, R., Jakobsson, J. (2022).

Distinct subcellular autophagy impairments in induced neurons from patients with Huntington's disease.

Brain: a journal of neurology, 145(9), 3035–3057

Gustavsson, E., Sethi, S., Gao, Y., Brenton, J., García-Ruiz, S., Zhang, D., **Garza, R.,** Reynolds, R., Evans, J., Chen, Z., Grant-Peters, M., Macpherson, H., Montgomery, K., Dore, R., Wernick, A., Arber, C., Wray, S., Gandhi, S., Esselborn, J., Blauwendraat, C., Douse, C., Adami, A., Atacho, D., Kouli, A., Quaegebeur, A., Barker, R., Englund, E., Platt, F., Jakobsson, J., Wood, N., Houlden, H., Saini, H., Bento, C., Hardy, J., Ryten, M. (2023).

The annotation and function of the Parkinson's and Gaucher disease-linked gene GBA1 has been concealed by its protein-coding pseudogene GBAP1.

BioRxiv.

Pandiloski, N., Horvath, V., Karlsson, O., Christoforidou, G., Dorazehi, F., Koutounidou, S., Matas, J., Gerdes, P., **Garza, R.,** Jönsson, M., Adami, A., Atacho, D., Johansson, J., Englund, E., Kokaia, Z., Jakobsson, J., Douse, C. (2023).

DNA methylation governs the sensitivity of repeats to restriction by the HUSH-MORC2 corepressor.

BioRxiv.

Research review articles not included in the thesis

Jönsson, M., **Garza, R.,** Johansson, P., Jakobsson, J. (2020).

Transposable elements: a common feature of neurodevelopmental and neurodegenerative disorders.

Trends in genetics: TIG, 36(8), 610–623.

Abbreviations

AD	Alzheimer's Disease
AGO	Argonaute
ALS	Amyotrophic Lateral Sclerosis
APOBEC	Apolipoprotein B mRNA-editing enzyme catalytic polypeptide
BAM	Binary Alignment Map
BED	Browser Extensible Data
Bp	Base pair
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CUT&RUN	Cleavage Under Targets & Release Using Nuclease
dsDNA	Double stranded DNA
dsRNA	Double stranded RNA
EM	Expectation Maximization
ERV	Endogenous Retrovirus
GSEA	Gene Set Enrichment Analysis
GTF	General Feature Format
H3K4me3	Tri-methylation of the 4th lysine of histone H3
H3K9me3	Tri-methylation of the 9th lysine of histone H3
hGPC	Human Glia Progenitor Cells
HUSH	Human Silencing Hub
IFN	Interferon
iPSC	Induced Pluripotent Stem Cells
kbp	Kilobase pairs
KO	Knock-out
KRAB-ZNF	Krüppel associated box zinc finger
LINE	Long Interspersed Nuclear Element
lncRNA	Long non coding RNA
LTR	Long terminal repeat
MAVS	Mitochondrial Antiviral Signalling
MHC	Major Histocompatibility Complex
miRNA	Micro RNA
MRI	Magnetic Resonance Imaging
mRNA	Mature RNA
MS	Multiple Sclerosis
NB	Northern Blot
NPC	Neural Progenitor Cells
ONT	Oxford Nanopore Technologies
ORF	Open Reading Frame

PacBio	Pacific Biosciences
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PD	Parkinson's Disease
piRNA	PIWI-interacting RNAs
Pol-II	Polymerase II
PRR	Pattern Recognition Receptors
RISC	RNA Induced Silencing Complex
RPKM	Reads Per Kilobase per Million mapped reads
SAM	Sequence Alignment Map
SINE	Long Interspersed Nuclear Element
siRNA	Small Interfering RNA
ssDNA	Single Stranded DNA
ssRNA	Single Stranded RNA
STING	Stimulator of interferon genes
SVA	SINE-VNTR-Alu like element
TBI	Traumatic Brain Injury
TE	Transposable Elements
TES	Transcription Ending Site
TLR	Toll-Like Receptor
TSD	Tandem Site Duplication
TSS	Transcription Start Site
UMAP	Uniform Manifold Approximation and Projection for Dimension Reduction
UMI	Unique Molecule Identifier
UTR	Untranslated Region
VNTR	Variable Number of Tandem Repeats
XDP	X-linked Dystonia Parkinsonism

Introduction

Preface

Our genome hosts an abundant and diverse group of mobile sequences called transposable elements (TEs, also called transposons). TEs are present in the genomes of nearly all species sequenced to date. The abundance of TEs across the tree of life may be attributed to their remarkable ability to expand within their host genomes and their vertical (and in some species even horizontal) transmission [5,6].

TEs account for around 50% of the human genome (Figure 1)[2]. We classify TEs according to their mobilisation mechanism – either as a retrotransposon (class I) or DNA transposon (class II)[5,6].

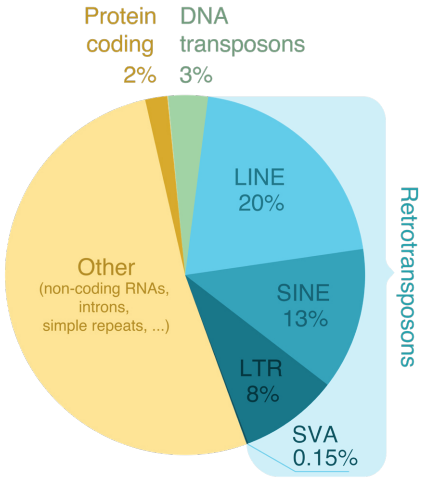


Figure 1 Pie chart showing sequence content of human DNA.

Data extracted from Hoyt et al. (2022)[2].

Retrotransposons mobilise in a copy-and-paste manner – transcribed and retrotranscribed into a different genomic location, creating an exact copy of themselves[5,6]. On the other hand, DNA transposons mobilise in a cut-and-paste manner – "moving" to a different genomic location. DNA transposons are not able to mobilise in the human genome, but a small subset of retrotransposons retain retrotransposition potential and continue to produce inter-individual variation in the human population[7, 49]. Retrotransposons are the focus of this thesis.

Retrotransposons are further classified as elements that either contain long terminal repeats (LTR) or do not (non-LTR). LTRs are regions of several hundred base pairs that typically flank retroviral ORFs and contain regulatory sequences sufficient to drive their expression. LTR retrotransposons include so-called endogenous retroviruses (ERV). However, retrotransposition-competent ERVs have not been identified in humans[8]. Non-LTR elements do not contain flanking LTR sequences but contain other regulatory sequences to initiate transcription. Non-LTR elements are classified into different superfamilies: Long Interspersed Nuclear Elements (LINEs), Short Interspersed Nuclear Elements (SINEs), and SINE-VNTR-Alu elements (SVAs) (Figure 1)[5,6]. Each non-LTR retrotransposon superfamily is further classified by evolutionary age into families and subfamilies, as exemplified in Figure 2.

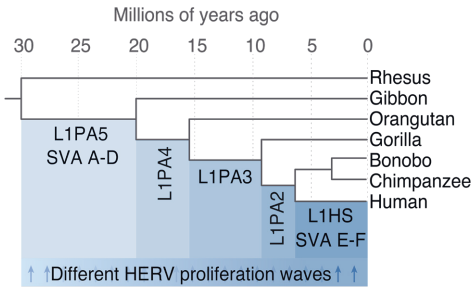


Figure 2 Phylogenetic tree of selected primates and incorporation times of evolutionary young L1, SVA and HERV subfamilies.

Adapted from Garza et al. (2023).

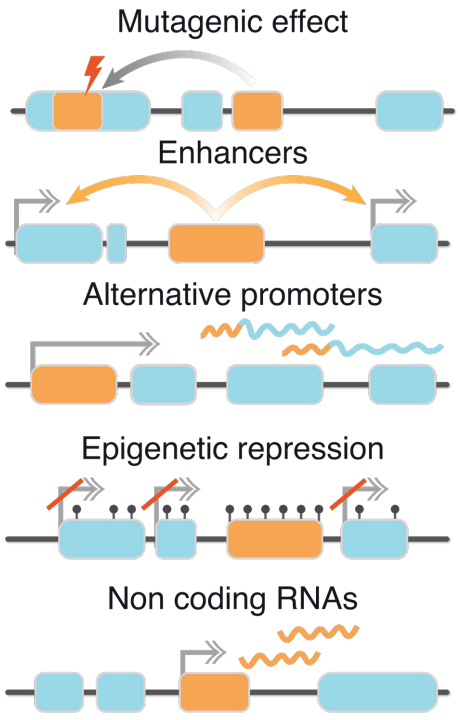


Figure 3 Genomic, transcriptional, and epigenetic changes introduced by TEs.

Adapted from Jönsson et al. (2020)[1].

TEs contain regulatory sequences that initiate transcription of the element and may act as alternative promoters or enhancers to nearby host genes (Figure 3)[6,9]. This can give rise to new isoforms of both protein-coding and non-coding RNA. Retrotransposons are also a source of alternative splice sites (e.g., exonized TEs or alternative 3' or 5' sites) and homologous recombination sites (common among LTR or SINE sequences), and are able to modulate the levels of gene transcription (e.g., via local epigenetic changes, or acting as enhancers)[1,9].

The expression and regulatory activity of TEs varies widely between tissues[10,11], developmental time points[12-14], and disease conditions[9,15]. Furthermore, several observations suggest a link between the activity of TEs and a cell's state[16] or type[17], and have been implicated in the evolutionary origin of entire cell lineages, such as the neural crest[18]. These have been challenging topics to address, particularly as single-cell technologies have severe limitations regarding TE analyses (reviewed below).

The work presented in this thesis focuses on bioinformatic approaches for analysing retrotransposons (specifically, L1NE-1, SVA, and ERV subfamilies) to account for their repetitive nature, and special considerations for each class of retrotransposon (hereon referred to as TEs).

Friends or foes

Given that new TE insertions are not always for the host's benefit and, on the contrary, are often at the expense of the host's genomic integrity, TEs have long been thought of as 'selfish' genetic elements. However, TEs have also played a major role in genome evolution. Owing to their repetitive nature and abundance, TEs are prone to undergo recombination, which has resulted in major genomic rearrangements[19-21]. More so, the

intrinsic regulatory properties of TEs have shaped gene regulatory networks involved in development[13,22], pluripotency[23,24], and inflammation[25]. Thus, despite their intrinsic 'selfish' nature, TEs drive genomic evolution and have resulted in the development of beneficial traits for the host. More so, TEs represent an important research avenue to study the evolution of species-specific traits[26], such as the expansion and increased complexity of the human brain[27-29].

The activity of TEs is correlated to the development of human disease and is associated with disruption of transcription, splicing and/or translation, or altering the epigenetic status of a region[15]. Importantly, several neurological disorders have been associated with an increased TE expression, including Multiple Sclerosis (MS)[30-32] and Amyotrophic Lateral Sclerosis (ALS)[33,34]. The role of TEs in these and other associated diseases is yet to be defined; however, an interesting hypothesis is that an increased TE expression triggers or is the result of an innate immune reaction. Other neurological disorders, on the other hand, are directly caused by a TE insertion, including X-linked Dystonia Parkinsonism (XDP)[1,15,35].

Mechanistic studies have shown that under some conditions, neural cell cultures support TE retrotransposition events[16,36], and several studies have tried to characterised de novo retrotransposition events from postmortem human brain material[37-41]. Little is known, however, about the extent to which TE activity has shaped the human brain transcriptome, or how the expression of TEs is modulated through life, and if this changes in a disease context.

Human TEs

The work presented in this thesis relates to LINE-1, SVA, and HERV subfamilies. Here, I summarise their sequence structure, evolutionary history, polymorphism rates, and known effects on gene expression – which are all relevant considerations for the experimental design, data analysis, and interpretation of the results.

LINE-1

Long interspersed nuclear element 1 (LINE-1 or L1) elements constitute 85% of LINE elements (~17% of the human genome)[2]. L1s are classified into subfamilies defined by evolutionary age, e.g., the L1HS subfamily consists of human-specific elements, while the L1PA2 subfamily is present in human, chimpanzee, and bonobo genomes (Figure 2). The structure of a full-length L1 includes two untranslated regions (3' and 5' UTRs) and three open reading frames (ORF0, ORF1, and ORF2) and a polyadenylation signal. The 5' UTR harbours essential binding motifs for several transcription factors required for transcription initiation, such as YY1[42], as well as a sense promoter (for RNA polymerase II, or 'Pol II') and a primate-specific antisense promoter (also Pol II) that initiates the transcription of ORF0 (Figure 4)[43]. ORF0's function is unknown, but it is suggested to enhance L1 mobility[43]. The antisense promoter upstream of ORF0 on the antisense strand can act as an alternative promoter of neighbouring genes, which can result in L1-gene fusion transcripts that may translate to L1-derived peptides[43]. ORF1 encodes for an RNA-binding protein, and ORF2 encodes for an endonuclease and reverse transcriptase[44]. The L1 structure ends with the 3' UTR and a poly-A tail[44,45].

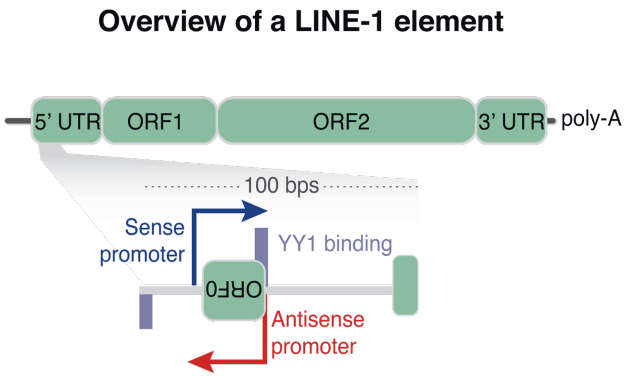


Figure 4 Schematic of L1 structure.
Figure taken from Garza et al. (2023).

Due to interrupted reverse transcription, most L1s in the human genome are 5' truncated (or mutated) and retrotranspositionally incompetent[46]. However, a small subset of L1 elements in the human genome (80-100 elements) retain an intact sequence and are, therefore, retrotranspositionally competent in the event that they escape restriction mechanisms such as epigenetic silencing[47,48]. It is estimated that a new germline polymorphic L1 insertion occurs once in every 63 births[49]. There are also documented somatic de novo L1 insertions, including in the brain, however, the frequency of these events remains a debated topic in the field[37-39,41,50].

L1 is the only autonomously mobilising family of non-LTR retrotransposons in humans, while SINE and SVA elements depend on the L1 machinery for their own retrotransposition. Consequently, L1s, SINEs, and SVAs are preferentially introduced into the loose L1-endonuclease recognition motif (5'-TTTT/AA-3'). Hallmarks of their retrotransposition include flanking target-site duplications (TSDs), and long poly-A tails[44,51].

Increased L1 expression or mobility has been documented in human diseases, including most epithelial-derived cancers[52]. L1s have been suggested to play a (still undefined) role in neurological disorders such as Rett Syndrome[53,54] and Ataxia Telangiectasia[55], as well as in neurodevelopmental and psychiatric disorders with a less defined genetic component, such as schizophrenia[56] and major depression disorder[57].

SVA

Known as "repeat of repeats", SVAs are hominoid-specific elements that span from 700 bp to 4kbp, with an average of 2kbp in the human genome. The structure of an intact SVA (likely transcribed by Pol II) starts with a hexameric repeat (CCCTCT), followed by an antisense sequence of two Alu fragments, a Variable Number of Tandem Repeats (VNTR), a SINE-R domain (derived from the env gene and 3' LTR of an HERV-K10), and ends in a poly-A tail of varying length; however, readthrough transcription past the poly-A tail is commonly observed (Figure 5)[51,58].

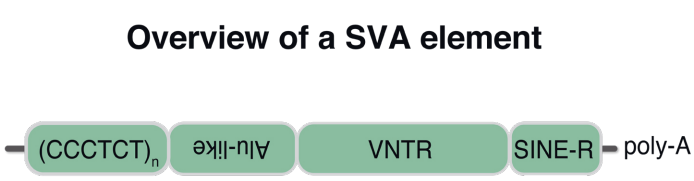


Figure 5 Schematic of SVA structure.

SVAs originated in hominoids, but the exact evolutionary history of their structure is unknown. Similarly to L1s, SVA subfamilies are defined by their evolutionary age (A to F; oldest to youngest). The human-specific subfamilies (SVA-E and SVA-F) remain active, with an estimated germline insertion rate of one in every 60 births[7,49,51].

The transcription start site (TSS) of an SVA is thought to reside in its 5' end, but an internal promoter has not yet been defined. However, SVAs have enhancer effects on nearby gene expression (commonly attributed to the SINE-R domain) which could serve to initiate the SVA's transcription from an external upstream promoter[51,58]. SVAs have the potential to affect nearby gene expression by acting as enhancers, exonizing, creating alternative splice sites in a host gene, or by changing the epigenetic landscape of a region[13,38,58]. Changes in the host transcriptome can have a detrimental effect, as illustrated by different brain diseases such as XDP[35], Neurofibromatosis 1 and 2, and Fukuyama-type muscular dystrophy[15,51].

HERV

Human endogenous retroviruses (HERVs) are remnant sequences of retroviral insertions that infected a germ cell and became "endogenised" into the host's offspring[8,59]. Unlike exogenous retroviruses, HERVs have lost the ability to horizontally transfer in humans and can only be transferred vertically[8,59]. A complete HERV provirus is composed of two flanking LTRs and an internal region containing viral genes encoding for the matrix and capsid of the retrovirus (gag), a protease (pro), a reverse transcriptase (pol), and a surface protein (env) (Figure 6). LTRs contain promoter sequences for Pol II which initiate the transcription of the internal region of the HERV[5,8], or can give rise to downstream chimeric transcripts[3].

Flanking LTRs are identical at the time of insertion, which makes them prone to undergo homologous recombination. This has resulted in thousands of solo LTRs in the human genome, as well as internal HERV sequences with one or no LTRs. Thus, the presence and sequence similarity between the flanking LTRs is commonly used for the estimation of the evolutionary age of a HERV[8,59]. HERVs in the human genome integrated as exogenous retroviral infections, most of which occurred between 30 and 70 mya. Thus, HERVs are primate-specific retrotransposons (not human-specific as their name suggests)[8,59].

HERVs and solo LTRs have regulatory sequences that can affect nearby gene expression or alter splicing formation[13,59,60]. Interestingly, increased HERV transcription and presence of HERV-derived proteins has been detected in several human diseases, including cancer[59,61], Alzheimer's Disease (AD;

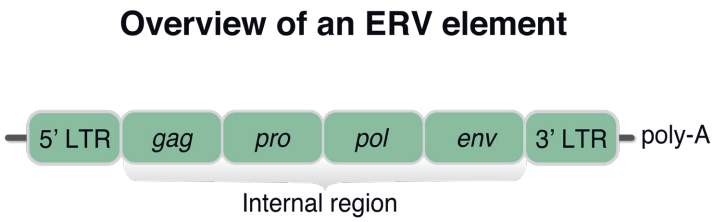


Figure 6 Schematic of ERV structure

associated with an elevated HERV-W expression)[62], ALS (associated with an elevated HERV-K expression)[33,34,62], and several autoimmune disorders including MS (associated with an elevated HERV-H and HERV-W expression)[31,63].

Bioinformatic considerations of TEs

Given their repetitive nature, TEs pose a challenge for common molecular technologies to accurately define their positions in the genome, to discriminate their transcription to that of genes, and to characterise their regulatory effect on nearby genes. Thus, TEs have been ignored in many scientific studies, including during the development of many bioinformatic tools where TEs are "masked" or discarded[3,6].

In the past decade however, the TE-field has benefited from several technological advancements, including the increased availability of long-read sequencing. Long-read sequencing has significantly improved the accuracy of genome assemblies, and has allowed a better annotation of repetitive regions, including TEs[2,38]. Nonetheless, short-read sequencing is still an affordable and common sequencing platform, and accounts for most of the publicly available datasets. The exclusion of TEs from scientific studies and tools has already resulted in a technical and scientific debt and, thus, it is crucial to establish bioinformatic approaches for analysing short-read sequencing from a 'TE-centric' view.

Technology status and limitations

Measuring the transcription of repetitive elements is not a trivial task. Commonly used technologies like Real Time quantitative Polymerase Chain Reactions (RT-qPCRs), microarrays, Northern Blots (NB), or short-read RNA sequencing all face the challenge of the repetitive nature of TEs, which often results in a lack of specificity of their measurements[3].

TEs are often part of chimeric transcripts, which primer-based technologies fail to account for – providing no distinction between passive transcription and one initiated by a TE promoter. This ambiguity will yield confusing and often over- or under-estimations of TE transcription. Locus-specific resolution is important to investigate TE's function, and over 33% of protein-coding genes contain at least one exon of TE-derived origin. Moreover, TEs have played an essential role in the emergence of many long non-coding RNAs (lncRNAs), with over 75% of lncRNAs containing one TE-derived exon[3,64].

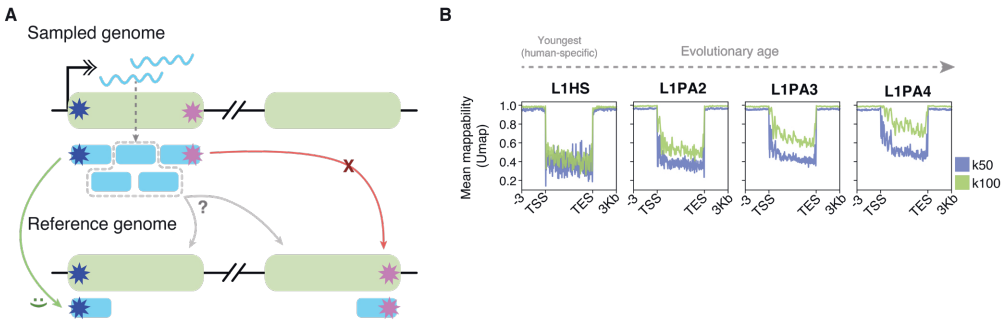


Figure 7 Mappability issues of TEs

A) Mapping ambiguity to reference genome (adapted from Lanciano & Cristofari, 2020)[3]. B) Mappability (Umap, genome mappability scores[4]) of selected evolutionary young L1 subfamilies (elements >6kbp) with 50 bp (blue) and 100 bp (green) read lengths (single end, unique mapping).

Given the high degree of sequence similarity and common repetitive domains within TEs, primer design for a specific TE (or TE subfamily) is challenging. Consequently, technologies such as RT-qPCR, microarrays, or NB, are likely to face off-target signals[3].

Contrastingly, RNA sequencing enables measuring TE expression on a genome-wide scope without the need for primer design. By including crucial information such as transcript orientation and junction reads, RNA sequencing can also partially resolve the ambiguity of co-transcription between genes and TEs. Nevertheless, short-read RNA sequencing technology is still limited by TE mappability and polymorphisms (Figure 7 and 8)[3,6].

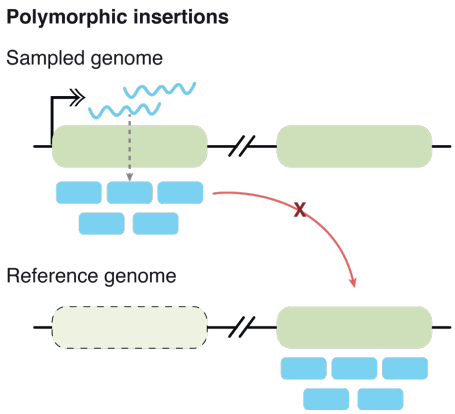


Figure 8 Mapping problem of polymorphic insertions to reference genome

Adapted from Lanciano & Cristofari, (2020)[3].

Mappability of TEs

Given the repetitive nature of TEs and their large number in the genome, the mapping of short reads to the reference genome can easily result in a large fraction of ambiguously-mapped reads. The accumulation of mutations over time, however, increases the percentage of divergence from a TE to its subfamily and, thus, the so-called 'mappability' to a single position in the genome. In other words, the evolutionary age of an element correlates with its mappability (Figure 7)[3,6].

TE subfamilies differ in structure: some contain conserved domains, suffer from common truncations, or contain simple repeats within them[5,50,51,59]. Thus, read length positively correlates with the mappability of a TE (Figure 7B).

Read length can also benefit the mapping of DNA-based sequencing data, where reads can span outside of the TE to its surrounding regions, increasing its likelihood to be mapped unambiguously to the genome.

Importantly, mapping sequencing data to a reference genome can be better phrased as mapping each read to the "best matching" location in the reference genome. Thus, reads originating from polymorphic insertions (not present in the reference genome) will be wrongly assigned to the best matching location in the reference (Figure 8). Moreover, not only polymorphic elements but also internal polymorphism events (e.g., variable length of VNTR regions on an SVA) may result in the wrong assignment or complete discard of a read when mapping to a reference genome[3].

Despite long read sequencing not being as commonly used by the scientific community as short read sequencing (yet), the advantages of using long reads are most apparent when studying repetitive regions of the genome, such as TEs[3,6]. Limitations from long read sequencing technologies, for example developed by Pacific Biosciences (PacBio), Oxford Nanopore Technologies (ONT), and Illumina, depend on the specific technology and include:

- Large amounts of input material, making it hard to couple with single-cell technologies, or to use with scarce material.
- Elevated cost when compared to short-read sequencing.
- Error rates in some platforms.
- Limited computational tools for 'TE-centric' analyses (so far).

These technologies, however, undergo rapid advancements.

Notably, PCR amplification and library preparation steps in both short and long read sequencing may result in technical artefacts that can lead to the detection of false positives (non-existing TE insertions)[41]. Long read sequencing, however, allows for a more careful validation of TE insertions, where reads span the full insertion, allowing for the manual curation of retrotransposition hallmarks [38,65].

TE analyses from single cell RNA sequencing

Single-cell technologies have become an integral component for many research studies focusing on complex tissues with heterogeneous populations of cells, such as the brain[66,67]. 10X libraries for RNA sequencing provide single-cell gene expression information in a high-throughput manner, easily reaching thousands of cells per sample. The technology isolates individual cells (or nuclei) on lipid droplets where identifying primers for each droplet (cell barcode) and molecule (unique molecule identifier (UMI)) are added, effectively tagging each molecule in each cell[68]. Sequencing of 10X libraries results in reads of 90bps of length (excluding cell barcodes and UMIs), with around 40K-100K UMIs per cell in our hands.

An important consideration for TE analysis using 10X data is the poly-A dependent capture, which results in an enrichment over the 3' end of the transcripts[69]. This inherent 3' enrichment can hinder the recognition of a transcript's isoform or TSS. Moreover, 3' enrichment will also further limit mappability to TEs, especially for evolutionary young elements or TE subfamilies with inherently more repetitive, or conserved, 3' ends.

10X sequencing captures a substantial amount of premature RNA (23% of total reads) due to oligo mispriming, which is sufficient to even predict the future fate of a cell by comparing the levels of mRNA vs premRNA (also known as cell velocities)[66,70]. This is likely of benefit for TE analyses, as mispriming may capture signal from the body of the element.

Given 10X technology's specifications, a unique mapping approach to quantify the expression of TEs (see methods: *Mapping strategies*) would likely result in a sparse matrix with low statistical power. Therefore, special considerations for TE analyses using this technology include multi-mapping of reads and quantification of TEs per subfamilies[71].

Assuming there are distinct groups of transcriptionally similar cells, one could increase quantification and statistical power by pooling reads from a group of cells into a pseudo-bulk (see results: *Accurate TE subfamily quantification from single cell RNA sequencing technologies using pseudo-bulks)*[72]. Notably, given the technology's limitations, orthogonal validation of TE expression – e.g., epigenetic status to validate usage of promoters, bulk RNAseq, SMRTseq, long-read sequencing, etc[3] – is valuable.

Epigenetic regulation of TEs

To maintain genomic integrity, sophisticated machinery has evolved to regulate the transcription of TEs. The work presented in this thesis relates to two types of heterochromatin formation that are deposited over TEs: H3K9me3 and DNA methylation (which, in many cases, co-exist).

Transcriptional silencing of TEs via H3K9me3 deposition

Tri-methylation of the 9th lysine of histone H3 (H3K9me3) is commonly found over lowly expressed or completely silent regions of the genome. The presence of H3K9me3 compacts DNA into heterochromatin, preventing transcription initiation of the genomic region[73]. TEs are enriched for H3K9me3 – a state dependent on SETDB1-associated complexes including Krüppel associated box (KRAB) zinc finger proteins (hereon referred to as KRAB-ZNF) and the Human Silencing Hub (HUSH)[74-76].

KRAB-ZNFs are the largest transcription factor family in higher vertebrates with over 350 genes in the human genome[77]. Their structure includes an N-terminal KRAB domain and DNA-binding zinc fingers that target TE-derived sequences. The evolution of KRAB-ZNFs and TEs is entwined and has been described as an evolutionary arms-race, where KRAB-ZNFs evolve to target younger TEs while TEs evolve to escape the repression of KRAB-ZNFs[58,78]. Individual KRAB-ZNFs show preferential binding to a group of target-TEs (a TE subfamily or small group of TE subfamilies). If these target-TEs accumulate enough mutations over time, or become incapable of transcription, the KRAB-ZNF can become co-opted for other purposes, e.g., to repress gene expression[77-79]. Studies in cell culture models have shown that, upon binding to the DNA, KRAB-ZNFs recruit TRIM28 (also known as KAP1). TRIM28 acts as a scaffold protein for SETDB1, HP1, and the NuRD complex. SETDB1 deposits H3K9me3, HP1 facilitates heterochromatin spreading, and the NuRD complex deacetylates local chromatin (Figure 9)[74,75,80].

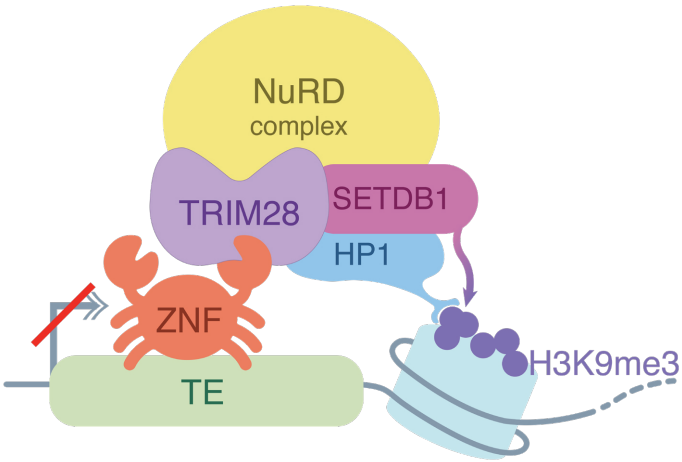


Figure 9 Cartoon representation of the KRAB-ZNF-mediated deposition of H3K9me3[148].

The HUSH complex can also silence TEs by establishing H3K9me3. HUSH is a vertebrate-specific complex of transcriptional suppressor proteins, i.e., transcriptional activator suppressor (TASOR or FAM208A), M-phase phosphoprotein 8 (MPP8), and periphilin (PPHLN1)[76]. The complex recognizes nascent RNA over H3K9me3-rich regions and recruits SETDB1 to further deposit H3K9me3[76]. HUSH binds and represses evolutionary young L1s[81-83], and has been suggested to also target young ERVs, SVAs, and certain repetitive genes[81,82].

Transcriptional silencing of TEs via DNA methylation

DNA methylation, specifically 5mC (hereon referred to as DNA methylation), refers to the addition of a methyl group to the 5-position of cytosines[84]. The presence of this epigenetic mark over CpG islands weakens or blocks the binding of methylation-sensitive transcription factors, which usually leads to the transcriptional silencing of a region[84]. Conversely, regions depleted from DNA methylation are more accessible, which allows for the binding of different transcription factors that in turn protect the CpG island from the deposition of DNA methylation. The presence of DNA methylation is enriched over TE sequences and has an inactivating effect over their promoters, which usually have CpG islands. DNA methylation has even been suggested to have evolved for the transcriptional silencing of TEs[85]. DNA methyltransferases are a group of enzymes responsible for the addition of the methyl group to the cytosine residues of the DNA. While DNMT3a and DNMT3b are responsible for the de novo deposition of DNA methylation, DNMT1 is responsible for the maintenance of DNA methylation upon cell division – thereby ensuring inheritance of the parent cell's DNA methylation pattern (Figure 10)[84].

In preimplantation development (and later on in primordial germ cells), DNA methylation is erased and reestablished by DNMT3a, DNMT3b[84]. During this process, some TEs are silenced via H3K9me3 and others become transcriptionally active. In vitro studies have shown that deletion of

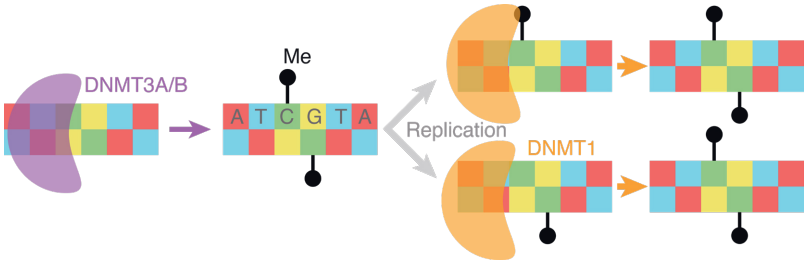


Figure 10 Cartoon representation of DNA methyltransferases establishing de novo DNA methylation (DNMT3A/B in purple) and maintenance upon cell division (DNMT1 in orange).
Adapted from Jönsson et al. (2020)[1].

DNMT1 in human embryonic stem cells results in cell death[86], while human neural progenitor cells (NPCs) remain viable[87]. Upon deletion of DNMT1, NPCs undergo a massive transcriptional activation of evolutionary young L1 elements – but not their older counterparts or other classes of TEs[87]. This is an important consideration for research concerning ageing, or disease contexts such as cancer, where DNA methylation undergo changes[52,88].

Other silencing or inactivating mechanisms

The work presented in this thesis relates to TE silencing via DNA methylation and H3K9me3 deposition. There are other mechanisms, however, through which TEs lose their transcriptional potential. Some examples are briefly summarized below.

APOBEC

Apolipoprotein B mRNA-editing enzyme catalytic polypeptide (APOBEC) proteins can introduce mutations from cytidines (C) to uridines (U) in ssDNA and RNA. Depending on the mutation site, this process can impede TE retrotransposition[89]. APOBEC proteins are thought to reduce TE mobility[90-92].

Small RNA silencing

The RNA Induced Silencing Complex (RISC) is composed of an Argonaut protein (AGO) and a small RNA molecule which serves as a guide to its target; these small RNA molecules include micro RNAs (miRNAs), small interfering RNAs (siRNAs), and PIWI-interacting RNAs (piRNAs)[93]. Upon binding to their targets, RISC can cut the mRNA, recruit proteins to shorten its poly-A tail, and block its translation[93]. Similarly, it can target complementary nascent RNA and recruit relevant proteins to deposit DNA methylation or establish H3K9me3 over the transcribed region[93,94]. This is particularly important for TEs and viruses, as dsRNAs can be fragmented and used as a guide to complementary RNA molecules[93,95].

PiRNAs are mostly expressed in the germline and are encoded from piRNA-clusters which are rich in TEs. RNA generated from these TE sequences is used to guide the silencing complex to transcriptionally active TEs (known as the "trap model"). Once a piRNA has found a complementary TE-RNA target, its surrounding regions are leveraged to create new piRNAs – extending the piRNA repertoire to sequences not present in the original piRNA cluster[93].

Accumulation of mutations

The natural accumulation of mutations over time can also result in the deterioration or loss of sequences required for the transcription of a TE and/or its retrotransposition – namely, promoter regions and coding sequences for the retrotransposition machinery. An estimated mutation rate of the human genome is 0.2% per million years[139, 140]. Different families of TEs undergo common mutations. For example, most L1 insertions are characterized by 5' UTR truncations of the element which lead to loss of the promoter[46]. LTR elements, on the other hand, can be fully inactivated via homologous recombination events, resulting in solo LTRs[59].

Potential consequences of TE transcription

If the silencing machinery fails to suppress TE transcription, TEs pose a mutagenic threat and have the potential to interfere with gene regulatory networks[87,88,97]. The polymorphic insertion of TEs can lead to the development of diseases such as XDP – a neurodegenerative disorder caused by an SVA insertion into the TAF1 gene[98]. The transcriptional activation of TEs may also trigger an innate immune response[99]; however, their role in the beginning of human neuroinflammation remains unknown.

Polymorphic TE insertions as a trigger for disease development

X-linked dystonia Parkinsonism (XDP)

XDP is a rare, recessive, neurodegenerative disorder endemic to Panay, Philippines. It is a motor disorder that develops into dystonia and Parkinsonism at a later stage. MRI studies and post-mortem examination of XDP patients have shown prevalent striatal-neuron death, which correlates with the progression of the disease[100,101].

Genetic linkage analyses have attributed the development of the disease to an SVA insertion into the TAF1 gene encoding an important transcriptional activator whose expression is mildly decreased in XDP patients. The ~2.6kbp long SVA element sits in the 32nd intron of TAF1 and has a variable length of its hexameric repeat (CCCTCT)n which negatively correlates with the patient's age at disease onset[98,102,103]. The presence of the SVA insertion results in the retention of the 32nd intron and the decreased expression of the TAF1 3' exons[35,104]. How the intronic SVA insertion affects TAF1 expression is not yet fully understood but exemplifies how a polymorphic insertion can cause a neurological disorder.

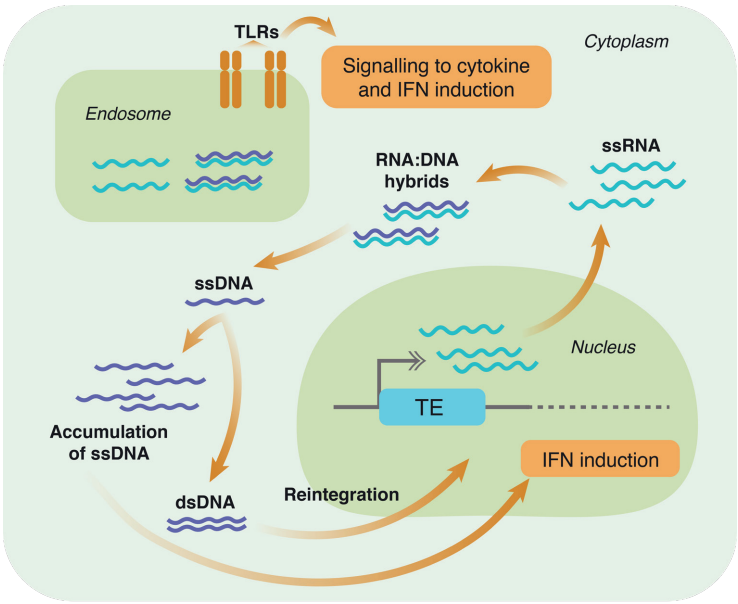


Figure 11 Overview of IFN induction via TE expression.
Adapted from Jönsson et al. (2020)[1].

TEs as immunogenic sequences

Toll-like (TLR) and other pattern recognition (PRRs) receptors are sensitive to the presence of intermediate or finished products from the retrotransposition process of TEs including single-stranded RNA (ssRNA), DNA:RNA hybrids, single stranded DNA (ssDNA), cytosolic double-stranded DNA (dsDNA), and TE-derived peptides such as L1-ORF1p or L1-ORF2p or Env, Gag, Pro and Pol from ERVs (Figure 11)[99,105,106]. TE-derived RNA molecules engaged by RIG-I and MDA5 (both cytosolic sensors of viral RNA) can activate MAVS[107]. MAVS activation triggers a type I interferon (IFN-I) response, specifically, IFN-α and IFN-β[107,108]. Cytosolic dsDNA or loops formed from DNA:RNA hybrids or ssRNAs can be recognized and bound by cGAS[109,110]. This reaction triggers the production of cGAMP, which directly activates STING. STING acts as the master regulator for response factors that are responsible for the triggering of an innate immune response in the form of an IFN-I response (IFN-α and IFN-β)[109]. IFNs (including IFN-γ (INF-II), another antiviral cytokine produced by activated immune cells[93,111]) are secreted and can bind to cell-surface receptors of neighboring cells. The recognition of IFNs engages the JAK-STAT signaling pathway, which further aggravates the innate immune response[93]. Several diseases that are postulated to have an immune component – including ALS[33,34] and MS[30,31] – have been documented to correlate with an elevated TE expression[63], which may have relevance to disease progression[112-114].

Interestingly, there are examples of TEs being co-opted and aiding the innate and adaptive immune system for the host's benefit[115,116]. Thus, the involvement of TEs in the evolution of our immune system remains an active field of research[25,117].

TEs in relation to a neuroinflammatory state

Traumatic brain injury (TBI)

As a leading cause of disability and morbidity worldwide, TBI is responsible for over 1.5 million hospitalisations and 57000 deaths a year in the European Union only[118]. A recent study including 16 European countries estimated TBI to cost the individual an average of 24.3 years of life[119]. The impact to the head causes inflammation and swelling of the brain that can cause increased intracranial pressure, which can require decompressive surgery[120]. The brain injury is characterised by cell death (both neuronal and glial), axon injury, the activation of microglia, and recruitment of other immune cells. Long-term outcomes for TBI survivors include white matter degeneration, persistent inflammation, and an increased risk to develop neurodegenerative disorders such as AD and Parkinson's disease (PD)[121-123].

Despite its prevalence and long-term consequences, little is known about the molecular response that triggers and maintains the neuroinflammatory state upon TBI and neurodegenerative diseases. Increased understanding of this molecular response would lead to better treatments to mitigate long-term consequences for patients. Research on the development of neurodegenerative disorders has been hindered by the inevitable late disease stage from the available post-mortem material. TBI samples, on the other hand, represent a unique model to study the molecular cascade being triggered at the very beginning of a neuroinflammatory response, which may elucidate important cues for research on neurodegenerative disorders. Notably, an increased TE expression has been reported to correlate with different brain diseases that hold a neuroinflammatory state[30,34,62]. As immunogenic sequences expressed in the human brain, it is relevant to understand the role of TEs at the beginning of human neuroinflammation, for which TBI tissue samples present a unique opportunity.

Aims of the thesis

The overall aim of this thesis is to describe the functional relevance of retrotransposons in human brain development and their role in neuroinflammation.

The specific aims of the thesis are to:

- Develop bioinformatic approaches to handle next and third generation sequencing data to study retrotransposons.
- Investigate whether retrotransposons play a role in neuroinflammation.
- Assess the impact of retrotransposons for the transcriptional complexity of the human brain.

Materials and Methods

The work I have performed in this thesis has been in relation to the bioinformatic analyses of transposable elements. In this section I summarise the methods I have learned and developed for this purpose, as well as the sequencing strategies that were used to produce the data used in this thesis.

Sequencing strategies

RNA-based

Short-read bulk RNAseq

Read length: 2x150bp.

Total RNA was isolated using RNeasy Mini Kit (Qiagen). Libraries were produced using Illumina TruSeq Stranded mRNA library preparation kits with poly-A selection and were sequenced on a NextSeq500 or a Novaseq6000 machine.

Long read bulk RNAseq

Read length: ~2kbp

Total RNA was isolated from the samples using miRNA Easy Mini Kit (Qiagen). Libraries were prepared as instructed by PacBio in the "Procedure & Checklist – Iso-seq™ Express Template Preparation for Sequel® and Sequel II Systems" (PacBio, PN-101763800 Version 02 (October 2019)). The libraries were sequenced on a Sequel II and Sequel IIe System, using a Sequel II Sequencing Plate 2.0.

Short-read single nuclei RNAseq

Read length: 90bp

Nuclei were isolated from brain tissue (fresh or frozen) or organoids, as previously described[124]. Nuclei (or whole cells) were loaded onto the 10X Genomics Single Cell 3' (or 5') Chip to create single-cell/nucleus gel beads in emulsion. Libraries were multiplexed and sequenced on a Novaseq6000 machine with a 150-cycle kit, using the read length recommended by the manufacturer (10X Genomics).

DNA-based

Whole genome Nanopore sequencing

Mean read length: ~15kbp

DNA was extracted from frozen cells using Nanobind HMW DNA Extraction kit from PacBio. Libraries were constructed using the SQK-LSK109 Ligation Sequencing kit from ONT and FLO-PRO002 PromethION Flow Cell R9 and sequenced on a PromethION machine (ONT).

Targeted Nanopore sequencing

Primers binding around the region of interest (two primers upstream and two primers downstream of the region in question) were used on a Cas9 sequencing kit (SQK-CS9109) as instructed by the manufacturer (ONT). Libraries were then sequenced using a MinION Mk1Mc using flow cell R9.4.1 (ONT).

CUT&RUN

Read length: 2x75bp

The experiments were performed following a previously described protocol[125]. Sequencing libraries were constructed using the Hyperprep kit (KAPA) with unique dual-indexed adapters and sequenced on an Illumina NextSeq500.

Bioinformatic analyses

Short-read bulk RNAseq

All libraries were demultiplexed and assigned to sample-specific fastq files using bcl2fastq (Illumina; RRID:SCR_015058). Quality control was assessed using FastQC (Babraham Bioinformatics, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Mapping strategies

Arguably the most important step for a transposon-focus bioinformatic analysis (besides the library preparation) is to choose a relevant mapping approach for the data. Given the repetitive nature of TEs, special considerations are to be taken to produce the most accurate TE-mapping out of short-read sequencing. There are two main mapping approaches used in this thesis:

Unique mapping

Unique mapping is a common approach to deal with mapping ambiguity. In this method, reads are allowed to map to a single best-matching location in the reference genome. The read is discarded if it maps equally well to two or more places in the genome. As previously introduced, the number of

ambiguously mapping reads will depend on the sequence similarity to other elements in the genome. Therefore, this method will likely underestimate the expression of evolutionary young TEs.

STAR[126] aligner, well-known for its splicing-aware mapping, soft-clipping, and short running time, was used to map reads from (bulk) short-read RNA sequencing to the human or mouse genome (hg38 or mm10).

Specific parameters for this mapping approach:

- outFilterMultimapNmax set to 1 (default: 20) to limit the number of allowed mapping loci of the read in question (unique mapping).
- outFilterMismatchNoverLmax set to 0.03 (default: 0.3) allowing only 3% of the mapped length to mismatch.
- sjdbGTFfile inputs Gencode's GTF (hg38 or mm10) file to guide the splicing-aware mapping of the reads (optional).
- outFilterMultimapScoreRange left on its default value of 1. Given two possible alignment sites, this parameter defines the difference between the alignment scores for a read to not be considered a multimapper.
- outSAMtype set to BAM SortedByCoordinate to output a BAM file sorted by genomic coordinates.

Multi mapping

A different approach to quantify TE expression is to group their expression by subfamily. This approach is most interesting when studying, for example, the effects of a gene that may affect a large amount of TEs (e.g., TRIM28), but not to study a particular locus (e.g., XDP-SVA).

Specific parameters for this mapping approach:

- outFilterMultimapNmax set to 100 (default: 20) to increase the amount of allowed ambiguous mapping loci for a read.
- outFilterMismatchNoverLmax left on its default value (default: 0.3).
- winAnchorMultimapNmax set to 200 (default: 50). One of the first steps when running STAR is to define "seeds" – pieces of the read that map exactly to the genome. By clustering seeds within the read, "anchor-seeds" are defined. winAnchorMultimapNmax defines the upper limit for the number of loci where an anchor-seed can map in the genome. General recommendations for this parameter is to double the amount of outFilterMultimapNmax[71].
- sjdbGTFfile inputs Gencode's GTF (hg38 or mm10) file to guide the splicing-aware mapping of the reads.
- outSAMtype set to BAM SortedByCoordinate to output a BAM file sorted by genomic coordinate.

All BAM files were indexed using SAMtools[127]. Files were converted to bigwig files using deeptools[128] normalizing by RPKM (normalizeUsingRPKM), and splitting files by strand (filterRNAstrand forward/reverse).

Quantification of reads

FeatureCounts (Subread package)[129] was used to quantify uniquely-mapped reads. Libraries were defined as reversely stranded (-s2) (TruSeq library preparation kits) to only quantify reads in the same direction as the feature in question.

To quantify gene expression, the feature parameter (-a) receives Gencode's GTF (hg38 or mm10).

To quantify TE expression, the feature parameter (-a) receives a GTF version of Repeatmasker's output file "parsed to filter out low complexity and simple repeats, rRNA, scRNA, snRNA, srpRNA and tRNA"[71] (hg38 or mm10).

To quantify ERV elements and proviruses, the feature parameter (-a) receives a GTF version of Retrotector's output (hg38 or mm10)[130].

The specialized software TEcount[71], was used to quantify the expression of TE subfamilies. In a nutshell, the program (--mode multi) performs an expectation-maximization (EM) algorithm to calculate the relative abundance of a particular TE locus as a function of the number of multi-mapping reads and relative abundance of the rest of the TEs in the genome. Count matrices for genes and TE subfamilies are produced summing the relative abundances of a TE subfamily (as calculated by the EM algorithm) and the uniquely-mapping reads of the individual TEs in the subfamily[71].

Differential expression analysis

Differential gene expression analyses were performed using DESeq2 (DESeqDataSetFromMatrix & DESeq)[131]. Shrinked log2FoldChanges (DESeq2:: lfcShrink) were used to perform Gene Set Enrichment Analyses (GSEA) using gseGO (minGSSize = 3, maxGSSize = 800, pAdjustMethod = BH) function of the clusterProfiler R package.

Differential TE expression analyses were performed using DESeq2 (DESeqDataSetFromMatrix & DESeq) using TEcount's output (subfamilies count matrices, only including TE subfamilies), or featureCount's output (individual elements count matrices). For visualization purposes, raw TE counts (by subfamily or by individual elements) were normalized using the sample-specific size factors as calculated by DESeq2 using gene expression (median of ratios).

Short-read single nuclei RNAseq

All libraries were demultiplexed and assigned to sample-specific fastq files using mkfastq (10X Genomics)[68]. Quality control was assessed with Cell Ranger (10X Genomics)[68] and Seurat[132].

Mapping and gene quantification of the reads was performed using Cell Ranger. A custom pre-mRNA genome index (generated following 10X Genomics guidelines for Cell Ranger version 3) was

used for the mapping of some nuclei samples. Latest studies were performed using the --include-introns parameter, as it was introduced in a later version of Cell Ranger.

Manifolds were calculated using UMAP (Seurat::RunUMAP) and clustering of the cells was performed from 10 to 20 precomputed principal components (dataset-dependent) using the nearest neighbour modularity optimization-based algorithm (Seurat::FindClusters). The resolution was defined on a dataset-dependent way (0.1 to 0.5).

Cells/nuclei were filtered out if mitochondria content was over 10% (Seurat's perc_mitochondrial). Thresholds for the number of genes detected were generally set on a sample-specific manner. In general, cells with more genes detected than two standard deviations above the mean and cells with less than a standard deviation from the mean were filtered out.

Special considerations were taken in the TBI study. Given the contrasting levels of data quality between TBI and control samples, we set a lower threshold of 1000 genes detected to all samples. Upper threshold was set to two standard deviations above the mean.

Statistical analyses and expression scores

Enrichment scores for different gene signatures (e.g., immune-related genes) were calculated using the AddModuleScore function from Seurat[132]. Briefly, the function selects a control set of genes per gene in the signature (five control genes randomly selected but matched on expression levels to the gene in question) and subtracts that from the mean expression of the entire gene signature (per cell).

Differential gene expression analyses per cell type were performed using Seurat::FindMarkers (Wilcoxon Rank Sum tests)[132]. Log2FoldChanges were used to perform Gene Set Enrichment Analyses (GSEA) or Gene ontology overrepresentation analyses (project-dependent) using gseGO (minGSSize = 3, maxGSSize = 800, pAdjustMethod = BH) from the clusterProfiler R package[133]. For overrepresentation analyses, differentially expressed genes were input for each cell type (padj < 0.01).

Cell-cycle scores were calculated using Seurat::CellCycleScoring function, using Seurat's cc.gene as input[132].

Velocity

Velocyto run10X[70] was used on default parameters to estimate RNA velocities from single nuclei data – masking for TEs (-m) and providing Gencode hg38 GTF as the guide for features. The resulting loom files were loaded to R. Velocities were calculated using velocyto.R::RunVelocity (kCells = 25, deltaT = 1, fit.quantile = 0.2). The calculated velocities were projected on previously defined UMAP coordinates.

Quantification of TEs

10X sequencing technologies carry important limitations for TE quantification including low number of reads per cell, short-reads, and a 3' enrichment[69]. To overcome some of these issues and perform a quantification of TEs, we sacrificed resolution in exchange of quantification power.

For the purposes of this thesis, a Python package named trusTEr was developed as a workflow to quantify the expression of TE subfamilies over defined groups of cells. Groups can be cell clusters (e.g., defined by gene expression in Seurat), or annotated cell types, and can be further grouped per conditions of samples (e.g., treatment vs control).

The workflow follows seven steps (Figure 12):

1. Collecting reads from a group of cells: Inputs a tabulated file per cluster (for each sample) containing the barcodes of cells to be collected. The function calls subset-bam from 10X Genomics (RRID:SCR_023216) and extracts said barcodes from the sample's BAM file (possorted_genome_bam.bam output from Cell Ranger). The step outputs a BAM file containing all the reads of the cells listed in the input file.

2. Filter duplicates: The function calls a custom script which inputs a BAM file per cluster (previous step's output) and filters out reads that do not have a unique combination of barcode, UMI and sequence. This step outputs a BAM file.

3. File conversion to fastq: Inputs BAM file (previous step's output) and calls bamtofastq from 10X Genomics (RRID: SCR_023215), to output a fastq file with the reads' sequences.

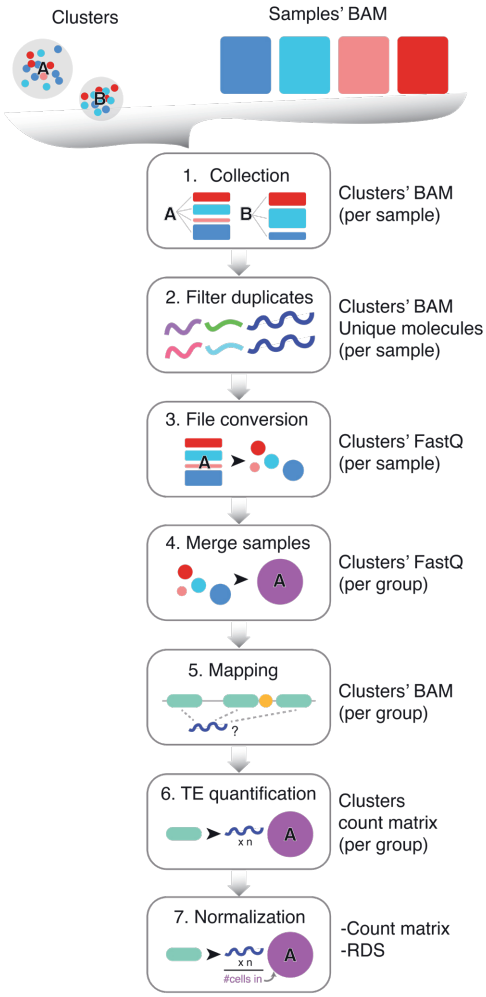


Figure 12. Schematic of trusTEr's workflow for the quantification of TE subfamilies per cells' clusters.

4. Merge samples per cluster: Concatenates fastq files (previous step's output) of each cluster across all samples and outputs a single fastq file per cluster. When the workflow is started, the user can input a dictionary to specify a sample grouping factor e.g., treatment / control groups.

5. Mapping of the reads: By default, multi mapping is performed for each cluster's fastq file (hence the term "pseudo-bulk". This step is performed from the previous step's output) as described above (see methods: *Short-read bulk RNAseq: Mapping strategies*). If the parameter "unique" is set to True, a unique mapping approach is taken. This step outputs a BAM file per cluster.

6. TE subfamily quantification: By default, this function will call TEcount[71] in "multi" mode for a cluster's BAM file. If the parameter "unique" is set to True, quantification is performed by featureCounts[129]. Quantification is performed as described at the methods section *Short-read bulk RNAseq: Quantification of reads*.

7. Normalization per cluster size: After all clusters of cells have been quantified, this function calls a custom R script to normalize the TEcount output matrices by the number of cells in a cluster. The function integrates this information as an assay of an input Seurat object. This step outputs the normalized count matrix as a tabulated file and saves an RDS file containing the Seurat object with the TE assay added.

Long read bulk RNAseq

Mapping of the reads to hg38 was performed using SMRTLink with the default settings for PacBio cDNA sequencing data (Isoseq).

To map reads to L1HS and L1PA2 consensus sequence, a minimap2 index was created to map (-a) the reads using the PacBio HiFi preset (-x map-hifi)[134]. We filtered antisense reads using SAMtools[127] view (-F16), sorted the resulting BAM file using SAMtools sort, and indexed it with SAMtools index. Coverage of mapped reads was visualized in the Integrative Genomics Viewer (IgV).

CUT&RUN

Reads were mapped to hg38 index using Bowtie2[135] for local alignment (--local --very-sensitive-local), only keeping alignments of paired reads (--no-mixed), without discordant alignments (--no-discordant), from 10-700 bps in length (--I 10 --X 700). We used SAMtools to convert from SAM to BAM (samtools view -Sb) and retain only uniquely-mapping reads (MAPQ > 10) (samtools view -q 10). BAM files were sorted and indexed using SAMtools. We used BamCoverage from deeptools[128] to generate bigwig files (--normalizeUsingRPKM).

Homer[136] was used to define regions where the CUT&RUN signal seemed to "peak". Tag directories were produced using makeTagDirectory on default parameters. For the purposes of the projects in this thesis, peak calling was performed using findPeaks with the style parameter set to "histone". annotatePeaks.pl was used to further characterize these regions – or directly intersected them to regions of interest using BEDtools[137] intersect (e.g., intersect with young L1 TSS coordinates).

Long read DNA sequencing

Format conversion

During the time of this thesis, Nanopore stopped using fast5 files and has adopted the pod5 format as output. However, some tools have not been updated to this new standard. Thus, to convert from pod5 to fast5, I used the pod5 python module (specifically, function to_fast5). Indexing of fast5 files was performed using nanopolish index[138].

Mapping to the XDP genome

A custom genome index with a previously reported sequence of the XDP-SVA[102] was created for the purposes of paper IV. Given the methodology used by the authors to extract this sequence, some of the surrounding regions were included. These surrounding regions were used to find the exact position within TAF1 where this SVA insertion was located.

The TAF1 sequence was extracted from hg38 using BEDtools getfasta. Using these two sequences (TAF1 and XDP-SVA), a multiple sequence alignment was performed using clustalw2 and used to identify the breaking point between the TAF1 sequence and the XDP-SVA (chrX:71,440,502).

Two fasta files containing the sequence of chrX before (chrX:1-71440502) and after the insertion break (chrX:71440503-156040895) were created using BEDtools getfasta. The resulting (three) fasta files (chrX-before, XDP-SVA, and chrX-after) were concatenated to produce the "complete" sequence of XDP-chrX.

Reads were mapped using minimap2 (version 2.24) on the Nanopore preset (-a -x map-ont) to the index of the XDP-chrX (built using minimap2 using -x map-ont). Methylation calls were performed using nanopolish call-methylation (version 0.13.2)[138].

Identification of polymorphic insertions

Transposons from Long Reads (TLDR) (version 1.2.2)[38] was used to identify unannotated TE insertions from ONT reads. Briefly, given a TE library (a collection of TE subfamilies' consensus sequences) the tool scans the mapped reads to find deviations from the reference genome (insertions); if the insertion has sufficient sequence similarity to an entry in the TE library, and the surrounding regions match the reference genome, the reads mapping to it are used to produce a local consensus. The tool outputs a table with the coordinates, consensus sequence, and scores of the insertions found.

TLDR was executed with a TE library including consensus sequences of L1Ta, L1preTa, L1PA2, SVA A-F, and HERVK (sequences provided by the developers). We input GRCh38.p13 as the reference genome (-r) and specified for detailed output (--detail_output).

For downstream analyses, insertions were only considered if they were found in the two individuals included in the analysis. Insertions were required to be at least 80% the TE in question (UnmapCover) and have over 80% sequence similarity to the TE consensus sequence (TEMatch), as well as a minimum of three reads to support it (SpanReads).

Custom genome with polymorphic insertions

To create a custom genome (in fasta) that would include all polymorphic insertions, a custom script was used to:

1. Read TLDR's output table.

Per chromosome:

2. Sort insertions from last to first (end to start of the chromosome).
3. Read the chromosome's fasta.

Per insertion (from the last to the first in the chromosome):

a. Split chromosome's fasta into two sequences: before and after the insertion in question.

b. Read local consensus from TLDR output.

c. Concatenate the chr-before + the insertion + and the chr-after.

d. Re-write the chromosome's fasta.

Similarly, to update gene annotation files to fit the custom genome:

1. Read TLDR's output table.

Per chromosome:

2. Sort insertions from last to first (end to start of the chromosome).
3. Subset (to the chromosome in question) and sort the gene annotation file.

Per insertion (from the last to the first in the chromosome):

a. If the polymorphic insertion starts before than the gene (TSS_{gene}) in question, move TSS_{gene} to (TSS_{gene} + $length_{insertion}$).

b. If the polymorphic insertion starts before than the gene ends (TES_{gene}), move TES_{gene} to (TES_{gene} + $length_{insertion}$).

An index using minimap2 index (-x map-ont) was created using the custom genome's fasta. Reads were mapped to this index (-a -x map-ont).

DNA methylation visualization

Methylation calls were performed using nanopolish call-methylation (version 0.13.2)[138]. Methylartist db-nanopolish (version 1.2.2) was used to produce modified base-calls databases and ran the methylartist locus function to visualise specific loci.

Other TE-related analyses

Deeptool's heatmaps

A commonly used approach to visualise large amounts of features at once is using deeptool's heatmaps[128]. This is especially useful for visualising e.g., an entire family or subfamily of TEs. To produce

these heatmaps we used computeMatrix as scale-regions (e.g., to visualize full-length L1s >6kbp) or reference-point (to center the signal over a particular side of the features e.g., --referencePoint center, or transcription start sites [TSS]). When scale-regions were used, the regions' body length (--regionBodyLength) was defined depending on the purpose of the heatmap (e.g., 6kbp for young L1s, 1kbp for SVAs, 9kbp for ERVs, etc). Surrounding windows were usually defined as the same size as the region in question (e.g., for young L1s -a 6000 -b 6000).

When working with stranded RNAseq, the groups of regions of interest, as well as the signal in the input bigwig files were split by strand. The computation of the matrices was performed twice: 1. Forward stranded regions (+) using forward and reverse stranded bigwig files (in that order), 2. Reverse stranded regions (-) using reverse and forward strands (in that order). The two matrices were then bound together using computeMatrixOperation rbind. The resulting matrix contain 2x number of columns (x being the number of samples. 2x meaning: x columns in sense and x in antisense), and y number of rows (y being the number of regions of interest of both transcriptional directions).

The visualisation of matrices was performed using plotHeatmap with specific parameters set on a project-specific manner (--sortUsingSamples and --yMax might vary. Generally, --zMax was set to 1).

Sense vs antisense quantification of TEs

SAMtools flags were used to isolate each strand from the samples' BAM files.

Forward transcription

- First mate in the pair mapping to the reverse strand: -f 80.
- Second mate in pair mapping to the forward strand: -f 128 -F 16.

Reverse transcription

- First mate in pair mapping to the forward strand: -f 64 -F 16.
- Second mate in pair mapping to reverse strand: -f 144.

Quantification of the reads was performed for uniquely mapped reads (see methods: *Short-read bulk RNAseq: Quantification of reads*) and used to compare sense and antisense transcription over the elements. Comparison for forward-stranded elements (+) was done between forward transcription (transcription "in sense" to the element), and reverse transcription (transcription "in antisense" to the element), and the opposite for reverse-stranded elements.

Defining intact YY1 motif presence over L1s

To characterise some of the expressed L1s in our samples, the sequences were extracted using BEDtools getfasta from GRCh38.p13 (-fi). Using a custom script, the YY1 motif was defined to be to be "intact" if CAAGATGGCCG was found in the first 100 bps of the element.

Age estimation of TEs

The evolutionary age of a given TE was calculated from the percentage of divergence to its subfamily's consensus sequence (information retrieved from RepeatMasker). Evolutionary age was calculated assuming a natural mutation rate of 0.2% per million years[139,140].

L1s as alternative promoters

To identify cases where young full-length L1 elements act as alternative promoters, we narrowed the search to overlapping elements in antisense to the first exon of a gene's transcript. The transcription start site coordinate of all transcripts of Gencode hg38 GTF were extracted and formatted as a BED file. We intersected our young full-length L1 elements to the genes' TSS BED file using BEDtools intersect, forcing for opposite strands between the features in question (-S).

To validate some of these candidates using PacBio Isoseq data, the alignment files were converted to BED format using BEDtools bamtobed (--split) and intersected to the candidates (bedtools intersect -a [candidates] -b [reads] -u).

Computer systems and workflow management

All methods (except for statistical analyses and visualization steps) were performed on a SLURM-based computer cluster. All pre-processing pipelines were developed using Snakemake except for trusTEr. trusTEr uses a more rudimentary workflow management based on Python's concurrent. futures ThreadPoolExecutor.

Summary of results

The papers included in the thesis can be found in the appendix as papers I-IV. A summary of the overarching results can be found below.

Accurate TE subfamily quantification from single-cell RNA sequencing technologies using pseudo-bulk (papers I, II, III)

To better handle the limitations of single-cell technologies (as reviewed in the introduction), I developed a pipeline to quantify the expression of TE subfamilies in pseudo-bulk (per cell clusters or cell types; see methods: *trusTEr*). To the best of our knowledge, this was the first pipeline for quantifying TE expression in single-cell data, along with scTE, published five days later[141].

TrusTEr remains the only pipeline that performs TE quantification in pseudo-bulk, which is the main difference to scTE. SoloTE, a more recently published pipeline, quantifies TE expression on a locus level using uniquely mapped reads, which is likely to be disadvantageous for evolutionarily young TE subfamilies[3,142]. The expression values in this pseudo-bulk approach represent the mean expression of a group of cells; this results in less sparse count matrices than if quantifying individual cells. Using pseudo-bulks represents an advantage for lowly-expressed or evolutionarily young TE subfamilies that would otherwise be scarce in uniquely mapped reads, effectively increasing the statistical power of these TE subfamilies. Despite it not yet being proven for TE expression specifically, methods relying on pseudo-bulk for measuring differences in gene expression data have been shown to greatly increase the accuracy of the detected differences from single-cell data[72].

A fundamental difference between trusTEr and other single-cell TE pipelines is the underlying EM algorithm from TEtranscripts[71]. This method uses multi-mapping reads to calculate the relative abundances of each TE locus prior to their sum by TE subfamilies. Contrastingly, scTE uses a more rudimentary allocation of multimapping reads, where reads are mapped to each TE subfamily sequence[141]. It is yet to be determined however, how much the usage of multi-mapping reads on a TE subfamily sequence may inflate the expression of TE subfamilies that present a high degree of sequence similarity (e.g., L1HS and L1PA2)[3].

Paper I: Quantification of LTR subfamilies is sensitive to transcriptional differences between cell types

We first developed trusTEr to test for differential TE expression between cell types present in cortical samples from Emx1-Cre Trim28-floxed mice. These animals underwent Trim28 deletion in cortical progenitors (expressing Cre) starting at day 10 of embryonic development. The cortex of the adult Emx1-Cre Trim28-floxed mice lacked Trim28 in excitatory neurons, astrocytes, oligodendrocytes, and oligodendrocyte precursors, but remained intact in migrating cells present later in development, such as inhibitory neurons and microglia.

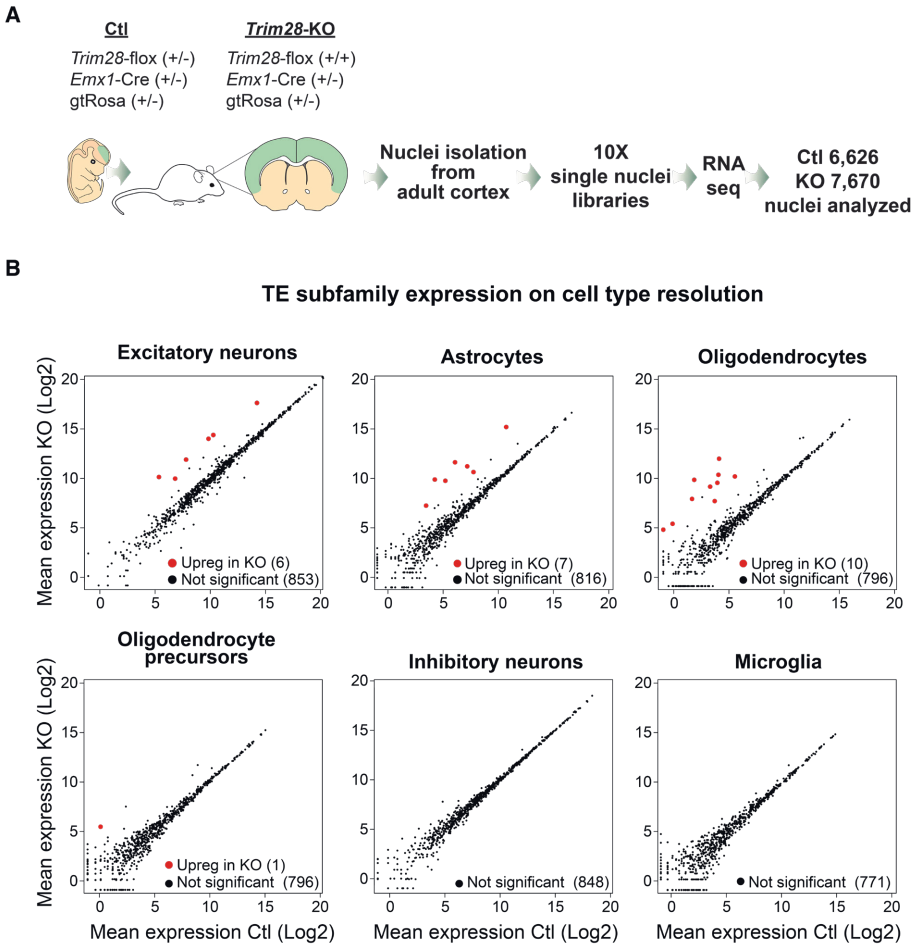


Figure 13 Cell type pseudo-bulk quantification of TE subfamilies upon TRIM28-KO

A) Breeding scheme for efficient conditional deletion of Trim28-KO during cortical development to analyse the adult tissue 3 months later by single-nuclei RNA-seq. B) Mean plots showing TE subfamilies upregulation upon TRIM28-KO using a pseudobulk approach per cell type.

trusTEr confirmed the upregulation of mouse-specific LTR elements (i.e., several IAP subfamilies and MMERVK10C) specifically in cell types that lacked Trim28 (Figure 13). This experiment confirmed that trusTEr's methodology accurately captured the differences in TE expression between the different cell types.

Paper II: Correct quantification of young L1 subfamilies is validated via epigenetic analyses

Using trusTEr, we analysed single-nuclei RNA sequencing data from postmortem adult human cortex, where we found a contrasting amount of young L1 expression between neurons and glia populations (Figure 14A). Given the technology's limitations and the prevalent co-transcription of L1s with genes, we performed CUT&RUN analysis of H3K4me3 (an active promoter histone mark) to validate the active status of evolutionarily young L1 promoters (Figure 14B). Our results show that trusTEr's methodology is sufficiently sensitive to capture the expression of even the youngest, most repetitive L1 subfamilies in the human genome.

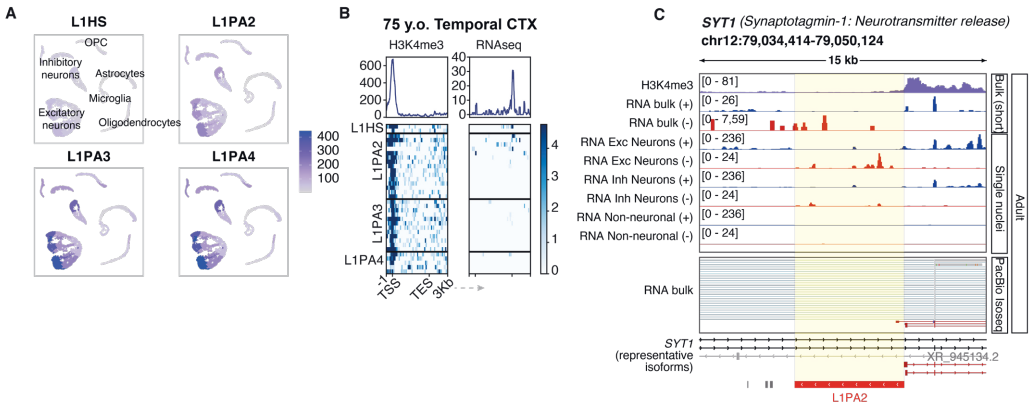


Figure 14 Expression of young L1 elements in the adult human brain

A) Pseudo-bulk cluster quantification of evolutionary young L1 subfamilies in adult cortical single nuclei RNA-seq. B) Heatmaps showing signal in confident H3K4me3 peaks called (left) over full-length evolutionary young L1s and their expression (right) in adult cortical neurons (NeuN positive nuclei). C) Genome browser showing an example of an adult-specific and neuron-specific L1-derived transcript creating a SYT1 isoform. Tracks from top to bottom: (1) H3K4me3 CUT&RUN in adult NeuN positive nuclei, bulk RNA-seq split by (2) forward (blue) and (3) reverse (red) transcription, pseudo-bulk of clusters in adult cortical samples single nuclei RNA-seq grouped by cell type and split by strand (4-9), PacBio Isoseq of adult cortical sample (10), genomic locus showing representative transcripts and L1s (11).

Paper III: Pseudo-bulk of gene expression validates interferon gamma response upon TBI

TBI tissue was obtained from emergency decompressive surgeries intended to lower the intracranial pressure of the patients. Given the type of injury and clinical setting, the TBI samples contained blood and a large number of damaged cells, which limited the quality of the sequencing data. Thus, cell types and levels of gene detection were heterogeneous among samples.

We used trusTEr's methodology to validate the main gene expression findings in our study, namely, an innate immune response in oligodendroglia characterised by a long list of upregulated genes, such as STAT1 and STAT2. Our results validate trusTEr's accuracy and sensitivity, which captures major transcriptional differences in gene expression data using a pseudo-bulk approach (Figure 15).

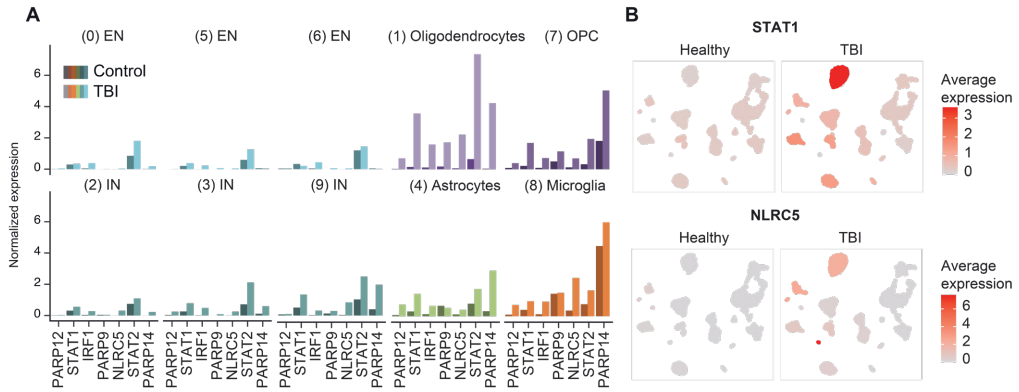


Figure 15 Pseudo-bulk expression of immune related genes.

A) Cluster pseudobulk expression of representative immune-related genes which were significantly upregulated when quantified per cell. B) UMAP projection coloured by normalised expression of STAT1 and NLRC5 and split by condition.

TRIM28-mediated silencing of TEs is required for the establishment of a stable silencing during early development (papers I, IV)

Increasing evidence suggests that the correct regulation of TE expression is essential for healthy mouse and human development[12,60,143]. Several studies have shown species-specific regulatory networks that depend on a timely TE repression via key regulators such as DNA methylation, TRIM28, and its targeting machinery, KRAB ZNFs[12,18,60,87,143,144]. Little is known, however, about the downstream effects of flawed TE repression during development.

Paper I: Correct epigenetic regulation of ERVs during mouse development is essential for stable silencing during adulthood

We investigated the function of TRIM28 in mouse brain development. The deletion of TRIM28 during early development – either modelled *in vitro* using a CRISPR-KO in mouse NPCs or *in vivo* from Emx1-Cre Trim28-floxed mice – resulted in a massive upregulation of ERVs. Contrastingly, upon the removal of TRIM28 in adult animals (CRISPR-KO), no ERV upregulation was detected. Emx1-Cre Trim28-floxed animals (which developed without TRIM28 in cortical progenitors) survived until adulthood and maintained the ERV transcriptional activation. Thus, the repression of ERVs depends on the presence of TRIM28 during development and its absence during development hinders ERV repression throughout life (Figure 16).

That the removal of TRIM28 in adulthood did not lead to an ERV upregulation points to the fact that TRIM28-mediated repression is not the sole regulator of these elements in adulthood – which is in line with previous observations in the field[143,145]. In mice, the recruitment of TRIM28 during the first days of development results in de novo hypermethylation of the locus[145]. However, the existence of this mechanism in human cells, and its relevance for normal development, is yet to be proven[143,144].

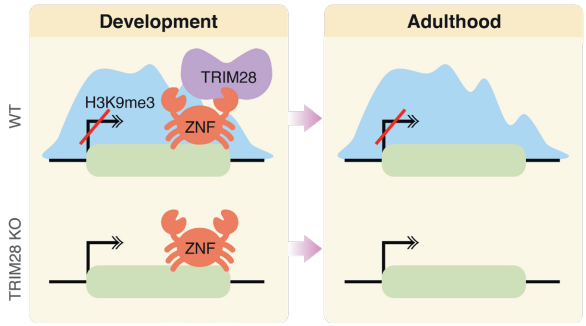


Figure 16 Schematic of the TRIM28-dependent establishment of H3K9me3 during development.

Paper IV: Stable epigenetic silencing of SVAs is accomplished by a dual layer of repression

We studied the role of ZNF91, a KRAB-ZNF that targets SVAs. We inhibited the expression of ZNF91, DNMT1, and a combination of the two (ZNF91+DNMT1) in human NPCs – where SVAs are transcriptionally repressed and mostly covered in DNA methylation (Figure 17A)[38]. Our results show that the inhibition of ZNF91 prevents the formation of H3K9me3 over ZNF91 targets, but that this H3K9me3-depletion is not sufficient for transcriptional activation (Figure 17C). Similarly, DNMT1-KO successfully removed DNA methylation over SVAs but did not result in the transcriptional activation of these elements (Figure 17C). The inhibition of ZNF91+DNMT1, however, resulted in a massive upregulation of SVAs (Figure 17B&C) – supporting the idea that TRIM28-mediated repression and DNA methylation simultaneously safeguard TE repression (Figure 17B).

To model an earlier developmental time point where DNA methylation has not been fully established, we used human induced pluripotent stem cells (iPSCs) to inhibit the expression of ZNF91 (Figure 17A). This experiment resulted in a massive upregulation of SVAs (Figure 17C).

Taken together, the observations drawn from these two projects strongly suggest that TRIM28-mediated silencing of TEs is required during early mouse and human development for the establishment of DNA methylation (Figure 17B).

Interestingly, there are exceptions to this model where the transcription of some SVAs is dependent on only one layer of repression (i.e., either TRIM28 or DNA methylation). Other exceptions include SVAs whose transcriptional repression is not dependent on a specific layer of repression; in other words, the removal of either one of these repressive layers is sufficient to cause their transcriptional activation. More studies will be required to answer the specific ways SVA escapes from these repressors.

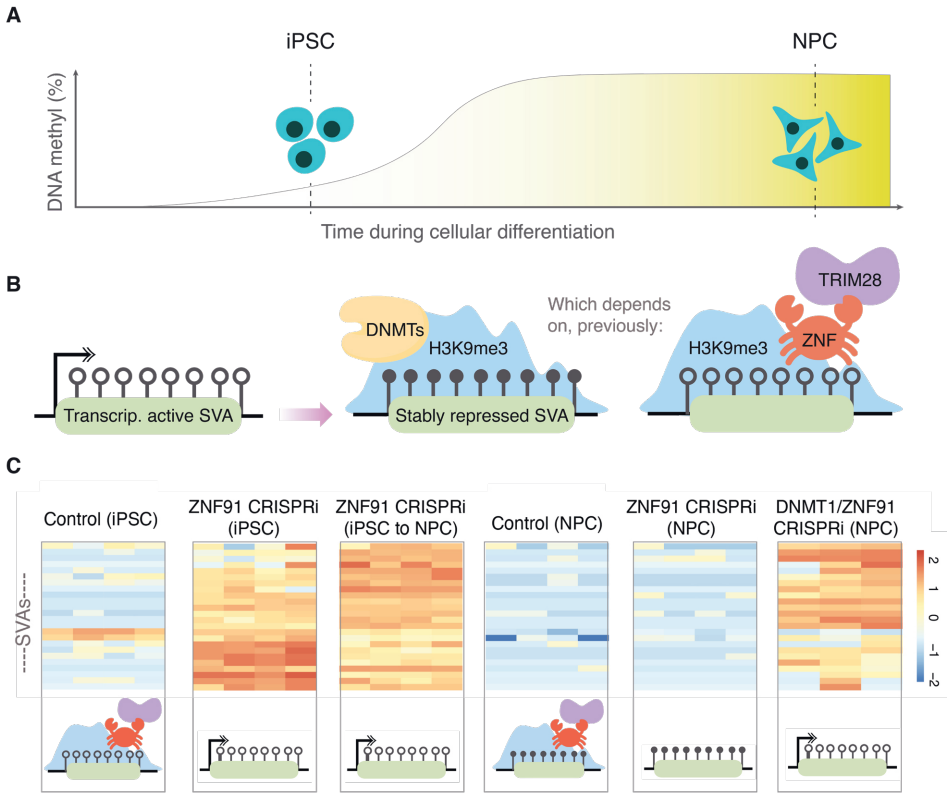


Figure 17 Double layer of repression achieved via H3K9me3 and DNA methylation.

A) Schematic of DNA methylation establishment in iPSC and NPC models. B) Schematic of H3K9me3 and DNA methylation establishment via a KRAB ZNF, TRIM28 and DNMTs. C) Top: heatmap showing expression of a subset of SVAs in the different datasets and bottom: schematic of their hypothetical epigenetic status.

TEs contribute to the human transcriptome complexity (papers II, IV)

Paper II: L1 retrotransposons drive human neuronal transcriptome complexity and functional diversification

Undoubtedly, TEs are an invaluable source of evolutionary innovation[2,64,78,140] whose activity may result in species-specific physiological effects[12,18]. Thus, genetic information rooted in human-specific TEs is likely to answer questions regarding human evolution – such as the size and complexity of our brains.

The regulatory potential of L1s might drive changes in the cells' transcriptome and even play a role in the fate and state of a cell[16,21,41,58,78]. However, studies regarding the contribution of L1s to human brain complexity have largely focused on L1 de novo insertions and somatic mosaicism in neurons, which has yet to be proven to have functional relevance. The role of L1 transcription in the brain, however, has not been thoroughly analysed[16,41,39].

This study is a multi-omic effort to characterise the expression of evolutionary young L1s in the human brain. Using our single-nuclei RNA-seq data, we investigated differences in L1 expression between the different cell types and cell states (cycling/non-cycling) in our datasets using trusTEr. Our results show that during adulthood, neurons express significantly higher levels of evolutionary young L1 elements when compared to glial cells (Figure 14A). Contrastingly, L1 expression is kept at similar levels throughout the different cell types in the fetal forebrain.

We further characterised the expressed L1 in these samples using different RNA sequencing approaches, as well as CUT&RUN experiments for H3K4me3 and H3K9me3. Our results show substantial contribution of these elements to the human brain transcriptome (Figure 18A). We prove that young L1s have active promoters and initiate transcription of many alternative gene isoforms – some of which are human-specific (Figure 18B). More so, we found that L1 transcripts are dynamically regulated throughout life, pointing towards a functional role of L1-derived transcripts in the human brain (Figure 18A).

To exemplify their potential function, we conducted detailed studies focusing on an L1-lncRNA (LINC01876) – a lncRNA originating from an antisense L1PA2 promoter (Figure 18B). The expression of this L1-lncRNA is human-specific and only present during brain development. Using cell culture models, we show that upregulated genes upon L1-lncRNA CRISPRi in human NPCs are also more expressed in chimpanzee NPCs when compared to human NPCs. Furthermore, inhibition of this L1-lncRNA at the iPSC state and differentiation towards cerebral organoids resulted in smaller organoids (Figure 18C-D). Transcriptional profiling of this experiment shows that the inhibition of this L1-lncRNA leads to premature differentiation of NPCs into neurons (Figure 18E-F).

Paper IV: Contribution of SVA retrotransposons to the human transcriptome

Many studies have directly implicated TEs in human brain disorders. Some brain disorders associated with TE activity have a documented disease-causing polymorphic insertion[15,146]. The SVA family includes some of the youngest retrotransposon subfamilies in our genome, encompassing many human-specific and polymorphic elements[51]. Paper IV focuses on the XDP-SVA, a disease-causing polymorphic insertion; however, our results also shed light on the contributions of SVAs as regulatory elements in the human transcriptome.

In this study, we performed a detailed transcriptional and epigenetic study of SVAs in human NPCs and iPSCs, including information about their expression, histone modifications, and DNA methylation status.

Our results show that SVAs, including polymorphic insertions like the XDP-SVA, are repressed in a ZNF91- (and consequently in a TRIM28-) dependent manner (Figure 19). As previously described, we show that DNA methylation is also required to repress some SVAs in the genome (Figure 19; see results: *Paper IV: Stable epigenetic silencing of SVAs is accomplished by a dual layer of repression*) – suggesting a dynamic modulation of transcription during early development .

Importantly, we provide evidence that SVAs have strong effects on nearby gene expression, generally as enhancers. Furthermore, our results show that fixed and polymorphic SVA insertions also create new gene isoforms via the creation of alternative splicing sites, 3' ends, and promoter sites.

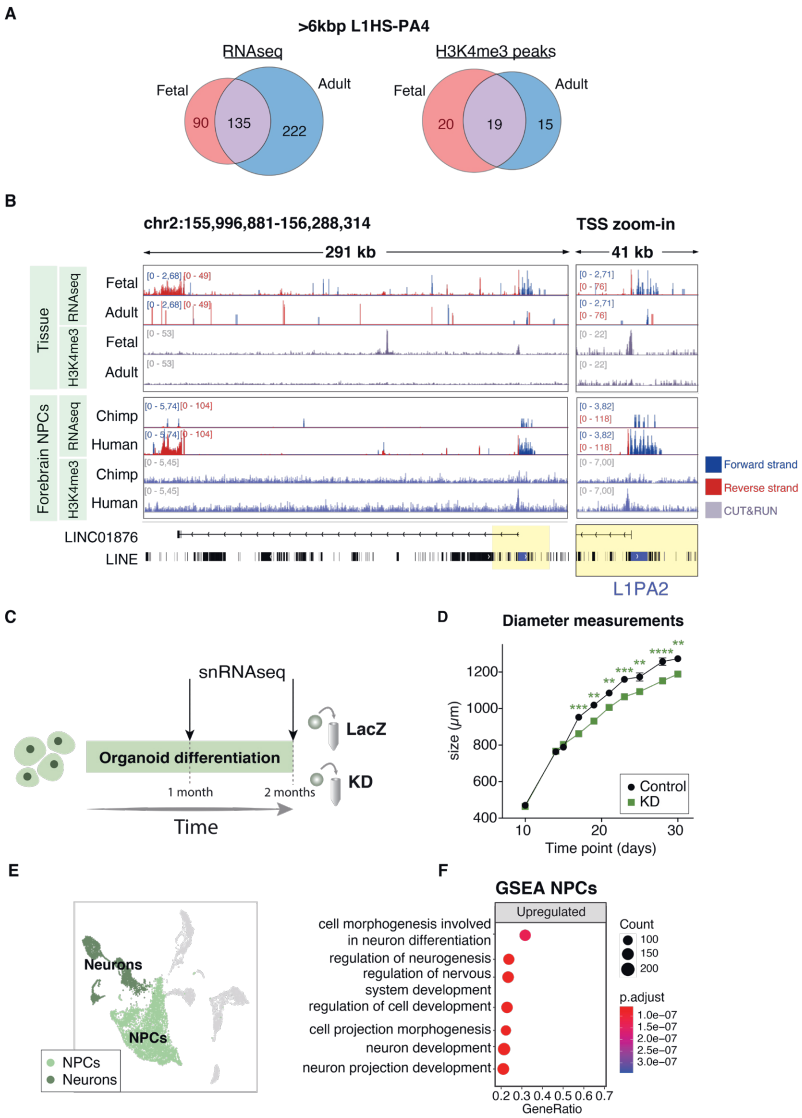


Figure 18 L1s are dynamically regulated and have functional roles in development of the human brain

A) Venn diagrams showing expressed (left) young full-length L1 elements and epigenetic status (H3K4me3, right) in fetal forebrain (pink) and adult cortex (blue). B) Genome browser tracks over the L1-lncRNA LINC01876. Tracks show RNA-seq signal split in forward (blue) and reverse (red) strand, and H3K4me3 CUT&RUN (purple) of fetal forebrain, adult cortex, chimpanzee forebrain neural progenitor cells (fbNPCs), and human fbNPCs. E) Schematic of organoid differentiation. F) Organoids diameter measurement (n=20-30 organoids per time point) (Mixed-effects analysis and Sidak correction for multiple comparisons). G) UMAP projection coloured highlighting NPC and neuronal clusters. H) Gene set enrichment analysis in NPC (L1-lncRNA CRISPRi vs LacZ).

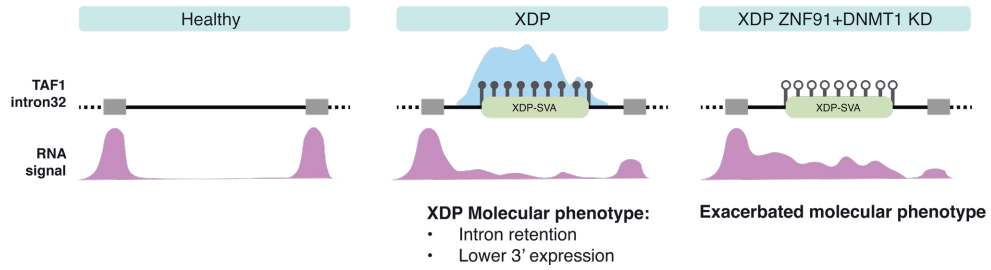


Figure 19 Summary schematic of the results of paper IV in relation to TAF1 expression.

Our results show that ZNF91 protects genes from the regulatory potential of SVAs. This is exemplified by TAF1 in XDP, where the molecular phenotype of the disease gets exacerbated upon the removal of the XDP-SVA-associated heterochromatin marks (Figure 19).

Aberrant expression of ERVs can lead to an immune response (papers I, III)

It has been previously described that aberrant TE expression can trigger an innate immune response via Cas-STING pathway recognition[99]. Interestingly, there is an increasing number of neurological diseases associated with an aberrant ERV expression, many of which have an immune component[25,30,34,63].

Paper I: ERV expression during mouse brain development correlates with a neuroinflammatory-like response that persists into adulthood

Studies on the activation of ERVs and their role in disease have been limited and technically challenging. In this study, we remove a key ERV regulator, TRIM28. As a result, we were able to investigate the effect of an aberrant transcription of ERVs without introducing a xeno-overexpression.

Results from our Emx1-Cre Trim28-floxed animals (which developed without TRIM28 in cortical progenitors) show that the transcriptional de-repression of ERVs during development correlates with a neuroinflammatory-like response which persists into adulthood. This response features protein aggregates of IAPs (mouse-specific ERV family) (Figure 20A), dysregulation of genes previously implicated in several neurodevelopmental and neurodegenerative disorders, and microglia changes. Importantly, TRIM28 was not knocked-out in microglia cells, suggesting that these changes are a reaction to the aberrant expression of ERVs in their neighbouring cells. Changes in microglia included an increased presence of CD68 and IBA1 – both present in low-levels in "resting" microglia and in high levels in "activated" microglia (Figure 20B-D).

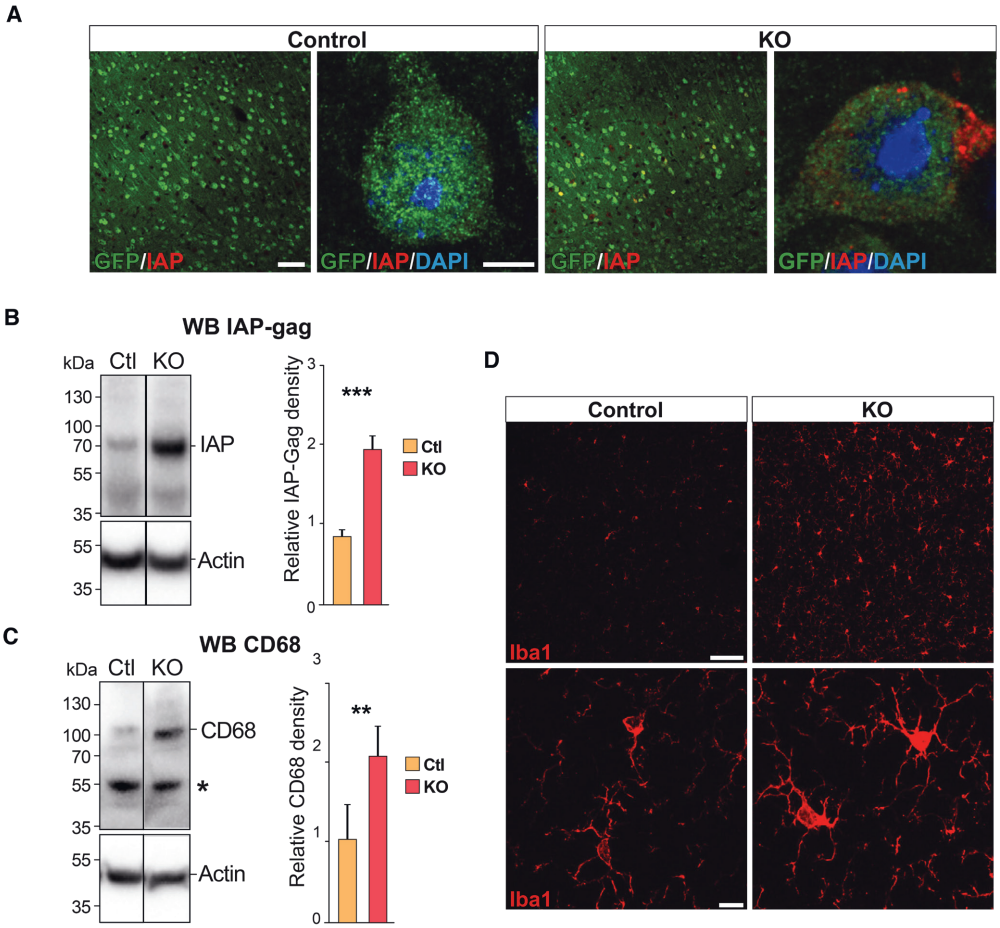


Figure 20 Inflammatory-like response upon TRIM28-KO.

A) Immunohistochemistry showing presence of IAP-gag in an aggregate-like in TRIM28-KO samples. B) Western blot (WB) showing presence of IAP-gag protein in TRIM28-KO samples C) WB showing increased presence of CD68 in TRIM28-KO samples. D) IHC showing increased presence of Iba1 in TRIM28-KO samples.

Paper III: ERV expression correlates with the beginning of human neuroinflammation

Neuroinflammation is a common and well-established symptom of several brain diseases, including TBI[122]. Yet, the molecular cascade in the different cell types of the brain at the beginning of a neuroinflammatory response remain poorly understood. Previous studies have relied on post-mortem material of neuroinflammatory-related pathologies, such as neurodegenerative disorders[147]. These samples represent a late stage of the disease, which limits their relevance in the understanding of the beginning of neuroinflammation.

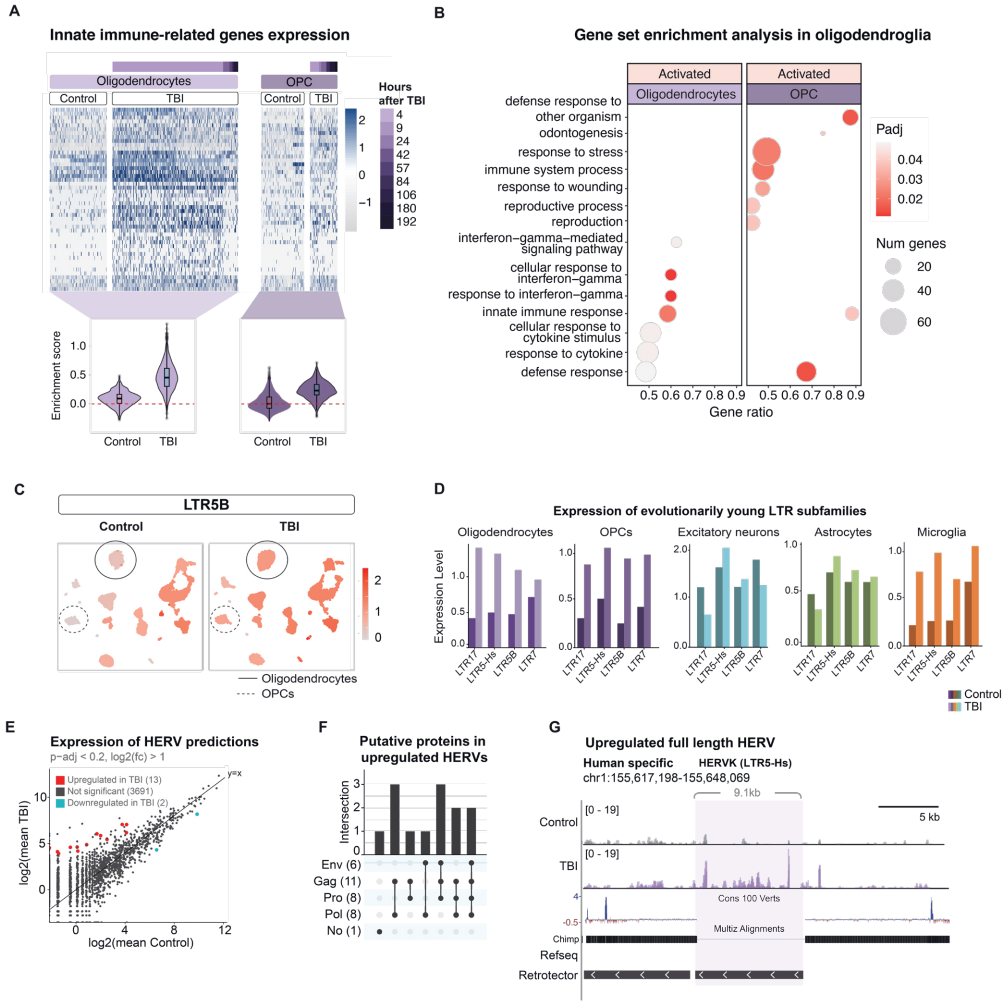


Figure 21 Innate immune response and upregulation of evolutionary young HERV elements in TBI oligodendroglia

A) Top: Expression of upregulated genes in TBI oligodendroglia related to innate immunity (TBI vs ctrl; padj < 0.01; log2FC > 0.05; n = 12 TBI samples, n = 5 ctrl samples). Bottom: Enrichment scores of genes shown in heatmap (AddModuleScore, Seurat) B) Overrepresentation test of differentially expressed genes in TBI vs ctrl oligodendroglia (padj < 0.01, Wilcoxon Rank Sum test (FindMarkers, Seurat); n = 12 TBI samples, n = 5 ctrl samples. Biological-process ontology). C) UMAP projection coloured by cluster pseudobulk expression of LTR5B split by sample condition. D) Cluster pseudobulk expression levels of evolutionary young HERV LTR families. E) Differential expression analysis of hg38 HERV predictions in bulk RNA-seq of TBI and control samples. F) Putative proteins identified in upregulated HERVs in TBI samples. G) Genome browser tracks showing increased expression of a human-specific HERV-K element. Tracks show RNA-seq signal of control and TBI samples, conservation scores of the region to 100 vertebrates, conservation track to chimpanzee, Refseq genes, and retrotector predictions.

In this study, we built a cohort of TBI patients with samples taken within hours after the injury. Using single-nuclei RNA sequencing, we were able to characterise the neuroinflammatory state in these samples, including the appearance of cycling microglia and a cell-type specific molecular cascade. We found genes related to synaptic terms to be dysregulated in excitatory neurons, and innate immune and defence response terms activated in oligodendroglia, with evidence for an IFNγ response specific to oligodendrocytes (Figure 21A&B). Furthermore, TBI oligodendroglia showed an upregulation of genes related to MHC class I and II (regulators from both classes, and MHC class II molecules). Taken together, these results strongly suggest oligodendroglia to undergo an immune-like shift, playing an unprecedented role in the beginning of neuroinflammation.

We assessed TE expression using trusTEr. Our results show an increased expression of the LTR subfamilies associated with evolutionary young HERV elements (LTR5/HERVK, LTR7/HERVH and LTR17/HERVW) in TBI oligodendroglia and microglia (Figure 21C-D). This expression was verified using bulk RNA sequencing using a unique mapping approach, where we narrowed down this upregulation to 13 HERV elements in the genome (Figure 21E) – most of which have protein coding potential (12 out of 13) (Figure 21F&G).

Paper III: IFNγ response is sufficient to activate ERV expression in human glia progenitor cells

To further characterise the relationship between the innate immune response observed in oligodendroglia and the ERV activation, we set up an in vitro model of immature human glia progenitor cells (hGPCs). The cells were treated for 48 hours with IFNγ and sequenced in bulk RNA sequencing.

We verified the IFNγ response in the treated hGPCs and observed a clear overlap between upregulated genes in IFNγ- versus non- treated hGPCs, and upregulated genes in TBI versus control samples (Figure 22A&B). Furthermore, we observed a transcriptional upregulation of several ERV elements in the IFNγ- vs non- treated cells, some of which were also upregulated in the TBI versus control samples (Figure 22C&D).

This experiment proves that the trigger of an IFNγ response is solely sufficient for the transcriptional activation of ERVs in hGPCs. Mechanistic studies activating HERVs in glia progenitor cells would be required to establish the directionality of the transcriptional activation of HERVs and IFNγ-related genes.

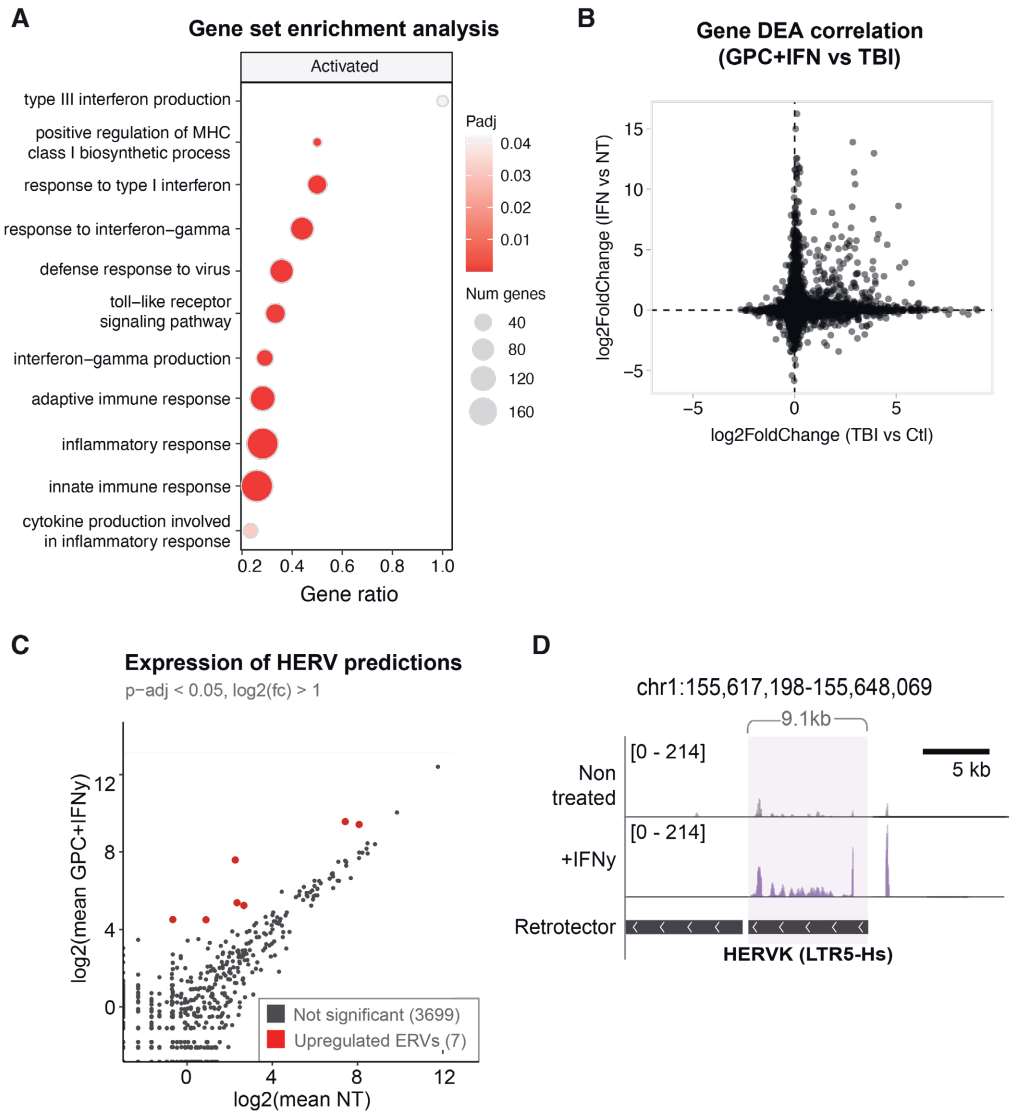


Figure 22 Transcriptional response of human glia progenitor cells (hGPC) upon IFNγ treatment.
A) Gene set enrichment analysis of hGPC post IFNγ- vs non- treated hGPCs. B) Correlation of gene expression effects (log2FoldChange) between TBI vs control samples (x-axis) and hGPC IFNγ- vs non- treated cells. C) Differential expression analysis of hg38 HERV predictions in bulk RNA-seq of IFNγ- and non- treated cells. D) Example (same as Figure 21E) showing upregulation of a human-specific HERV-K element in INFγ-treated hGPCs.

Conclusions and future perspectives

Research performed over the past decades has proven that TEs are not "junk DNA", but rather agents in genome evolution with functional impact in many organisms – including humans.

The work in this thesis shows that TEs have shaped and keep shaping our genome, a process we found to impact our brain transcriptome. It highlights the importance of human-specific TEs as a research avenue to explain human-specific traits and disease. Furthermore, our results show that the dynamic epigenetic status of TEs in the human brain likely modulates transcriptional networks during brain development, throughout life, and in disease contexts.

From the work conducted in these studies, I conclude that TEs are highly integrated agents in the transcriptional regulation of the human brain, as well as drivers of transcriptional innovation. I suspect that the reputation of TEs to mainly act as disease drivers merely reflects our understanding rather than the true net worth TEs have in the human genome. This is likely to change in the near future, as recent technological advances enable a better characterisation of TE-derived transcripts (e.g., through long-read sequencing technologies) and their interaction with their surrounding genome and transcripts (e.g., via integrated transcriptional and epigenetic analyses of the same cell). These technologies will solve some of the main bioinformatic limitations in the field, such as integrating data types and addressing mappability issues – ideally moving away from having a reference genome and towards building sample-specific genomes and transcriptomes from whole genome or transcriptome long-read sequencing data.

Several questions are left open to explore in relation to our research. Although several studies have shown a relationship between HERVs and an inflammatory state, we still need to answer if HERVs are initiating or boosting the observed response in the brain. This will require mechanistic studies which transcriptionally activate HERVs without an added stressor, overexpression of these elements, or depletion of their transcriptional regulators (which would likely come with side effects that are hard to distinguish from those directly related to the transcriptional activation of HERVs). Similarly, open questions regarding the function of the observed expression of evolutionary young L1s during brain development will need to be addressed mechanistically.

As for SVA retrotransposition events in early development, it will be crucial to investigate how SVAs are still creating polymorphic insertions if ZNF91 targets most (if not all) SVAs in the genome? When is ZNF91-dependent H3K9me3 not present to transcriptionally repress SVAs? Or is ZNF91 enabling SVA retrotransposition by controlling the gene regulatory potential of SVAs and thus mitigating its consequences? The answer to this last question might confirm or completely transform our understanding of the relationship between KRAB-ZNFs and TEs.

From our results, it is clear to me that there are multiple epigenetic switches that enable, for example, the change of epigenetic status over L1s between development and adulthood, or between SVAs

in iPSC and NPCs. How these epigenetic switches are modulated and timed during developmental processes and adult life, including over polymorphic and somatic insertions, remains an important open question.

The TE field is at a crucial point in time, where it has recently caught the attention of a wider (and growing) scientific research community. This interest has inspired many methodological and technological advances for the study of TEs, which I foresee exponentially widening our understanding of the role of TEs in human brain evolution, inter-individual variation, and disease.

References

1. Jönsson, M.E., Garza, R., Johansson, P.A. & Jakobsson, J. Transposable Elements: A Common Feature of Neurodevelopmental and Neurodegenerative Disorders. Trends in Genetics 36, 610-623 (2020).
2. Hoyt, S.J. et al. From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. Science 376, eabk3112 (2022).
3. Lanciano, S. & Cristofari, G. Measuring and interpreting transposable element expression. Nat Rev Genet 21, 721-736 (2020).
4. Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M.M. Umap and Bismap: quantifying genome and methylome mappability. Nucleic Acids Res 46, e120 (2018).
5. Wells, J.N. & Feschotte, C. A Field Guide to Eukaryotic Transposable Elements. Annu Rev Genet 54, 539-561 (2020).
6. Bourque, G. et al. Ten things you should know about transposable elements. Genome Biology 19(2018).
7. Bennett, E.A., Coleman, L.E., Tsui, C., Pittard, W.S. & Devine, S.E. Natural Genetic Variation Caused by Transposable Elements in Humans. Genetics 168, 933-951 (2004).
8. Bannert, N. & Kurth, R. The evolutionary dynamics of human endogenous retroviral families. Annu Rev Genomics Hum Genet 7, 149-73 (2006).
9. Chuong, E.B., Elde, N.C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet 18, 71-86 (2017).
10. Xie, M. et al. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. Nat Genet 45, 836-41 (2013).
11. Trizzino, M., Kapusta, A. & Brown, C.D. Transposable elements generate regulatory novelty in a tissue-specific fashion. BMC Genomics 19, 468 (2018).
12. Pontis, J. et al. Primate-specific transposable elements shape transcriptional networks during human development. Nat Commun 13, 7178 (2022).
13. Pontis, J. et al. Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. Cell Stem Cell 24, 724-735 e5 (2019).
14. Playfoot, C.J. et al. Transposable elements and their KZFP controllers are drivers of transcriptional innovation in the developing human brain. Genome Research 31, 1531-1545 (2021).
15. Hancks, D.C. & Kazazian, H.H., Jr. Roles for retrotransposon insertions in human disease. Mob DNA 7, 9 (2016).
16. Coufal, N.G. et al. L1 retrotransposition in human neural progenitor cells. Nature 460, 1127-1131 (2009).
17. Philippe, C. et al. Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. Elife 5(2016).
18. Prescott, S.L. et al. Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. Cell 163, 68-83 (2015).
19. Batzer, M.A. & Deininger, P.L. Alu repeats and human genomic diversity. Nat Rev Genet 3, 370-9 (2002).

20. Han, K. et al. L1 recombination-associated deletions generate human genomic variation. Proc Natl Acad Sci U S A 105, 19366-71 (2008).
21. Erwin, J.A. et al. L1-associated genomic regions are deleted in somatic cells of the healthy human brain. Nature Neuroscience 19, 1583-1591 (2016).
22. Frost, J.M. et al. Regulation of human trophoblast gene expression by endogenous retroviruses. Nat Struct Mol Biol 30, 527-538 (2023).
23. Theunissen, T.W. et al. Molecular Criteria for Defining the Naive Human Pluripotent State. Cell Stem Cell 19, 502-515 (2016).
24. Lu, X. et al. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. Nat Struct Mol Biol 21, 423-5 (2014).
25. Chuong, E.B., Elde, N.C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science 351, 1083-7 (2016).
26. Konkel, M.K., Walker, J.A. & Batzer, M.A. LINEs and SINEs of primate evolution. Evol Anthropol 19, 236-249 (2010).
27. Lui, J.H., Hansen, D.V. & Kriegstein, A.R. Development and evolution of the human neocortex. Cell 146, 18-36 (2011).
28. Kronenberg, Z.N. et al. High-resolution comparative analysis of great ape genomes. Science 360(2018).
29. King, M.C. & Wilson, A.C. Evolution at two levels in humans and chimpanzees. Science 188, 107-16 (1975).
30. Christensen, T. et al. Molecular characterization of HERV-H variants associated with multiple sclerosis. Acta Neurol Scand 101, 229-38 (2000).
31. Christensen, T. Association of human endogenous retroviruses with multiple sclerosis and possible interactions with herpes viruses. Rev Med Virol 15, 179-211 (2005).
32. Laska, M.J. et al. Expression of HERV-Fc1, a human endogenous retrovirus, is increased in patients with active multiple sclerosis. J Virol 86, 3713-22 (2012).
33. Tam, O.H. et al. Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia. Cell Rep 29, 1164-1177 e5 (2019).
34. Li, W. et al. Human endogenous retrovirus-K contributes to motor neuron disease. Sci Transl Med 7, 307ra153 (2015).
35. Pozojevic, J. et al. Transcriptional Alterations in X-Linked Dystonia-Parkinsonism Caused by the SVA Retrotransposon. Int J Mol Sci 23(2022).
36. Muotri, A.R. et al. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 435, 903-910 (2005).
37. Upton, R., Kyle et al. Ubiquitous L1 Mosaicism in Hippocampal Neurons. Cell 161, 228-239 (2015).
38. Ewing, A.D. et al. Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling. Molecular Cell 80, 915-928.e5 (2020).
39. Evrony, D., Gilad et al. Single-Neuron Sequencing Analysis of L1 Retrotransposition and Somatic Mutation in the Human Brain. Cell 151, 483-496 (2012).
40. Baillie, J.K. et al. Somatic retrotransposition alters the genetic landscape of the human brain. Nature 479, 534-7 (2011).
41. Evrony, G.D., Lee, E., Park, P.J. & Walsh, C.A. Resolving rates of mutation in the brain using single-neuron genomics. Elife 5(2016).

42. Athanikar, J.N. A YY1-binding site is required for accurate human LINE-1 transcription initiation. Nucleic Acids Research 32, 3846-3855 (2004).
43. Denli, A.M. et al. Primate-specific ORF0 contributes to retrotransposon-mediated diversity. Cell 163, 583-93 (2015).
44. Kazazian, H.H. & Moran, J.V. The impact of L1 retrotransposons on the human genome. Nature Genetics 19, 19-24 (1998).
45. Belancio, V.P., Whelton, M. & Deininger, P. Requirements for polyadenylation at the 3' end of LINE-1 elements. Gene 390, 98-107 (2007).
46. Viollet, S., Monot, C. & Cristofari, G. L1 retrotransposition: The snap-velcro model and its consequences. Mob Genet Elements 4, e28907 (2014).
47. Brouha, B. et al. Hot L1s account for the bulk of retrotransposition in the human population. Proceedings of the National Academy of Sciences 100, 5280-5285 (2003).
48. Beck, C.R. et al. LINE-1 retrotransposition activity in human genomes. Cell 141, 1159-70 (2010).
49. Feusier, J. et al. Pedigree-based estimation of human mobile element retrotransposition rates. Genome Res 29, 1567-1577 (2019).
50. Sanchez-Luque, F.J. et al. LINE-1 Evasion of Epigenetic Repression in Humans. Mol Cell 75, 590-604 e12 (2019).
51. Hancks, D.C. & Kazazian, H.H., Jr. SVA retrotransposons: Evolution and genetic instability. Semin Cancer Biol 20, 234-45 (2010).
52. Burns, K.H. Transposable elements in cancer. Nat Rev Cancer 17, 415-424 (2017).
53. Muotri, A.R. et al. L1 retrotransposition in neurons is modulated by MeCP2. Nature 468, 443-446 (2010).
54. Zhao, B. et al. Somatic LINE-1 retrotransposition in cortical neurons and non-brain tissues of Rett patients and healthy individuals. PLoS Genet 15, e1008043 (2019).
55. Coufal, N.G. et al. Ataxia telangiectasia mutated (ATM) modulates long interspersed element-1 (L1) retrotransposition in human neural stem cells. Proceedings of the National Academy of Sciences 108, 20382-20387 (2011).
56. Bundo, M. et al. Increased L1 Retrotransposition in the Neuronal Genome in Schizophrenia. Neuron 81, 306-313 (2014).
57. Liu, S. et al. Inverse changes in L1 retrotransposons between blood and brain in major depressive disorder. Scientific Reports 6, 37530 (2016).
58. Jacobs, F.M.J. et al. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature 516, 242-245 (2014).
59. Jern, P. & Coffin, J.M. Effects of retroviruses on host genome function. Annu Rev Genet 42, 709-32 (2008).
60. Fuentes, D.R., Swigut, T. & Wysocka, J. Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. eLife 7(2018).
61. Gonzalez-Cao, M. et al. Human endogenous retroviruses and cancer. Cancer Biol Med 13, 483-488 (2016).
62. Johnston, J.B. et al. Monocyte activation and differentiation augment human endogenous retrovirus expression: implications for inflammatory brain diseases. Ann Neurol 50, 434-42 (2001).
63. Posso-Osorio, I., Tobon, G.J. & Canas, C.A. Human endogenous retroviruses (HERV) and non-HERV viruses incorporated into the human genome and their role in the development of autoimmune diseases. J Transl Autoimmun 4, 100137 (2021).

64. Fort, V., Khelifi, G. & Hussein, S.M.I. Long non-coding RNAs and transposable elements: A functional relationship. Biochim Biophys Acta Mol Cell Res 1868, 118837 (2021).
65. Groza, C., Chen, X., Wheeler, T.J., Bourque, G. & Goubert, C. GraffiTE: a Unified Framework to Analyze Transposable Element Insertion Polymorphisms using Genome-graphs. (Cold Spring Harbor Laboratory, 2023).
66. Braun, E. et al. Comprehensive cell atlas of the first-trimester developing human brain. (Cold Spring Harbor Laboratory, 2022).
67. La Manno, G. et al. Molecular architecture of the developing mouse brain. Nature 596, 92-96 (2021).
68. Zheng, G.X.Y. et al. Massively parallel digital transcriptional profiling of single cells. Nature Communications 8, 14049 (2017).
69. Wang, X., He, Y., Zhang, Q., Ren, X. & Zhang, Z. Direct Comparative Analyses of 10X Genomics Chromium and Smart-seq2. Genomics Proteomics Bioinformatics 19, 253-266 (2021).
70. La Manno, G. et al. RNA velocity of single cells. Nature 560, 494-498 (2018).
71. Jin, Y., Tam, O.H., Paniagua, E. & Hammell, M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics 31, 3593-3599 (2015).
72. Squair, J.W. et al. Confronting false discoveries in single-cell differential expression. Nat Commun 12, 5692 (2021).
73. Becker, J.S., Nicetto, D. & Zaret, K.S. H3K9me3-Dependent Heterochromatin: Barrier to Cell Fate Changes. Trends Genet 32, 29-41 (2016).
74. Schultz, D.C., Friedman, J.R. & Rauscher, F.J., 3rd. Targeting histone deacetylase complexes via KRAB-zinc finger proteins: the PHD and bromodomains of KAP-1 form a cooperative unit that recruits a novel isoform of the Mi-2alpha subunit of NuRD. Genes Dev 15, 428-43 (2001).
75. Schultz, D.C., Ayyanathan, K., Negorev, D., Maul, G.G. & Rauscher, F.J., 3rd. SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. Genes Dev 16, 919-32 (2002).
76. Tchasovnikarova, I.A. et al. GENE SILENCING. Epigenetic silencing by the HUSH complex mediates position-effect variegation in human cells. Science 348, 1481-1485 (2015).
77. de Tribolet-Hardy, J. et al. Genetic features and genomic targets of human KRAB-zinc finger proteins. Genome Res 33, 1409-1423 (2023).
78. Imbeault, M., Helleboid, P.Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. Nature 543, 550-554 (2017).
79. Johansson, P.A. et al. A cis-acting structural variation at the ZNF558 locus controls a gene regulatory network in human brain development. Cell Stem Cell 29, 52-69 e8 (2022).
80. Matsui, T. et al. Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. Nature 464, 927-31 (2010).
81. Robbez-Masson, L. et al. The HUSH complex cooperates with TRIM28 to repress young retrotransposons and new genes. Genome Res 28, 836-845 (2018).
82. Pandiloski, N. et al. DNA methylation governs the sensitivity of repeats to restriction by the HUSH-MORC2 corepressor. (Cold Spring Harbor Laboratory, 2023).
83. Liu, N. et al. Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators. Nature 553, 228-232 (2018).
84. Moore, L.D., Le, T. & Fan, G. DNA methylation and its basic function. Neuropsychopharmacology 38, 23-38 (2013).

85. Yoder, J.A., Walsh, C.P. & Bestor, T.H. Cytosine methylation and the ecology of intragenomic parasites. Trends Genet 13, 335-40 (1997).

86. Liao, J. et al. Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. Nat Genet 47, 469-78 (2015).

87. Jönsson, M.E. et al. Activation of neuronal genes via LINE-1 elements upon global DNA demethylation in human neural progenitors. Nature Communications 10(2019).

88. Horvath, S. DNA methylation age of human tissues and cell types. Genome Biol 14, R115 (2013).

89. Koito, A. & Ikeda, T. Intrinsic immunity against retrotransposons by APOBEC cytidine deaminases. Front Microbiol 4, 28 (2013).

90. Ikeda, T. et al. Intrinsic restriction activity by apolipoprotein B mRNA editing enzyme APOBEC1 against the mobility of autonomous retrotransposons. Nucleic Acids Res 39, 5538-54 (2011).

91. Kinomoto, M. et al. All APOBEC3 family proteins differentially inhibit LINE-1 retrotransposition. Nucleic Acids Res 35, 2955-64 (2007).

92. Marchetto, M.C.N. et al. Differential L1 regulation in pluripotent stem cells of humans and apes. Nature 503, 525-529 (2013).

93. Alberts, B., Heald, R., Johnson, A., Morgan, D., Raff, M., Roberts, K., Walter, P. Molecular Biology of the Cell, 1404 (W. W. Norton & Company, 2022).

94. Siomi, M.C., Sato, K., Pezic, D. & Aravin, A.A. PIWI-interacting small RNAs: the vanguard of genome defence. Nat Rev Mol Cell Biol 12, 246-58 (2011).

95. Soifer, H.S., Zaragoza, A., Peyvan, M., Behlke, M.A. & Rossi, J.J. A potential role for RNA interference in controlling the activity of the human LINE-1 retrotransposon. Nucleic Acids Res 33, 846-56 (2005).

96. Johnson, W.E. & Coffin, J.M. Constructing primate phylogenies from ancient retrovirus sequences. Proc Natl Acad Sci U S A 96, 10254-60 (1999).

97. Belancio, V.P., Roy-Engel, A.M. & Deininger, P.L. All y'all need to know 'bout retroelements in cancer. Seminars in Cancer Biology 20, 200-210 (2010).

98. Bragg, D.C. et al. Disease onset in X-linked dystonia-parkinsonism correlates with expansion of a hexameric repeat within an SVA retrotransposon in TAF1. Proc Natl Acad Sci U S A 114, E11020-E11028 (2017).

99. Kassiotis, G. & Stoye, J.P. Immune responses to endogenous retroelements: taking the bad with the good. Nature Reviews Immunology 16, 207-219 (2016).

100. Arasaratnam, C.J., Singh-Bains, M.K., Waldvogel, H.J. & Faull, R.L.M. Neuroimaging and neuropathology studies of X-linked dystonia parkinsonism. Neurobiol Dis 148, 105186 (2021).

101. Hanssen, H. et al. Basal ganglia and cerebellar pathology in X-linked dystonia-parkinsonism. Brain 141, 2995-3008 (2018).

102. Reyes, C.J. et al. Brain Regional Differences in Hexanucleotide Repeat Length in X-Linked Dystonia-Parkinsonism Using Nanopore Sequencing. Neurol Genet 7, e608 (2021).

103. Westenberger, A. et al. A hexanucleotide repeat modifies expressivity of X-linked dystonia parkinsonism. Ann Neurol 85, 812-822 (2019).

104. Aneichyk, T. et al. Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. Cell 172, 897-909 e21 (2018).

105. Roulois, D. et al. DNA-Demethylating Agents Target Colorectal Cancer Cells by Inducing Viral Mimicry by Endogenous Transcripts. Cell 162, 961-73 (2015).

106. Chiappinelli, K.B. et al. Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. Cell 169, 361 (2017).
107. Seth, R.B., Sun, L., Ea, C.K. & Chen, Z.J. Identification and characterization of MAVS, a mitochondrial antiviral signaling protein that activates NF-kappaB and IRF 3. Cell 122, 669-82 (2005).
108. Yoneyama, M. et al. The RNA helicase RIG-I has an essential function in double-stranded RNA-induced innate antiviral responses. Nat Immunol 5, 730-7 (2004).
109. Sun, L., Wu, J., Du, F., Chen, X. & Chen, Z.J. Cyclic GMP-AMP synthase is a cytosolic DNA sensor that activates the type I interferon pathway. Science 339, 786-91 (2013).
110. Gazquez-Gutierrez, A., Witteveldt, J., S, R.H. & Macias, S. Sensing of transposable elements by the antiviral innate immune system. RNA 27, 735-52 (2021).
111. Murray, P.D., McGavern, D.B., Pease, L.R. & Rodriguez, M. Cellular sources and targets of IFN-gamma-mediated protection against viral demyelination and neurological deficits. Eur J Immunol 32, 606-15 (2002).
112. Wang-Johanning, F. et al. Immunotherapeutic potential of anti-human endogenous retrovirus-K envelope protein antibodies in targeting breast tumors. J Natl Cancer Inst 104, 189-210 (2012).
113. Ng, K.W. et al. Antibodies against endogenous retroviruses promote lung cancer immunotherapy. Nature 616, 563-573 (2023).
114. Ishak, C.A., Classon, M. & De Carvalho, D.D. Deregulation of Retroelements as an Emerging Therapeutic Opportunity in Cancer. Trends Cancer 4, 583-597 (2018).
115. Hasenkrug, K.J. & Chesebro, B. Immunity to retroviral infection: the Friend virus model. Proc Natl Acad Sci U S A 94, 7811-6 (1997).
116. Mi, S. et al. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403, 785-9 (2000).
117. Chang, Y.H. & Dubnau, J. Endogenous retroviruses and TDP-43 proteinopathy form a sustaining feedback driving intercellular spread of Drosophila neurodegeneration. Nat Commun 14, 966 (2023).
118. Majdan, M. et al. Epidemiology of traumatic brain injuries in Europe: a cross-sectional analysis. Lancet Public Health 1, e76-e83 (2016).
119. Majdan, M. et al. Years of life lost due to traumatic brain injury in Europe: A cross-sectional analysis of 16 countries. PLoS Med 14, e1002331 (2017).
120. Ghajar, J. Traumatic brain injury. Lancet 356, 923-9 (2000).
121. Gardner, R.C. et al. Mild TBI and risk of Parkinson disease: A Chronic Effects of Neurotrauma Consortium Study. Neurology 90, e1771-e1779 (2018).
122. Stocchetti, N. & Zanier, E.R. Chronic impact of traumatic brain injury on outcome and quality of life: a narrative review. Crit Care 20, 148 (2016).
123. Sundman, M.H., Hall, E.E. & Chen, N.K. Examining the relationship between head trauma and neurodegenerative disease: A review of epidemiology, pathology and neuroimaging techniques. J Alzheimers Dis Parkinsonism 4(2014).
124. Jönsson, M.E. et al. Activation of endogenous retroviruses during brain development causes an inflammatory response. The EMBO Journal 40(2021).
125. Skene, P.J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. Elife 6(2017).
126. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21 (2013).
127. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079 (2009).

128. Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Research 42, W187-W191 (2014).
129. Liao, Y., Smyth, G.K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923-930 (2014).
130. Sperber, G.O., Airola, T., Jern, P. & Blomberg, J. Automated recognition of retroviral sequences in genomic data—RetroTector©. Nucleic Acids Research 35, 4964-4976 (2007).
131. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15(2014).
132. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. Cell 177, 1888-1902.e21 (2019).
133. Wu, T. et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation 2, 100141 (2021).
134. Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094-3100 (2018).
135. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. Nature Methods 9, 357-359 (2012).
136. Heinz, S. et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell 38, 576-589 (2010).
137. Quinlan, A. & Hall, I. BEDTools: a flexible suite of utilities for comparing genomic features. Vol. 26 841-842 (Bioinformatics, Oxford Univesity Press, 2010).
138. Loman, N.J., Quick, J. & Simpson, J.T. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nature Methods 12, 733-735 (2015).
139. Hardison, R.C. Comparative Genomics. PLoS Biology 1, e58 (2003).
140. Khan, H., Smit, A. & Boissinot, S. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Research 16, 78-87 (2006).
141. He, J. et al. Identifying transposable element expression dynamics and heterogeneity during development at the single-cell level with a processing pipeline scTE. Nat Commun 12, 1456 (2021).
142. Rodriguez-Quiroz, R. & Valdebenito-Maturana, B. SoloTE for improved analysis of transposable elements in single-cell RNA-Seq data using locus-specific expression. Commun Biol 5, 1063 (2022).
143. Rowe, H.M. et al. TRIM28 repression of retrotransposon-based enhancers is necessary to preserve transcriptional dynamics in embryonic stem cells. Genome Research 23, 452-461 (2013).
144. Brattas, P.L. et al. TRIM28 Controls a Gene Regulatory Network Based on Endogenous Retroviruses in Human Neural Progenitor Cells. Cell Rep 18, 1-11 (2017).
145. Wiznerowicz, M. et al. The Krüppel-associated Box Repressor Domain Can Trigger de Novo Promoter Methylation during Mouse Early Embryogenesis. Journal of Biological Chemistry 282, 34535-34541 (2007).
146. Ostertag, E.M., Goodier, J.L., Zhang, Y. & Kazazian, H.H. SVA Elements Are Nonautonomous Retrotransposons that Cause Disease in Humans. The American Journal of Human Genetics 73, 1444-1451 (2003).
147. Smajic, S. et al. Single-cell sequencing of human midbrain reveals glial activation and a Parkinson-specific neuronal state. Brain 145, 964-978 (2022).
148. Crab icon retreived from Flaticon.com

Acknowledgements

This is going to be so long. I am incredibly grateful for each and every person that has supported me and my research during these past years. Here, I highlight the main stars that made this possible:

To the entire **Molecular Neurogenetics lab**, present and past members. This is probably the most authentic "I could not have done this without you" statement ever. You have taught me so much. Plus, I would have a total amount of zero bytes of data if it were not for you. Your scientific discussions, your kindness, and your feedback on my projects made me grow in an unparalleled way during these years. I will forever be thankful to you for that. **Chromatin Dynamics lab**, I thank you for closely collaborating with us in these past few years. Your expertise and enthusiasm have perfectly complemented our nerdy circle and has advanced our understanding of human TEs. I thank you for all our scientific discussions; they were essential for this thesis.

To **both labs**, thank you for all the fikas, retreats, afterworks, dinners, conference trips, and our fantastic running club "The Jumping Genes". I have had a BLAST (ba dum tss). From here on, I will directly refer to crucial co-workers, friends, and family:

Johan (Jakobsson), my main supervisor, for answering my never-ending questions since my first day as a master's student in your lab. Thank you for challenging me, for sharing your scientific ambitions, and for all the opportunities and infrastructure you have offered to explore all my interests and curiosities. Thank you for keeping your office door open, for brainstorming with me, and for always being available for a chat. Most importantly, I truly thank you for believing and investing in my scientific development. I could not have asked for a more challenging yet believing supervisor. Needless to say, this work has only been possible because of your support.

Chris (Douse), for sharing your expertise and for having such a contagious scientific enthusiasm – it truly is a sunshine! Thank you for all the kindness you give, for always being up for a scientific chat, and for reviewing this kappa. So thankful for having you around!

Marie (Jönsson), for your friendship during all these years. I cherish it so much. You are a source of sunshine to everyone that is lucky enough to work with you. Thank you for explaining the TRIM28-KO mice breeding systems a million times to the new computer scientist in the lab back then. Thank you for teaching me Illustrator and sharing your enthusiasm for science with me. Thank you for all the laughs, life advice, for taking me to the ski slopes (in Skåne, right?), the wine, and that helmet. I could not have done it without you!

Yogita (Sharma), for always being there for any of my bioinformatics questions, scientific discussions, and for being there to share some of the frustrations that come with bioinformatics. Thank you for always being there for me and for including me in your ideas, projects, workshop organization, and meetings. You were an anchor during these years. Thank you for your friendship and scientific collaboration.

Per (Brattås), for taking me as your master's student back in 2018. I learned so much from you. Thank you for your patience back then, for all the fikas, and shared music. Thank you for all your support and for including me in your projects.

Pia (Johansson), for getting me into running! It was a major help during the PhD. Thank you for all the projects we have worked together, for answering all my questions about brain development, and for your patience throughout it. I have learned so much from you. This thesis would not have been possible without you.

Karolina (Pircs), for telling me to open myself to the people in the lab. It stuck with me and made the lab start feeling like friends. Your attitude, ambitions, and enthusiasm are like no other. Thank you!

Jenny (Johansson), this thesis would definitely not exist without the immense amount of sequencing data you have produced. Thank you for the excellence in your work, for keeping me in the loop for the design of our experiments, and for sharing your insights on it with me. Thank you for being such a chill and kind person.

Diahann (Atacho), for your collaboration and support. What a cool paper we made! Thank you for the encouragement and kindness during the highs and lows in these past years. Thank you for everything you taught me!

Vivien (Horvath), for your endless nerdiness, dedication, and enthusiasm around SVAs and ZNF91, I enjoyed so much the process of building this story. Thank you for the good laughs, and the huge amount of crazy-exciting data.

Patricia (Gerdes), for your scientific insights into my projects, for the discussions, and for being so supporting during the writing of my thesis. For always being a kind co-worker. Thank you for being a great office mate, for sharing your knowledge, and for reviewing this kappa!

Laura (Castilla), for being such a caring, authentic, and kind person. Truly happy to have you around. Thank you for all the scientific discussions and for being such a lovely office mate.

Carrie (Davis-Hansson), you have very quickly become an esteemed part of the lab. I admire your authenticity as a person, and your bravery during scientific discussions. Thank you for reviewing this kappa!

Malin Parmar, my co-supervisor, thank you for all the support during these years.

Paulina (Pettersson), because this lab would not run without you. Thank you for all the administrative help, the fika conversations, and the tiny plants!

To the PhD students from MN and CD labs: **Anita Adami, Ninoslav Pandiloski, Ofelia Karlsson**, and **Chandramouli Muralidharan**. The PhD became easier when you guys arrived. I have enjoyed so much your company. Thank you for your companionship during these years, for all our scientific discussions, and for the support in my projects.

Ofelia (Karlsson), thank you for being such a lovely office mate, co-worker, and friend. I admire your dedication, and how true you are to yourself. Thanks for being the voice of reason and kindness in all discussions in the lab, and for helping me translating the popular summary of this thesis!

Ninoslav (Pandiloski), thank you for all our zoom fikas during COVID, for helping me analyse the crazy amount of data we get, for all our scientific discussions, the bioinfo rants, the whiteboarding, for your never-ending questions, and your courage to ask them. Thank you for your friendship during this time, I truly cherish it.

Anita (Adami), you made me cry of laughter so many times during these past years. I am so lucky to have gotten such a close friend as a co-worker. Thank you for being with me during the hardest, but also the happiest bits of these years. Thank you for always lifting me up. I am so grateful for all the projects we have worked together! For all the discussions and our shared excitement for science.

Thank you to **Petter (Storm)**, for being such a kind co-worker and for all the bioinformatical support you have given me through these years. Thank you for joining me teaching in the Unistem day and the basic R course. **Shamit (Soneji)**, thank you for inviting me to assist in, hands-down, the coolest R course I have been at. **Stefan (Lang)**, thank you for guarding the server's integrity. To all of you, plus Yogita and Per, thanks for sharing and spreading the bioinfo hype!

Thank you to **Niklas Marklund** for building and sharing a precious cohort of human traumatic brain injury samples (TBI), and for collaborating with us in this project. Thank you to **Agnete Kirkeby** and **Arun Thiruvalluvan** for your support during the TBI revisions and our ongoing collaborations. Thank you to **Elisabeth Englund, Zaal Kokaia, Molly Hammel, Evan Eichler, Roger Barker, Mattias Belting, Patric Jern**, and all their teams, for all the support and collaboration in these projects – it would not have been possible without you. Thank you to my mentors **Markus Rigner** and **Katharina Herzog** for all your advice. Thank you **to all the technicians and administrative staff** for your excellence and support during this time.

From here on, I will address friends and family, but I want to finish this part of the acknowledgements saying that, coming from computer science, having so many women as co-workers really made a deep impact on me. I never felt more integrated and empowered. This thesis is dedicated to all women in STEM, many of which have lifted me up and supported me during these years. Who would have thought I would find so much empowerment simply by gaining your company? Thank you. I feel extremely lucky of the team I have had the pleasure to work with. It is full of ambitious and strong people.

Thank you to **Joel (Ströbaek)**, for being the funniest and kindest friend (competing only with Anita). Thank you for all the good times during these years and for helping me with the Swedish translation of the thesis' popular summary. I cherish your friendship so much!

Thank you to **Deborah (Figueiredo), Suze (Roostee), Lisa (Bodily), and Anna (Righini**), for your friendship during these years and the many wine nights.

Gracias a mi fiel amigo, **Alfredo (Dueñas)**, por mantener nuestra amistad intacta a pesar del tiempo y la distancia – ¡Irremplazable!

Gracias a mi familia por todo su amor y apoyo en esta decisión de casi siete años. **Rebeca (Garza)**, la vida no me pudo haber dado a una mejor hermana. Gracias por ser cercana a pesar de la distancia. Por siempre estar ahí para mí a una llamada de distancia. Rebe y **Gustavo (Cota)**, gracias por darme a la sobrina más hermosa del universo, **Helena (Cota)**. Gracias por apoyar a mis padres y a Mealy durante estos años. Han sido clave para mi paz aquí. **Magdalena (Gómez)**, sé que estos años han

sido difíciles – te doy las gracias por darme el ejemplo de nunca darse por vencida. Gracias por creer en mí, por alentarme, y por priorizar mi educación y crecimiento personal toda mi niñez. **Fernando (Garza)**, gracias por siempre haber valorado mi educación y entusiasmarme desde niña a aprender de todo tema que yo quisiera. Gracias por apoyarme a aventurarme a Suecia, aunque todo fuera incierto. Gracias por apoyarme en toda situación que se me atravesó durante estos años. **Rene (Gonzalez), Victoria (Larrea), y Daniel (Gonzalez)**, la vida no me pudo haber dado una mejor familia añadida. Qué suerte es tenerlos en mi vida. Gracias por todo su apoyo y amor durante estos años.

Finally, to David, my best friend and life-partner. You have unceasingly encouraged my ideas and have propelled me to reach my goals every single day since we met. Thank you for being sail and anchor during all these years. For all the peace and light you give me, thank you. Also, thanks for reviewing this kappa!

PAPER I

Activation of endogenous retroviruses during brain development causes an inflammatory response

Marie E Jönsson[1], Raquel Garza[1], Yogita Sharma[1], Rebecca Petri[1], Erik Södersten[2], Jenny G Johansson[1], Pia A Johansson[1], Diahann AM Atacho[1], Karolina Pircs[1], Sofia Madsen[1], David Yudovich[3], Ramprasad Ramakrishnan[4], Johan Holmberg[2], Jonas Larsson[3], Patric Jern[5] & Johan Jakobsson[1,*]

Abstract

Endogenous retroviruses (ERVs) make up a large fraction of mammalian genomes and are thought to contribute to human disease, including brain disorders. In the brain, aberrant activation of ERVs is a potential trigger for an inflammatory response, but mechanistic insight into this phenomenon remains lacking. Using CRISPR/Cas9-based gene disruption of the epigenetic co-repressor protein Trim28, we found a dynamic H3K9me3-dependent regulation of ERVs in proliferating neural progenitor cells (NPCs), but not in adult neurons. *In vivo* deletion of *Trim28* in cortical NPCs during mouse brain development resulted in viable offspring expressing high levels of ERVs in excitatory neurons in the adult brain. Neuronal ERV expression was linked to activated microglia and the presence of ERV-derived proteins in aggregate-like structures. This study demonstrates that brain development is a critical period for the silencing of ERVs and provides causal *in vivo* evidence demonstrating that transcriptional activation of ERV in neurons results in an inflammatory response.

Keywords brain development; CRISPR; microglia; transposable elements; Trim28
Subject Categories Immunology; Neuroscience
DOI 10.15252/embj.2020106423 | Received 4 August 2020 | Revised 22 January 2021 | Accepted 26 January 2021 | Published online 1 March 2021
The EMBO Journal (2021) 40: e106423

Introduction

About one–tenth of the human and mouse genomes is made up of endogenous retroviruses (ERVs) (Jern & Coffin, 2008). This is a result of the cumulative infection of the germ line by retroviruses over millions of years. ERVs are dynamically silenced at the transcriptional level during early development via epigenetic modifications, including histone methylation and deacetylation as well as DNA methylation (Yoder *et al*, 1997; Rowe *et al*, 2010). Together, these repressive mechanisms suppress ERV expression in somatic tissues. However, it is becoming increasingly clear that ERVs are aberrantly activated in various human diseases, including a number of neurological disorders. For example, ERV expression has been found to be elevated in the cerebrospinal fluid and in post-mortem brain biopsies from patients with multiple sclerosis, amyotrophic lateral sclerosis, Alzheimer's disease, Parkinson's disease, and schizophrenia (Perron *et al*, 1997; Garson *et al*, 1998; Andrews *et al*, 2000; Karlsson *et al*, 2001; Steele *et al*, 2005; MacGowan *et al*, 2007; Perron *et al*, 2008; Douville *et al*, 2011; Li *et al*, 2015; Guo *et al*, 2018; Sun *et al*, 2018; Tam *et al*, 2019b).

Aberrant activation of ERVs in the brain has been proposed to be directly involved in the disease process through a number of different mechanisms, including the activation of an innate immune response, direct or indirect neurotoxicity or by modulating endogenous gene networks (Saleh *et al*, 2019; Tam *et al*, 2019a; Jonsson *et al*, 2020). However, causal studies of ERV activation in the brain are challenging, since this phenomenon is difficult to model in the laboratory. Most experimental studies rely on ectopic expression of ERV-derived transcripts, often using xeno-overexpression at non-physiological levels, making it hard to interpret the results (see e.g, (Antony *et al*, 2004; Li *et al*, 2015)). Still, while the role of ERVs in neurological disorders remains unclear (Tam *et al*, 2019a), ERV activation may constitute a new type of disease mechanism that could be exploited to develop much needed therapy for these disorders. Direct experimental evidence on the mechanisms underlying ERV repression and the consequences of ERV activation in the brain is therefore needed.

We recently found that Trim28, an epigenetic co-repressor protein, silences ERV expression in mouse and human neural progenitor cells (NPCs) (Fasching *et al*, 2015; Brattas *et al*, 2017).

1 Laboratory of Molecular Neurogenetics, Department of Experimental Medical Science, Wallenberg Neuroscience Center and Lund Stem Cell Center, Lund University, Lund, Sweden
2 Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden
3 Division of Molecular Medicine and Gene Therapy, Department of Laboratory Medicine and Lund Stem Cell Center, Lund University, Lund, Sweden
4 Division of Clinical Genetics, Lund University, Lund, Sweden
5 Science for Life Laboratory, Department for Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden
*Corresponding author. Tel: +46 46 2224225; Fax: +46 46 2220559; E-mail: johan.jakobsson@med.lu.se

© 2021 The Authors. Published under the terms of the CC BY 4.0 license

Trim28 is recruited to genomic ERVs via Krüppel-associated box-zinc finger proteins (KRAB-ZFPs), a large family of sequence-specific transcription factors (Imbeault *et al*, 2017). Trim28 attracts a multiprotein complex that establishes transcriptional silencing and deposition of the repressive histone mark H3K9me3 (Sripathy *et al*, 2006). Trim28 is highly expressed in the brain and has been linked to behavioral phenotypes reminiscent of psychiatric disorders (Jakobsson *et al*, 2008; Whitelaw *et al*, 2010; Fasching *et al*, 2015).

In this study, we have investigated the consequences of *Trim28* deletion in the developing and adult mouse brain. We found that while Trim28 is needed for the repression of ERVs during brain development, it is redundant in the adult brain. Our results demonstrate the presence of an epigenetic switch during brain development, where the dynamic Trim28-mediated ERV repression is replaced by a different more stable mechanism. Interestingly, conditional deletion of *Trim28* during brain development resulted in ERV expression in adult neurons leading to an inflammatory response, including the presence of activated microglia and ERV-derived proteins in aggregate-like structures. In summary, our results provide direct experimental evidence *in vivo* for a link between aberrant ERV expression in the brain and an inflammatory response.

Results

CRISPR/Cas9-mediated deletion of *Trim28* in mouse NPCs

To evaluate the consequences of acute loss of *Trim28* in NPCs, we used CRISPR/Cas9 gene disruption. We generated NPC cultures from Rosa26-Cas9 knock-in transgenic mice (Fig 1A) (Platt *et al*, 2014), in which Cas9-GFP is constitutively expressed in all cells. These Cas9-NPCs were transduced with a lentiviral vector expressing gRNAs (LV.gRNAs) designed to target either exon 3, 4, or 13 of *Trim28* (g3, g4, g13) or to target *lacZ* (control). The vector also expressed a nuclear RFP reporter gene (H2B-RFP). Cas9-NPCs transduced with LV.gRNAs were expanded for 10 days, at which point RFP expressing cells were isolated by FACS (Fig 1A). To assess gene editing efficiency, we extracted genomic DNA from the RFP^+ Cas9-NPCs and performed DNA amplicon sequencing of the different gRNA target sites. We found that all three gRNAs (g3, g4 and g13) were highly effective, generating indels at a frequency of 98–99% at their respective target sequences (Fig 1B). The majority of these indels caused a frameshift in the Trim28 coding sequence that is predicative of loss-of-function alleles (Fig 1B) resulting in a near complete loss of Trim28 protein (Fig 1C). These results demonstrate that CRISPR/Cas9-mediated gene disruption is an efficient way to investigate the functional role of *Trim28*.

Acute loss of *Trim28* in mouse NPCs results in upregulation of ERVs

We next queried if acute *Trim28* deletion in NPCs influences the expression of ERVs and other transposable elements (TEs). We performed strand-specific 2 × 150 bp RNA-seq on LV.gRNA-transduced Cas9-NPCs and investigated the change of expression in different ERV families using a TE-oriented read quantification software, TEtranscripts (Jin *et al*, 2015), while individual elements were analyzed using a unique mapping approach. Both of these analyses revealed an upregulation of ERVs upon the CRISPR-mediated *Trim28*-KO in mouse NPCs (Figs 1D and E, and EV1A). We found 13 upregulated ERV families, including IAPs and MMERVK10C. Both IAPs and MMERVK10C are recent additions to the mouse genome, and these ERV families include many full length, transposition-competent elements with the potential to produce long transcripts and ERV-derived peptides. We confirmed increased transcription of MMERVK10C elements using quantitative RT–PCR (qRT–PCR) (Fig EV1B). We also investigated the expression of other classes of TEs such as LINE-1s and SINEs but found no evident evidence of significant upregulation. Thus, acute deletion of *Trim28* in NPCs causes transcriptional upregulation of ERVs.

Trim28 attracts a repression complex containing the histone methyltransferase SETDB1 that deposits H3K9me3 (Sripathy *et al*, 2006). CUT&RUN epigenomic analysis (Skene & Henikoff, 2017) of this histone modification in NPCs revealed that *Trim28*-controlled ERVs were covered by H3K9me3. For example, almost all full length members of the MMERVK10C family were covered by H3K9me3 (Fig 1F). Notably, only a handful of these individual elements were transcriptionally activated after *Trim28*-KO. This suggests that *Trim28* binds to many full length ERVs in NPCs but is only responsible for transcriptional silencing of a small subset of them. However, these elements are highly expressed upon *Trim28* deletion.

TEs have the potential to change the surrounding epigenetic landscape and consequently influence the expression of protein coding genes in their vicinity (Fasching *et al*, 2015; Brattas *et al*, 2017; Chuong *et al*, 2017). Accordingly, we found, for example, that genes located in the close vicinity of an upregulated MMERVK10C element also displayed a significant upregulation (Fig EV1C). In some instances, this was due to the activated ERV acting as an

Figure 1. CRISPR/Cas9-based deletion of Trim28 in NPCs results in upregulation of ERVs.

A A schematic of the workflow for *Trim28*-KO in mouse NPC cultures. Scale bars: embryo 1 mm, NPCs 20 μm.
B Estimation of gene editing at the *Trim28* loci using NGS-sequencing of amplicons. Black bars indicate % of frameshift indels. Columns show an average of two biological replicates per guide RNA and error bars show mean ± SD.
C Western blot confirmed the loss of Trim28 expression upon *Trim28*-KO in mouse NPCs.
D RNA-seq analysis of the expression of TE families using TEtranscripts
E The significantly upregulated TE families with a fold change larger than 0.5 upon *Trim28*-KO in mouse NPCs. The dashed line indicates significance.
F RNA-seq analysis of the *Trim28*-KO and control samples, visualizing full length MMERVK10C elements (left panels) and CUT&RUN analysis of H3K9me3 in mouse NPCs (right panel). The location of the full length MMERVK10Cs is indicated as a thick black line under each histogram.
G Example of transcriptional readthrough outside a full length MMERVK10C into a nearby gene.

Source data are available online for this figure.

© 2021 The Authors

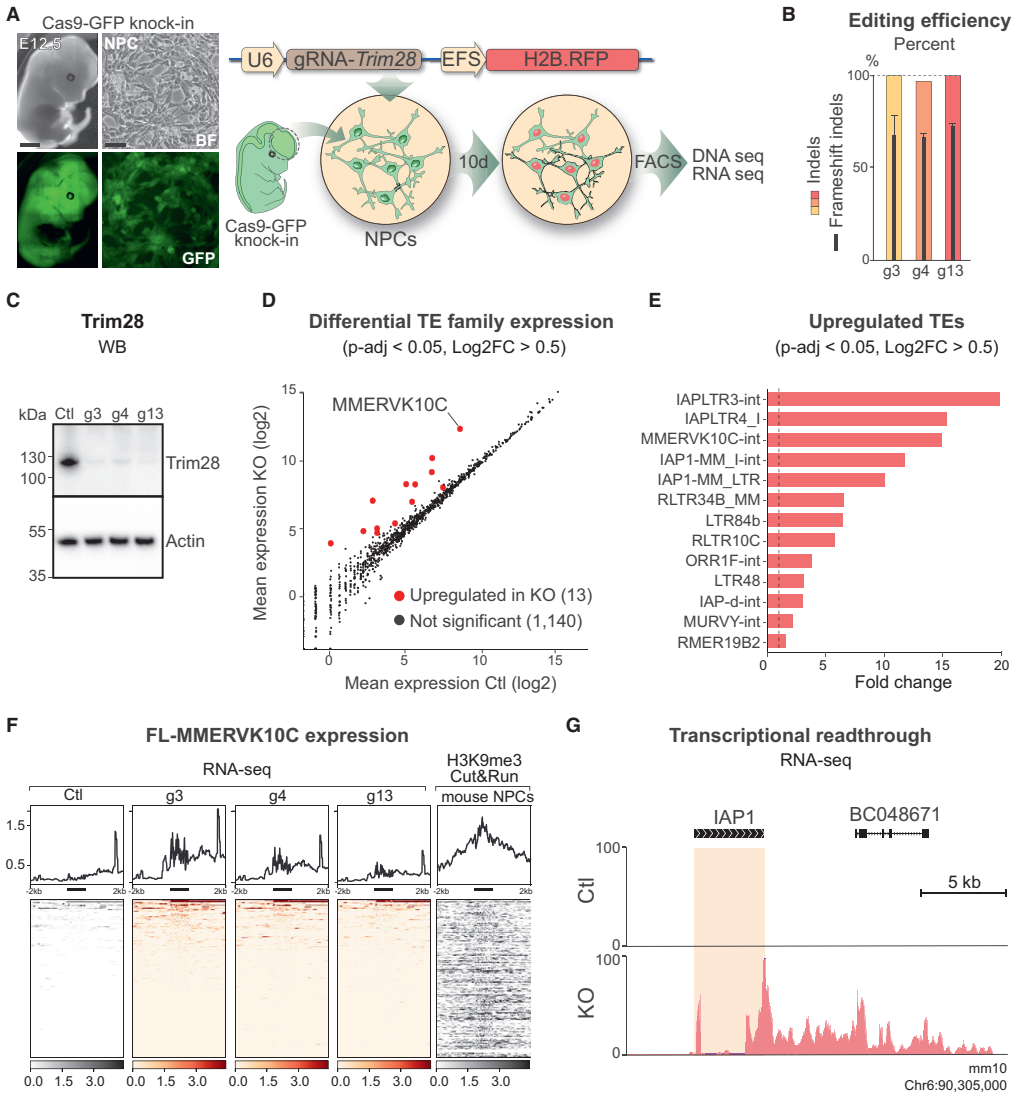


Figure 1.

alternative promoter (Fig 1G). Since Trim28 has additional functions in the cell (Quenneville *et al*, 2011) (Ziv *et al*, 2006) (Bunch *et al*, 2014), we also investigated the expression of all protein coding genes in *Trim28*-KO NPCs. This analysis revealed that acute loss of *Trim28* in NPCs had only a modest effect on protein coding genes (115 up- and 31 downregulated genes, respectively, Fig EV1D, Table EV1). In addition, PCA analyses of differentially expressed protein coding genes and TEs revealed that the *Trim28*-KO cells separated from control cells based on TE expression rather than gene expression (Fig EV1E and F). Together, these results demonstrate that Trim28 silence the transcription of ERVs in NPCs but has a modest direct effect on protein coding genes.

© 2021 The Authors

Deletion of *Trim28* in adult neurons *in vivo*

Alongside our previous data (Fasching *et al*, 2015; Brattas *et al*, 2017), these results confirm that *Trim28* is critical to repress ERVs in NPCs. However, it remains unclear how relevant these findings are to the situation *in vivo*. Therefore, we next investigated the consequences of deleting *Trim28* in mature neurons of adult mice. We designed adeno-associated viral vectors (AAV) vectors to drive expression of the *in vitro* verified gRNAs (AAV.gRNA) and used them separately or combined in the forebrain of Cas9-GFP knock-in mice. The AAV.gRNA vectors expressed H2B-RFP under the neuronal-specific synapsin promoter, allowing us to visualize and isolate transduced neurons (Fig 2A).

We injected AAV.gRNA into the forebrain of adult Cas9-GFP mice and sacrificed the animals after 8 weeks. The RFP expressing neuronal nuclei were isolated by FACS and amplicon sequenced to estimate the gene editing efficiency. All three gRNAs resulted in highly efficient gene editing (indel frequencies of 73–89%) where the majority of the indels were frameshift mutations (Fig 2B). Efficient deletion of *Trim28* was subsequently verified by immunohistochemistry (IHC) analysis, where quantification of Trim28 protein in RFP^+ cells showed loss of Trim28 expression in the majority of neurons (83–97%) in all of the groups (Fig 2C and D).

We next queried ERV expression in adult neurons lacking Trim28. We sequenced the RNA from the isolated RFP^+ nuclei and investigated the expression of ERV families as well as individual elements, using the same bioinformatical methods used for the NPCs. Remarkably, and in contrast to the NPC experiment, we observed no activation of ERVs upon *Trim28* deletion in adult neurons (Fig 2E). We also found no transcriptional activation of any other TE classes.

These results were particularly striking since Trim28 and many of its KRAB-ZFP adaptors are expressed in the brain, suggesting it is a significant organ for Trim28-mediated TE silencing (Imbeault *et al*, 2017). We therefore performed a series of additional control experiments to verify this finding. To ensure that the lack of ERV activation was not due to a bystander effect of nearby glial cells in which *Trim28* could be inactivated using our experimental setup, we developed an additional CRISPR-approach for cell type-specific deletion of *Trim28* in neurons using transgenic mice that conditionally express Cas9-GFP upon Cre expression (Stop-Cas9-GFP knock-in) (Platt *et al*, 2014) (Fig EV2A). We generated AAV vectors expressing the gRNAs and a Cre-inducible H2B-RFP reporter and an AAV vector expressing Cre under the control of the neuron-specific Synapsin1 (Syn) promoter. Upon transduction, Cre expression and subsequent expression of RFP and Cas9-GFP resulted in highly efficient neuron-specific gene editing of *Trim28* (Fig EV2B–D). RNA-seq analysis for TE expression revealed no ERV activation upon *Trim28* removal (Fig EV2E), which were in line with our results from the ubiquitous Cas9-GFP knock-in mice. To further verify that the lack of ERV expression was not caused by potential Cas9-mediated side effects which may occur *in vivo*, we injected AAV vectors expressing Cre into the forebrain of adult floxed *Trim28* animals (Cammas *et al*, 2000) (Fig EV2F). Again, we obtained a highly efficient *Trim28* deletion in adult mouse neurons but did not observe ERV activation (Fig EV2G–I). Furthermore, ChIP-seq from adult mouse forebrain (Jiang *et al*, 2017) showed a lack of H3K9me3 accumulation on MMERVK10C sequences in adult neurons, in line with our observed lack of transcriptional activation upon *Trim28* deletion (Fig 2F). Taken together, these results demonstrate that Trim28 is not required to silence the transcription of ERVs in adult neurons.

Deletion of *Trim28* in NPCs *in vivo*

Our results demonstrate that Trim28 is essential for transcriptional repression of ERVs in NPCs, but not in mature neurons. This suggests the existence of an epigenetic switch, where the dynamic and reversible Trim28/H3K9me3-mediated repression found in brain development is replaced by a different stable silencing mechanism in the adult brain. This is similar to what has been observed in early development where Trim28 participates in the establishment of DNA methylation to stable silence transposable elements (Wiznerowicz *et al*, 2007). To test this hypothesis, we deleted *Trim28* in dividing neural progenitors *in vivo*, which give rise to mature neurons in adulthood. We bred *Emx1*-Cre transgenic mice (Iwasato *et al*, 2000) with *Trim28*-flox mice (Cammas *et al*, 2000), resulting in *Trim28* deletion in cortical progenitors starting from embryonic day 10 (Figs 3A and EV3A). For better visualization of *Trim28*-excised cells by IHC, we included a *Cre*-inducible GFP reporter (gtRosa26-Stop-GFP) in the breeding scheme. With this setup, GFP expressing cells will correspond to cells in which Trim28 was deleted during development.

Emx1-Cre (+/−), *Trim28*-flox (+/+) mice were born at the expected ratio and survived into adulthood. Their overall brain morphology and size was not affected by the loss of *Trim28* during cortical development. IHC analysis of adult brains revealed that Trim28 protein was absent in virtually all pyramidal cortical neurons and that these cells also expressed GFP, demonstrating a highly efficient Cre-mediated excision of *Trim28* during development

Figure 2. CRISPR/Cas9 deletion of Trim28 in adult neurons *in vivo*.

A A schematic of the workflow targeting *Trim28* in the mouse forebrain using AAV vectors expressing the gRNA and a nuclear RFP reporter. 8 weeks later, the injected animals were analyzed either by immunohistochemical analysis or nuclei isolation by FACS prior to DNA/RNA-sequencing.

B Estimation of gene editing at the *Trim28* loci using NGS-sequencing of amplicons from DNA isolated from 50,000 RFP^+ nuclei per animal. One animal per group was analyzed. Black bars indicate % of the detected indels that disrupted the frameshift.

C, D Gene editing of the *Trim28*-loci resulted in a robust loss of Trim28 protein, as evaluated by IHC where the expression of Trim28 in RFP^+ cells was quantified and is displayed as mean ± SEM. Approximately 600 RFP^+ cells per animal and group was evaluated. Scale bar 30 μm.

E RNA-seq analysis of the expression of TE families using TEtranscripts.

F RNA-seq analysis of the *Trim28*-KO and control samples, visualizing full length MMERVK10C elements (left panels) and ChIP-seq analysis of H3K9me3 in adult forebrain neurons (right panel). The location of the full length MMERVK10Cs is indicated as a thick black line under each histogram.

Source data are available online for this figure.

© 2021 The Authors

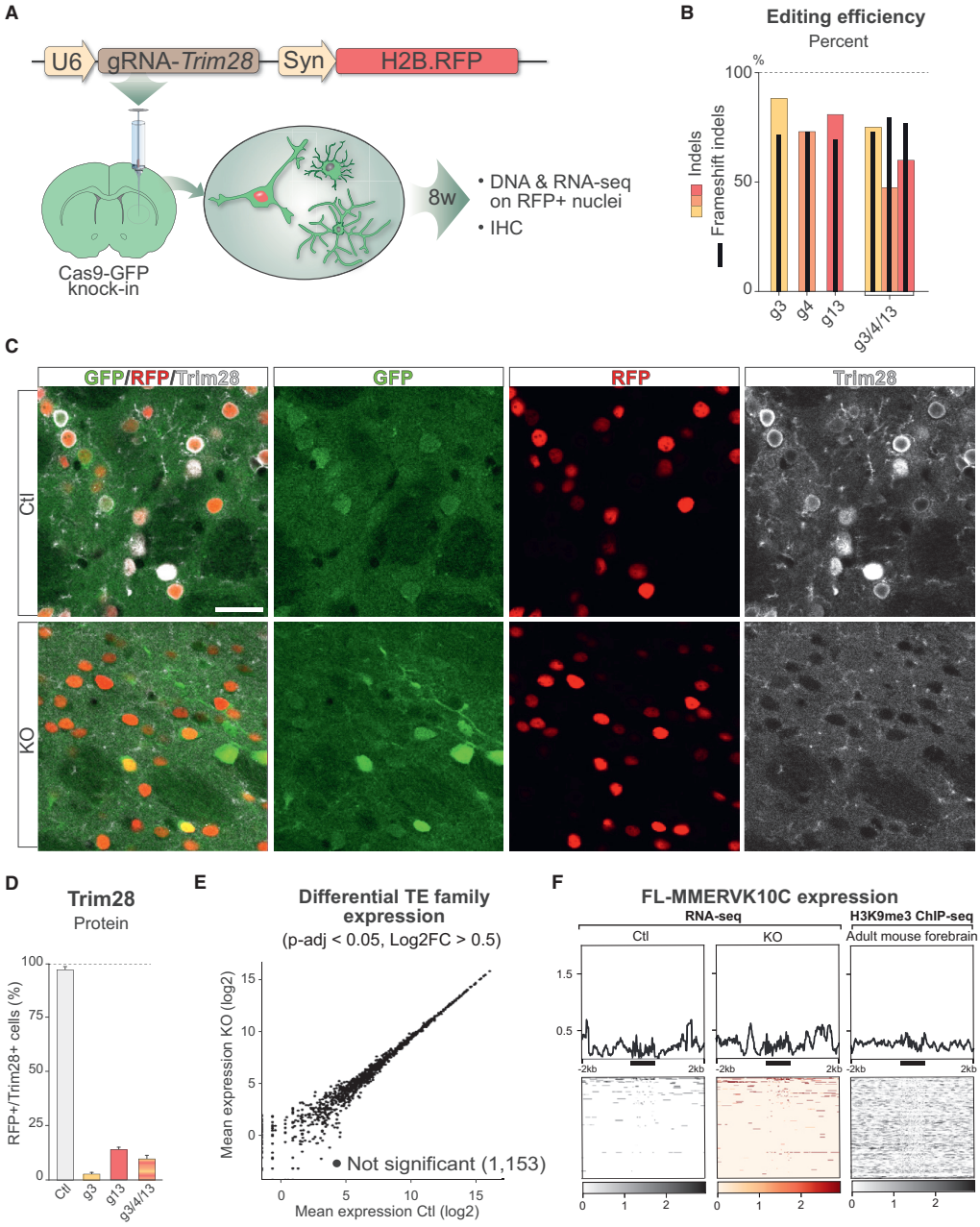


Figure 2.

© 2021 The Authors

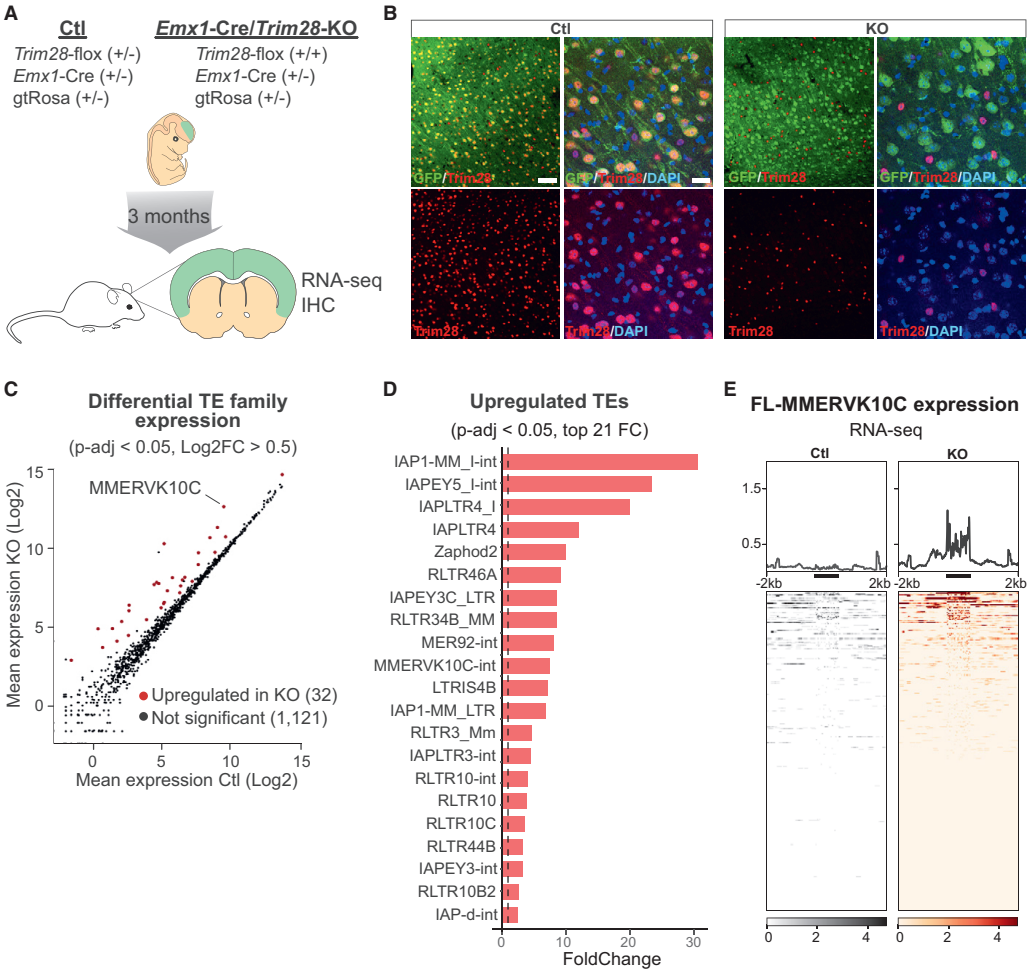


Figure 3. Deletion of Trim28 during brain development results in aberrant TE expression in the adult brain.

A A schematic of the breeding scheme resulting in highly efficient conditional deletion of *Trim28*-KO during cortical development and analyzing the adult tissue 3 months later by IHC and RNA-seq.

B IHC for Trim28 in the adult cortex revealed that the protein was lost in cells exposed to Cre-activity during brain development (GFP^+ cells). Scale bars: low magnification 75 μm, high magnification 20 μm.

C RNA-seq analysis of the expression of TE families using TEtranscripts

D Significantly upregulated TE families upon the *Trim28*-KO, in which the families with the highest fold change are listed.

E RNA-seq analysis of full length MMERVK10Cs in the adult tissue. The location of the full length MMERVK10Cs is indicated as a thick black line under each histogram.

(Figs 3B and EV3A). Cells that did not express GFP—for example, microglia and interneurons—expressed Trim28 as expected.

We performed RNA-seq on adult cortical tissue from these animals and control siblings to analyze ERV expression. In adult cortical tissue from *Emx1*-Cre (+/−), *Trim28*-flox (+/+), we observed a robust upregulation of several ERV families, many of which were also upregulated *in vitro* in the NPC experiment, including, e.g, MMERVK10C (Figs 3C and EV3B—individual TEs). Similar

© 2021 The Authors

to the NPC experiment, we found that only a few full length elements were activated, but that these were highly expressed (Fig 3D and E). Thus, the deletion of *Trim28* in NPCs *in vivo* during brain development results in the expression of ERVs in the adult brain. Notably, and in contrast to the *Trim28*-KO in NPCs *in vitro*, we did not detect an effect on nearby gene expression by activation of ERVs (Fig EV3C) nor observe any transcriptional readthrough from activated ERV elements into neighboring genes. This also included the very same elements that had this property *in vitro* in NPCs (Figs 3E and EV3D). These results demonstrate that deletion of *Trim28* during brain development *in vivo* results in high-level expression of ERVs in the adult brain.

Downstream transcriptional consequences of ERV activation in the brain

Our finding that *Emx1*-Cre (+/−), *Trim28*-flox (+/+) mice survive—despite high levels of ERV expression in the brain—raises the question about the downstream consequences of ERV activation *in vivo*. We first compared the expression of protein coding genes in *Emx1*-Cre (+/−), *Trim28*-flox (+/+) mice to their control littermates. We included RNA-seq data from animals in which *Trim28* was deleted in adult neurons (AAV.Syn-Cre, *Trim28*-flox (+/+)) in this analysis since these samples provide an important control setting in which the loss of *Trim28* does not impact ERV expression but only other Trim28 targets (Fig EV2F–I). Analysis of the *Emx1*-Cre (+/−), *Trim28*-flox (+/+) mice revealed 164 significantly upregulated and 86 significantly downregulated genes, of which 26 of the upregulated and 13 of the downregulated were similarly changed in the AAV.Cre, *Trim28*-flox (+/+) mice (Fig EV3E–G, Table EV1). This demonstrates that the loss of Trim28 during brain development causes substantial downstream effects on gene expression. Gene ontology (GO) analysis on molecular and biological pathways of genes specifically altered in *Emx1*-Cre (+/−), *Trim28*-flox (+/+)—but not in AAV.Cre, *Trim28*-flox (+/+) mice—revealed significant changes in genes related to cell adhesion (Fig EV3H).

Single-nuclei RNA-seq analysis of ERV-expressing brain tissue

These results indicate that the activation of ERVs in neurons results in downstream transcriptional effects that could have an impact on neuronal function. However, the brain is a complex tissue composed of several cell types located in close vicinity. To separate cell-intrinsic effects from cell-extrinsic effects, we performed single-nuclei RNA-seq analysis on forebrain cortical tissue dissected from *Emx1*-Cre (+/−), *Trim28*-flox (+/+) and control littermates. High-quality single-nuclei sequencing data were generated from a total of 14,296 cells, including 7,670 from *Emx1*-Cre (+/−), *Trim28*-flox (+/+) mice and 6,626 from control littermates (Fig 4A).

We first performed an unbiased clustering analysis to identify and quantify the different cell types present in the brain tissue. We detected seven different clusters (Figs 4B and EV4A–H), including excitatory and inhibitory neurons as well as several different glial populations, with excitatory neurons making up the largest cluster with more than half of the cells (Fig 4B). Overall, there was no major difference in cell number proportions between *Emx1*-Cre (+/−), *Trim28*-flox (+/+) and control animals (Fig 4C and D). For example, we found no reduction of excitatory neurons in the *Emx1*-Cre (+/−), *Trim28*-flox (+/+) animals even though this cell population completely lacked *Trim28*, as demonstrated by IHC of Trim28 and GFP. We conclude that deletion of *Trim28* during brain development in neuronal progenitors does not result in significant cell death.

Next, we analyzed transcriptional differences between the two genotypes. Among the dysregulated genes in excitatory neurons, we found altered expression of several lncRNAs and protein coding genes (Table EV2), many of which are linked to neurological disorders. Notably, we observed reduced expression of *Hecw2*, a ubiquitin ligase linked to neurodevelopmental delay (Berko *et al*, 2017) and *Sgcz*, a transmembrane protein linked to mental retardation (Piovani *et al*, 2014) (Fig 4E). We also found transcriptional alterations in astrocytes and oligodendrocytes (Fig 4F and G), two additional cell types in which *Emx1* is expressed during brain development and therefore should lack Trim28 expression. Among the dysregulated genes in astrocytes, we detected upregulation of *Apoe*, the key risk variant gene for Alzheimer's disease, *Rora*, a nuclear hormone receptor linked to intellectual disability, epilepsy and autism (Guissart *et al*, 2018) and downregulation of *Auts2*, a transcriptional regulator linked to several neurodevelopmental disorders (Oksenberg & Ahituv, 2013). In oligodendrocytes, we observed transcriptional changes of *Fth1*, a ferritin gene linked to neurodegeneration (Muhoberac & Vidal, 2019), and *Ngr3*, a ligand to tyrosine kinase receptors that has been linked to schizophrenia (Kao *et al*, 2010). In contrast, interneurons, a neuronal subtype that maintains *Trim28* expression in *Emx1*-Cre (+/−), *Trim28*-flox (+/+) mice, displayed no evidence of altered gene expression in interneurons.

Interestingly, we noted that also microglial cells displayed transcriptome alterations (Fig 4H). For example, microglia showed a reduced expression of *Csmd1*, a complement regulatory gene linked to familial epilepsy (Naseer *et al*, 2016) and schizophrenia (Schizophrenia Psychiatric Genome-Wide Association Study & C, 2011), and of the protocadherin *Pcdh9*, which is a risk factor for major depressive disorder (Xiao *et al*, 2018). Similar to oligodendrocytes, microglia showed a downregulation of *Nrg3* (Fig 4H). Microglia are immune cells of endodermal origin that do not express Emx1 during development and therefore maintains *Trim28* expression in *Emx1*-Cre (+/−), *Trim28*-flox (+/+) animals. Taken together, these results demonstrate that the ERV activation in excitatory neurons, due to the loss of Trim28 during brain development, results in both cell autonomous and non-cell autonomous effects, specifically on microglia where some of the downstream dysregulated genes have been previously linked to psychiatric disorders.

Cell type-specific analysis of ERV activation in the brain

The single-nuclei RNA-seq analysis indicated that microglia display transcriptional alterations despite maintaining Trim28 expression. Thus, microglia should be affected through cell-extrinsic mechanism mediated by derepressed ERVs from adjacent cells lacking Trim28. To verify this observation, we devised a strategy to analyze the expression of ERVs in the different cell populations in the *Emx1*-Cre (+/−), *Trim28*-flox (+/+) animals and control littermates. Since current pipelines available for high-throughput single cell RNA-seq analysis do not allow for estimation of TE expression, we developed an approach that initially uses the cell clusters established based on

© 2021 The Authors

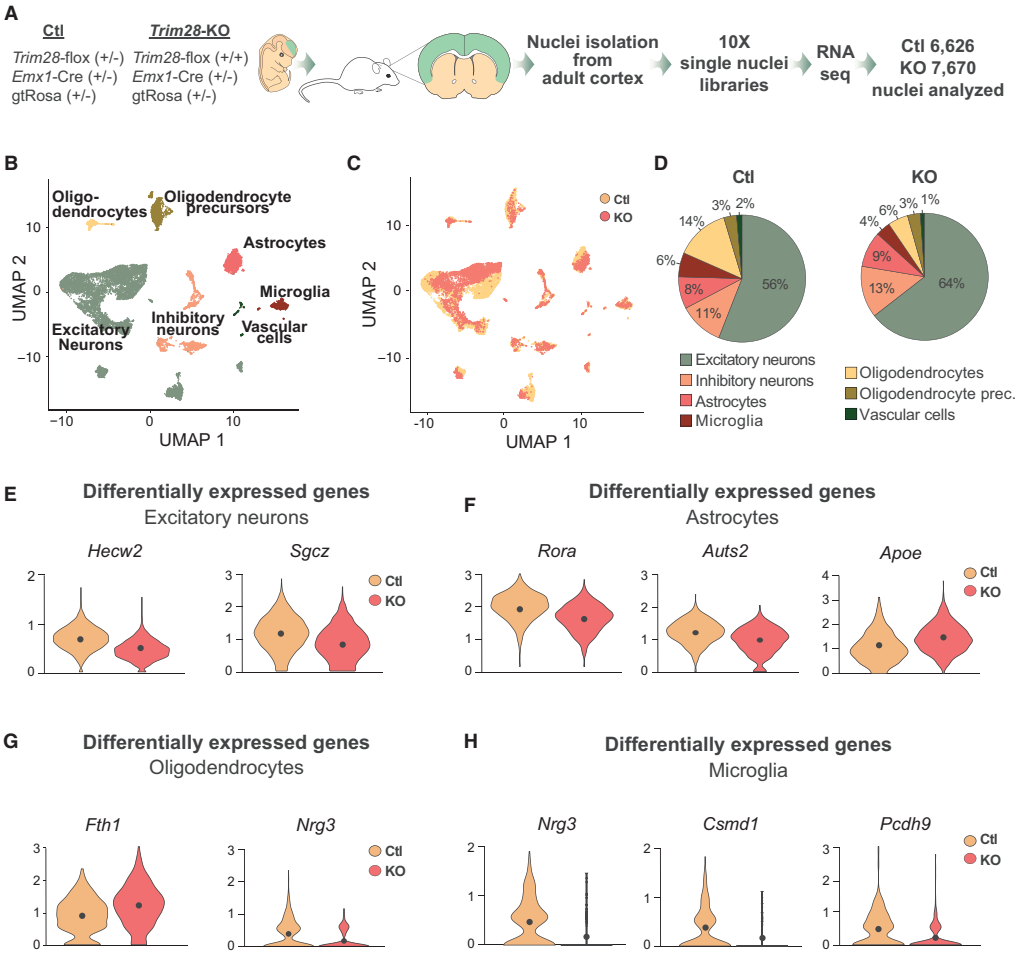


Figure 4. Single-nuclei RNA-seq of cortical tissue from *Emx1*-Cre/*Trim28*-KO animals and their littermates.

A A schematic of the workflow for the single-nuclei RNA-seq of cortical tissue from *Emx1*-Cre/*Trim28*-KO animals and their littermates.

B UMAP showing the unbiased clustering analysis with seven different cell clusters.

C, D UMAP and pie charts showing the distribution of *Trim28*-KO and control cells over the seven different clusters. There were no major differences in proportions of the different cell clusters between *Emx1*-Cre/*Trim28*-KO animals and controls.

E–H A selection of significant cell-type-specific changes in gene expression between *Emx1*-Cre/*Trim28*-KO animals and controls as revealed by single-nuclei RNA-seq. The black dots represent the mean value (Wilcoxon rank sum test, (*P*-adj value < 0.01), $n = 2$. For a full list, see Table EV2.

the gene expression (Fig 4A and B). Then, by back-tracing the reads from cells forming each cluster we were able to analyze the expression of ERVs and other TEs using TEtranscripts in distinct cell populations (Fig 5A) (Jin *et al*, 2015), increasing the sensitivity of TE expression at single cell type-resolution.

When using this bioinformatical approach, we found that different ERV families (e.g, MMERVK10C) were expressed in excitatory neurons, astrocytes, oligodendrocytes, and oligodendrocytes progenitors, i.e, the cell types in which *Emx1* is expressed during brain development and thus lack Trim28 (Fig 5B). Notably, we

© 2021 The Authors

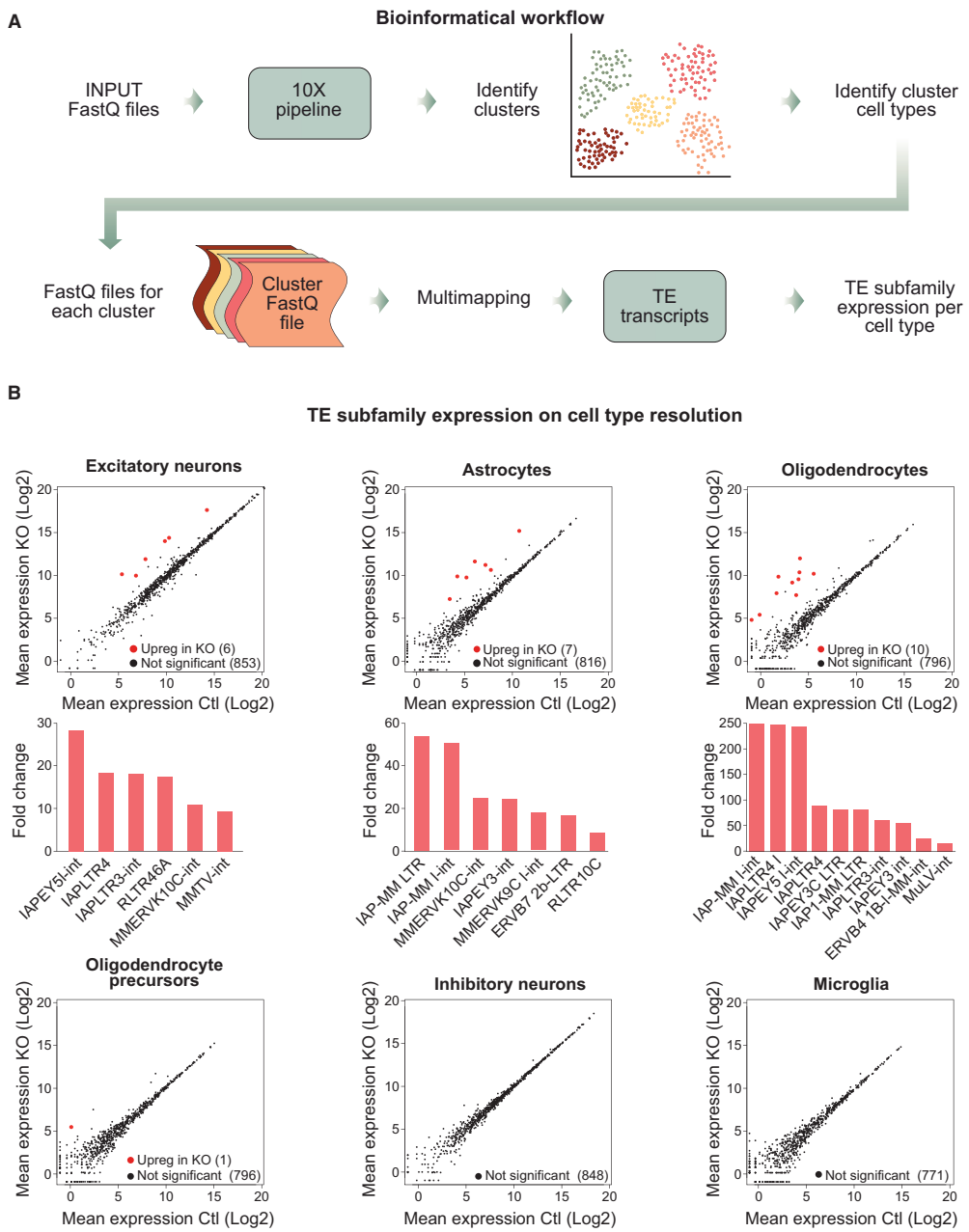


Figure 5.

© 2021 The Authors

Figure 5. Cell type-specific analysis of ERV activation in *Emx1*-Cre/*Trim28*-KO animals.

A The workflow used to analyze ERV expression in single-nuclei RNA-seq data.

B Mean plots show the changes of TE subfamily expression in each cell cluster upon the *Emx1*-Cre/*Trim28*-KO. A differential expression analysis performed with DESeq2 (described in material and methods) showed upregulated TEs in cell types in which *Trim28* was deleted (excitatory neurons, astrocytes, oligodendrocytes, and oligodendrocyte precursors), (indicated with red dots: *P*-adj < 0.01, log2FC > 3). The specific upregulated elements and their fold changes are listed in bar graphs under each mean plot. In cell types in which *Trim28* was not deleted, TE expression remained unaffected (inhibitory neurons and microglia).

found differences in ERV expression in distinct cell types, including activation of different families, demonstrating a cell type-specificity in the ERV-response to *Trim28* deletion. However, we found no upregulation of ERV expression in interneurons and microglia where Trim28 expression is maintained (Fig 5B). This analysis confirms that the transcriptional alterations observed in microglia are due to downstream cell-extrinsic effects.

Expression of ERVs in the brain is linked to an inflammatory response

The microglia response to the *Trim28*-KO in neurons is intriguing as the aberrant expression of ERVs and other TEs have been linked to inflammatory responses (Hurst & Magiorkinis, 2015; Lim *et al*, 2015; Roulois *et al*, 2015; Thomas *et al*, 2017; Ishak *et al*, 2018; Saleh *et al*, 2019; Tam *et al*, 2019a). To study this further, we analyzed microglial cells by IHC for the pan-myeloid marker Iba1 (ionized calcium-binding adapter molecule 1, encoded by the Allograft inflammatory factor 1 (*Aif1*) gene). We found that the microglial cells in *Emx1*-Cre (+/−), *Trim28*-flox (+/+) mice displayed signs of activation, including higher expression of Iba1 (Fig 6A). Automated high-content screening microscopy analysis revealed that, although the density of Iba1 + cells was unaffected (Fig EV5A), the microglia present in the cortex of *Emx1*-Cre (+/−), *Trim28*-flox (+/+) mice displayed a morphology that is typical for an activation phenotype, including longer and thicker processes with increased numbers of branches (Fig 6B). Interestingly, we only found activated microglia in the cortex, where excitatory neurons lack Trim28, but not in the nearby forebrain structure striatum, where Trim28 is still expressed in all neurons (Fig 6C). The inflammatory response was therefore spatially restricted to the area with increased ERV expression. We also detected an increased expression of CD68, a lysosomal protein upregulated in activated microglia, in the cortex of *Emx1*-Cre (+/−), *Trim28*-flox (+/+) animals, a further indication of an ongoing inflammatory response (Fig 6D). In addition, we found no signs of gliosis or an inflammatory response in animals where *Trim28* was deleted in mature neurons, a setting where Trim28 is removed but no ERVs are activated (AAV.Syn-Cre/*Trim28*-flox) (Fig EV5B). These results verify that the inflammatory response was not activated by the loss of *Trim28 per se* or by direct *Trim28*-targets in neurons, but more likely caused by the expression of ERVs.

ERV-derived proteins are found in areas of microglia activation

A number of recent studies have demonstrated that the presence of ERV-derived nucleic acids can activate the innate immune system, such as double-stranded RNAs or single-stranded DNA (Hurst & Magiorkinis, 2015; Roulois *et al*, 2015; Ishak *et al*, 2018). According to this model, the host cells misinterpret the expression of ERVs as a viral infection, and this triggers an autoimmune response through the activation of interferon signaling. Therefore, we searched for evidence of an interferon response and activation of a viral defense pathway in the transcriptome data (bulk RNA-seq and single-nuclei RNA-seq) from *Emx1*-Cre (+/−), *Trim28*-flox (+/+) mice. However, we found no evidence of activation of these pathways suggesting that the expression of genes linked to the innate immune response and viral response were not transcriptionally activated despite high levels of ERVs in the brain (Fig EV5C and D). Similarly, we found no evidence of this response in the *Trim28*-KO NPCs *in vitro*, in which ERVs were activated. Thus, deletion of *Trim28* and subsequent ERV activation in neural cells does not result in a detectable interferon response, suggesting that other cellular mechanisms are responsible for triggering the observed inflammatory response.

An alternative mechanism for ERV-mediated triggering of the inflammatory response is the expression of ERV-derived peptides and proteins, which could have neurotoxic properties (Li *et al*, 2015). To search for evidence of ERV-derived proteins, we performed WB and IHC analysis using an antibody detecting IAP-Gag protein in the brain of *Emx1*-Cre (+/−), *Trim28*-flox (+/+) mice. We found high levels of IAP-Gag protein expression in the brain of mice expressing elevated ERV transcripts, demonstrating their efficient translation into proteins (Fig 6E and F). The IAP-Gag labeling was restricted to cortical excitatory neurons lacking Trim28, as visualized by the co-expression of the Cre-dependent GFP reporter. Notably, the IAP-Gag labeling was not uniform, as some neurons expressed higher levels of IAP-Gag and some contained IAP-Gag in aggregate-like structures (Fig 6F), suggesting that the expression of ERVs in the brain results in the accumulation of ERV-derived proteins.

Discussion

In this study, we define epigenetic mechanisms that control the expression of ERVs during brain development and investigate the consequences of their inactivation. Previous work has implicated ERVs and other TEs in several neurological disorders, such as MS, AD, ALS, PD, and schizophrenia, where an increased expression of TEs has been reported along with the speculation of their contribution to the disease process (Perron *et al*, 1997; Garson *et al*, 1998; Andrews *et al*, 2000; Karlsson *et al*, 2001; Steele *et al*, 2005; MacGowan *et al*, 2007; Perron *et al*, 2008; Douville *et al*, 2011; Li *et al*, 2015; Guo *et al*, 2018; Sun *et al*, 2018; Tam *et al*, 2019b; Jonsson *et al*, 2020). However, these clinical observations have been difficult to interpret since the results are correlative. Although there are studies that indicate a causality between upregulated TEs and neurodegeneration using both *Drosophila* and mouse model systems (Li *et al*, 2012; Li *et al*, 2013; Krug *et al*, 2017; Guo *et al*, 2018; Kremer *et al*, 2019; Sankowski *et al*, 2019; Dembny *et al*,

© 2021 The Authors

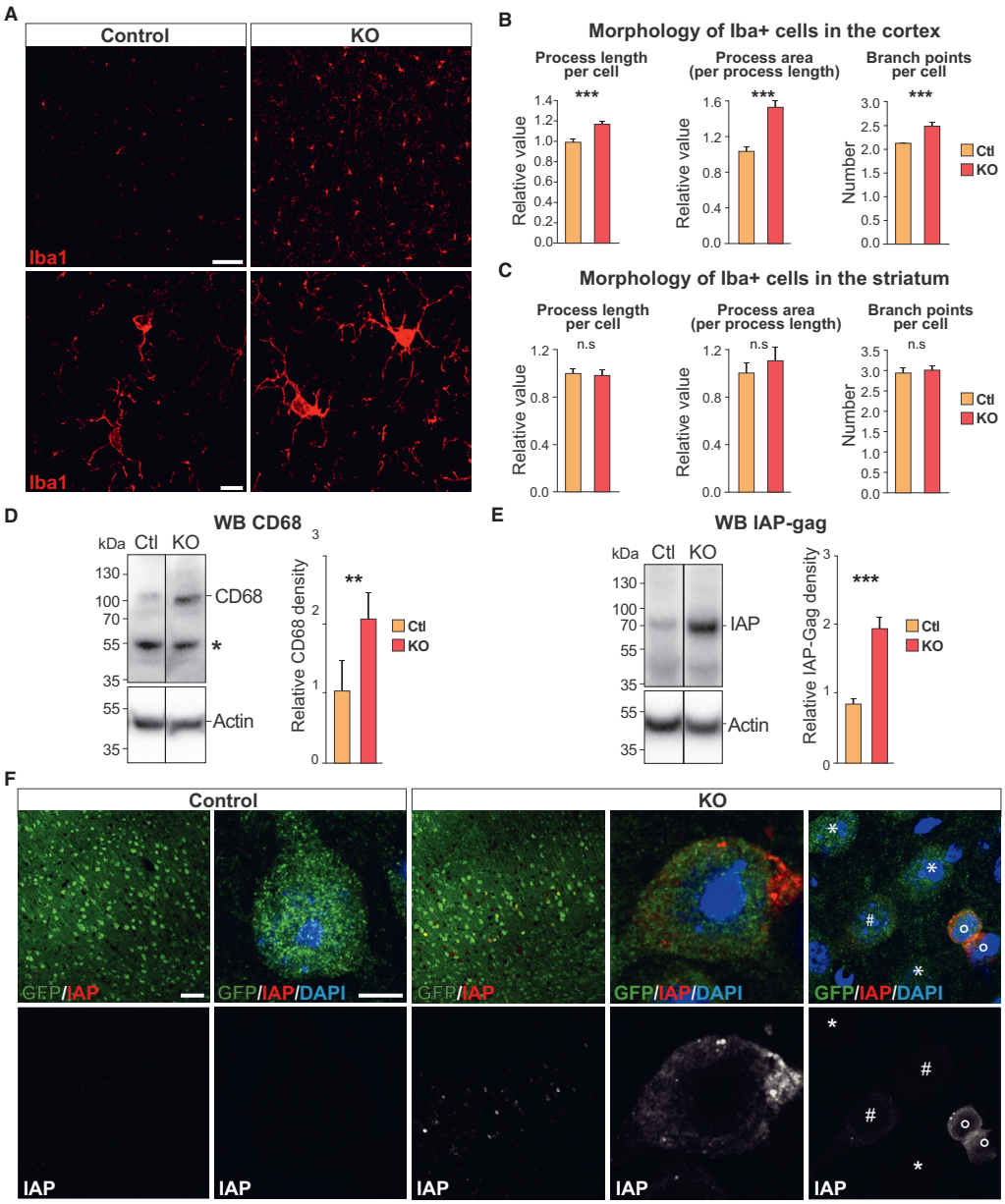


Figure 6.

© 2021 The Authors

Figure 6. Activation of ERVs during brain development results in the presence of ERV proteins in adult brain tissue and signs of an inflammatory response.

A IHC analysis for the microglia marker Iba1 in *Emx1*-Cre/*Trim28*-KO animals and controls. Scale bars: Low magnification 75 μm, high magnification 10 μm.

B, C The morphology of microglia in cortex and striatum (control region) of *Emx1*-Cre/*Trim28*-KO animals and controls were quantified by high-content screening of Iba1 immunoreactivity, revealing differences in process length, area, and number of branch points specifically in cortex, error bars mean ± SEM (unpaired *t*-test. *P* values: 1.8×10^{-4}, 9.6×10^{-4}, and 5.2×10^{-6}, respectively). A large number of photographs from three control and two KO animals were analyzed, see material and methods for details.

D Western blot revealed an increased expression of CD68 in the *Emx1*-Cre/*Trim28*-KO animals ($n = 5$ per group), unpaired *t*-test, $^{**}P = 0.0036$, error bars mean ± SEM. The star indicates an unspecific band.

E An increased expression of the ERV-derived protein IAP-Gag was detected in the cortex of *Emx1*-Cre/*Trim28*-KO animals ($n = 5$ per group), unpaired *t*-test, $^{***}P = 0.0008$, error bars mean ± SEM.

F Immunohistochemistry for IAP-Gag visualized the presence of ERV proteins in the cortex of *Emx1*-Cre/*Trim28*-KO animals. The distribution of IAP-Gag was heterogenous among *Trim28*-KO neurons, where it was either accumulated in a low (#) or high (°) number of aggregate-like structures in the cytoplasm, or as weak homogenous staining throughout the cytoplasm (#) or not present at all (*). Scale bars: low magnification 75 μm, high magnification 5 μm.

Source data are available online for this figure.

2020), modeling of ERV activation in an experimental setting is challenging. Thus, the putative involvement of ERVs in brain disorders has remained controversial and direct experimental evidence on the consequences of ERV activation in the brain has been needed (Tam *et al*, 2019a; Jonsson *et al*, 2020). This study demonstrates that aberrant activation of ERVs during brain development *in vivo* triggers an inflammatory response linked to the presence of ERV-derived proteins present in aggregate-like structures in the adult brain. This increases our understanding of the consequences of ERV activation in the brain and provides new mechanistic insights opening up for further research into the role of ERVs and other TEs in various brain disorders.

Brain development is a critical period for the repression of ERVs and other TE, during which several chromatin-associated factors, such as DNMTs, HUSH, and Trim28/KRAB-ZFPs, work in parallel to mediate transcriptional silencing (Fasching *et al*, 2015; Brattas *et al*, 2017; Liu *et al*, 2018; Jonsson *et al*, 2019). Our data demonstrate that Trim28 dynamically controls the expression of ERVs through a mechanism linked to the repressive histone mark H3K9me3. Remarkably, we also found that Trim28 is required for the establishment of a different, stable silencing mechanism that is present in adult neurons—most likely DNA methylation (Wiznerowicz *et al*, 2007). Importantly, once established this system is independent of Trim28 and H3K9me3. This observation suggests that the ERV silencing machinery is particularly vulnerable during brain development and that perturbation of the system during this period could have life-long consequences. If Trim28-mediated silencing of ERVs is impaired during brain development, for example through mutations in KRAB-ZFPs or by environmental influence, ERVs may be derepressed in adult neurons, resulting in inflammatory response.

It is well established that many psychiatric disorders are a consequence of neurodevelopmental alterations (Bale *et al*, 2010; Horwitz *et al*, 2019). A combination of genetic components and environmental exposures are thought to contribute to the appearance and progress of these diseases, indicating that epigenetic alterations play a key role in the disease process (see e.g, (Khashan *et al*, 2008; Susser *et al*, 2008)). Many psychiatric disorders also have an immune component that is thought to play an important role in the disease process (Bauer & Teixeira, 2019). This immune response includes both peripheral activation and glial activation, in the brain. The underlying cause for this immune response remains largely unknown. However, the combination of epigenetic dysregulation and immune activation in psychiatric disorders, together with the observations made in the current study, suggest that ERVs are directly involved in the disease process. We have previously demonstrated that the deletion of *Trim28* during postnatal forebrain development or heterozygous deletion of *Trim28* during early brain development results in complex behavioral changes including hyperactivity and an impaired stress-response (Jakobsson *et al*, 2008) (Fasching *et al*, 2015). In addition, heterozygous germ line deletion of *Trim28* in mice has been described to provoke an abnormal exploratory behavior (Whitelaw *et al*, 2010). These findings demonstrate that disruption of Trim28 levels in the mouse brain results in behavioral changes, similar to impairments found in psychiatric disorders. However, Trim28 is a multi-role protein which executes functions in the cell that is unrelated to the control of ERVs and we cannot formally exclude that such mechanisms contribute to the observed phenotypes (Ziv *et al*, 2006; Quenneville *et al*, 2011; Bunch *et al*, 2014). Still, aberrant ERV expression in settings which do not involve the loss of Trim28 has also indicated that ERV activation in the brain results in an inflammatory response as well behavioral impairments that are reminiscent to those observed in Trim28-mutant mice (Li *et al*, 2015; Kremer *et al*, 2019; Sankowski *et al*, 2019; Dembny *et al*, 2020; Jonsson *et al*, 2020).

In the current study, we found no evidence that activation of ERVs in NPCs, *in vitro* or *in vivo*, triggers an interferon response. This was unexpected, since a number of previous studies have demonstrated that activation of ERVs and other TEs through pharmacological inhibition of DNMTs, aging or by genetic alterations results in an interferon response through the recognition of double-stranded RNA or other TE-derived nucleic acids (Van Meter *et al*, 2014; Lim *et al*, 2015; Roulois *et al*, 2015; Thomas *et al*, 2017; Ahmad *et al*, 2018; De Cecco *et al*, 2019). On the other hand, there are mouse models, such as mice deficient for Toll-like receptors or antibodies (Young *et al*, 2012; Yu *et al*, 2012), that express high levels of ERVs without the appearance of an interferon response. Thus, the link between TE-activation and interferon response is likely context dependent, both in regard to the cell and tissue type as well as the identity of the TEs. Our results on ERV activation in the brain rather point to an alternative mechanism. One possibility is that ERV-derived peptides and proteins are implicated in the observed inflammatory response. We found numerous neurons displaying expression of ERV-derived proteins in the transgenic mice that expressed high levels of ERV-derived transcripts. These proteins were not distributed in a uniform pattern, but rather tended to form aggregate-like structures in a subset of neurons often located in close vicinity. This observation is interesting given the well-established link between protein aggregation and neurodegenerative disorders (Ross & Poirier, 2004). It is plausible that the presence of

© 2021 The Authors

ERV proteins directly or indirectly causes an inflammatory response, or they may serve as a trigger for further protein aggregation. Future in-depth studies are needed to understand this phenomenon.

In summary, these results demonstrate that Trim28 is required to silence ERVs during brain development and that the perturbation of this system results in an ERV-mediated inflammatory response in the adult brain. These results provide a new perspective to the potential cause and progression of neurodevelopmental and neurodegenerative disorders and further research into ERV-dysregulation in these types of disorders is therefore warranted.

Materials and Methods

Generation of Cas9-GFP mouse NPC cultures

All animal-related procedures were approved by and conducted in accordance with the committee for use of laboratory animals at Lund University.

The forebrain was dissected on embryonic day 13.5 from embryos obtained by breeding homozygote Cas9-GFP knock-in mice (Platt *et al*, 2014). The tissue was mechanically dissociated and plated in gelatin coated flasks and maintained as a monolayer culture (Conti *et al*, 2005) in NSA medium (Euromed, Euroclone) supplemented with N2 hormone mix, EGF (20 ng/ml; Gibco), bFGF (20 ng/ml; Gibco), 2 nM L-glutamine and 100 μg/ml Pen/Strep. Cells were then passaged 1:3–1:6 every 2–3 days using Accutase (Gibco).

Targeting *Trim28 in vitro*

Guides were designed at crispr.mit.edu and are listed in the Appendix. Lentiviruses were produced according to Zufferey *et al*, (1997), and titers were 10^9 TU/ml, which was determined using qRT–PCR. Cas9 mouse NPCs were transduced at a MOI 40 and allowed to expand for 10 days prior to FACS (FACSAria, BD Biosciences). Cells were detached and resuspended in basic culture media (media excluding growth factors) with propidium iodide (BD Biosciences) and strained (70 μm filters, BD Biosciences). RFP cells were FACS isolated at 4°C (reanalysis showed > 99% purity) and pelleted at 400 *g* for 5 min, snap frozen on dry ice and stored at −80°C until RNA/DNA were isolated. All groups were performed in biological triplicates.

Targeting *Trim28 in vivo* using CRISPR/Cas9 in the adult brain

All animal-related procedures were approved by and conducted in accordance with the committee for use of laboratory animals at Lund University.

The production of AAV5 vectors has been described in detail elsewhere (Ulusoy *et al*, 2009), and titers were in the order of 10^{13} TU/ml, which was determined by qRT–PCR using TaqMan primers toward the ITR. Prior to injection, the vectors were diluted in PBS; the vectors containing the guide RNAs were diluted to 30% except upon co-injection of guides 3, 4, and 13 where the vectors were diluted to 10% each. Rosa26 Cas9 knock-in mice were anesthetized by isoflurane prior to the intra-striatal injections (coordinates from bregma: AP + 0.9 mm, ML + 1.8 mm, DV −2.7 mm) of 1 μl virus solution (0.1 μl / 15 s). The needle was kept in place for additional 2 min post-injection to avoid backflow. Animals were sacrificed after 2 months and analyzed either by IHC or nuclei isolation (see details below) followed by DNA- or RNA-seq.

Targeting *Trim28 in vivo* during neural development

Male *Emx1*-Cre (+/−); *Trim28*-flox (+/−); gtRosa (+/−) were bred with *Trim28*-flox (+/+) females to generate animals in which one (*Emx1*-Cre +/−; *Trim28*-flox +/−) or both (*Emx1*-Cre +/−; *Trim28*-flox +/+) *Trim28* alleles had been excised, used as control and *Trim28*-KO, respectively. Animals used for IHC were additionally heterozygote for gtRosa, in order to visualize the cells in which Cre had been expressed. Animals were genotyped from tail biopsies according to previous protocols (Cammas *et al*, 2000) and sacrificed at 3 months of age for either IHC or RNA-sequencing.

Immunohistochemistry

Mice were given a lethal dose of phenobarbiturate and transcardially perfused with 4% paraformaldehyde (PFA, Sigma); the brains were post-fixed for 2 h and transferred to phosphate buffered saline (PBS) with 25% sucrose. Brains were coronally sectioned on a microtome (30 μm) and put in KPBS. IHC was performed as described in detail elsewhere (Sachdeva *et al*, 2010). Antibodies: Trim28 (Millipore, MAB3662, 1:500), Trim28 (Abcam, ab10484, 1:1,000), NeuN (Sigma-Aldrich, MAB377, 1:1,000), IAP-Gag (a kind gift from Bryan Cullen and described in (Dewannieux *et al*, 2004), 1:2,000), Iba1 (WAKO, no.019-19741, 1:1,000). All sections were counterstained with 4',6-diamidino-2-phenylindole (DAPI, Sigma-Aldrich, 1:1,000). Secondary antibodies from Jackson Laboratories were used at 1:400.

Nuclei isolation

Animals were sacrificed by cervical dislocation and brains quickly removed. The desired regions were dissected and snap frozen on dry ice and stored at −80°C. The nuclei isolation was performed according to (Sodersten *et al*, 2018). In brief, the tissue was thawed and dissociated in ice-cold lysis buffer (0.32 M sucrose, 5 mM $CaCl_2$, 3 mM MgAc, 0.1 mM Na_2EDTA, 10 mM Tris–HCl, pH 8.0, 1 mM DTT) using a 1 ml tissue douncer (Wheaton). The homogenate was carefully layered on top of a sucrose cushion (1.8 M sucrose, 3 mM MgAc, 10 mM Tris–HCl, pH 8.0, and 1 mM DTT) before centrifugation at 30,000 ×*g* for 2 h and 15 min. Pelleted nuclei were softened for 10 min in 100 μl of nuclear storage buffer (15% sucrose, 10 mM Tris–HCl, pH 7.2, 70 mM KCl, and 2 mM $MgCl_2$) before resuspended in 300 μl of dilution buffer (10 mM Tris–HCl, pH 7.2, 70 mM KCl, and 2 mM $MgCl_2$) and run through a cell strainer (70 μm). Cells were run through the FACS (FACS Aria, BD Biosciences) at 4°C with low flowrate using a 100 μm nozzle (reanalysis showed > 99% purity). Sorted nuclei intended for either DNA or RNA-sequencing were pelleted at 1,300 ×*g* for 15 min and snap frozen, while nuclei intended for single-nuclei RNA-sequencing were directly loaded onto the 10× Genomics Single Cell 3′ Chip—see *Single-nuclei sequencing*.

Analysis of CRISPR/Cas9-mediated *Trim28*-indels

Total genomic DNA was extracted from all *Trim28*-KO and control groups using DNeasy blood and tissue kit (Qiagen) and a 1.5 kb

© 2021 The Authors

fragment surrounding the different target sequences were amplified by PCR (see Table EV3 and EV4 for target and primer sequences, respectively) before subjected to NexteraXT fragmentation, according to manufacturer recommendations. Indexed tagmentation libraries were sequenced with 2 × 150 bp PE reads and analyzed using an in-house TIGERq pipeline to evaluate CRISPR/Cas9 editing efficiency.

RNA-sequencing

Total RNA was isolated from frozen cell/nuclei pellets and brain tissue using the RNeasy Mini Kit (Qiagen) and used for RNA-seq (tissue pieces were run in the tissue lyser for 2 min, 30 Hz, prior to RNA isolation). Libraries were generated using Illumina TruSeq Stranded mRNA library prep kit (poly-A selection) and sequenced on a NextSeq500 (PE 2 × 150 bp).

The reads were mapped with STAR (2.6.0c) (Dobin *et al*, 2013), using gencode mouse annotation GRCm38.p6 vM19 as a guide. Reads were allowed to map to 100 loci with 200 anchors, as recommended by (Jin *et al*, 2015) to run TEtranscripts.

Read quantification was performed with TEtranscripts version 2.0.3 in multimode using gencode annotation GRCm38.p6 vM19 for gene annotation, as well as the curated GTF file of TEs provided by TEtranscripts authors (Jin *et al*, 2015). This file differs to RepeatMasker as it excludes simple repeats, rRNAs, scRNAs, snRNAs, srpRNAs, and tRNAs. The output matrix was then divided between TE subfamilies and genes to perform differential expression analysis (DEA) with DESeq2 (version 1.22.2) (Love *et al*, 2014) contrasting *Trim28*-KO against control samples. DESeq2 creates a general linear model assuming a negative binomial distribution using the condition of a sample and the normalized values of a gene. The resulting coefficients are tested between conditions using a Wald test. *P* values are then adjusted using Benjamini and Hochberg correction. For more information about the package methods, see (Love *et al*, 2014).

We report TE subfamilies as significantly different if their *P* adjusted value is below 0.05 and the absolute value of its log2 fold change is higher than 0.5.

To show the expression levels per condition, samples from the different guides targeting *Trim28* were pooled together and tested against the LacZ controls. The data were normalized using sizeFactors from the DESeq2 object (median ratio method described in (Anders & Huber, 2010) to account for any differences in sequencing depth.

In order to define differentially expressed elements and study their effects on gene expression, reads were uniquely mapped with STAR (2.6.0c). Full length mouse ERV predictions were done using the RetroTector software (Sperber *et al*, 2007), and read quantification of them was performed using featureCounts (Subread 1.6.3) (Liao *et al*, 2014). Differential expression analysis (DEA) was done with DESeq2. An intersection of the gencode annotation GRCm38.p6 vM19 with windows of 10, 20, and 50 kbp up and downstream of the upregulated elements was made with BEDtools intersect (Quinlan & Hall, 2010); same intersection was done for non-upregulated elements to compare their nearby gene dysregulation.

Bigwig files were normalized by RPKM using bamCoverage from deeptools and uploaded to USCS Genome Browser (release GCF_000001635.25_GRCm38.p5 (2017-08-04)).

Differential gene expression analysis was performed using DESeq2. Up- and downregulated genes (*P*-adj < 0.05, log2FC > 0.5) were used to test for GO terms overrepresentation using the web-based tool PANTHER (Mi *et al*, 2019). 30407594 We tested for overrepresentation of terms in their GO-Slim biological process dataset using Fisher's exact test with false discovery rates. Terms shown in main figures were those with more than four genes among the group of genes we were testing (up or downregulated), with an absolute log2 fold change value higher than 0.5 and a false discovery rate less than 0.05.

Single-nuclei RNA-sequencing

Nuclei were isolated from the cortex of *Emx1*-Cre(+/−)/*Trim28*-flox (+/−, +/+) animals (Ctl $n = 2$, KO $n = 2$) as described above. 8,500 nuclei per sample were sorted via FACS and loaded onto 10× Genomics Single Cell 3′ Chip along with the Reverse Transcription Master Mix following the manufacturer's protocol for the Chromium Single Cell 3′ Library (10× Genomics, PN-120233) to generate single cell gel beads in emulsion. cDNA amplification was done as per the guidelines from 10× Genomics, and sequencing libraries were generated with unique sample indices (SI) for each sample. Libraries for samples were multiplexed and sequenced on a Novaseq using a 150-cycle kit.

The raw base calls were demultiplexed and converted to sample specific fastq files using cellranger mkfastq[1] that uses bcl2fastq program provided by Illumina. The default setting for bcl2fastq program was used, allowing 1 mismatch in the index, and raw quality of reads was checked using FastQC and multiQC tools. For each sample, fastq files were processed independently using cellranger count version 3.0 pipeline (default settings). This pipeline uses splice-aware program STAR[5] to map cDNA reads to the transcriptome (mm10). Since in nuclei samples it is expected to get a higher fraction of pre-mRNA, a pre-mRNA reference was generated using cellranger guidelines.

Mapped reads were characterized into exonic, intronic, and intergenic if at least 50% of the read intersects with an exon, intronic if it is non-exonic and it intersects with an intron and intergenic otherwise. Only exonic reads that uniquely mapped to transcriptome (and the same strand) were used for the downstream analysis.

Low-quality cells and genes were filtered out based on fraction of total number of genes detected in each cell (±3 nmads). From the remaining 16,671 nuclei, 6,472 came from control samples (Ctl) and 7,199 from knockout (KO).

For downstream analysis, samples were merged together using Seurat (version 3) R package (Dobin *et al*, 2013). Clusters have been defined with Seurat function FindClusters using resolution 0.1 and visualized with UMAP plots. Cell type annotation was performed using both known marker-based expression per cluster and a comparison of the expression profiles of a mouse brain Atlas (Zeisel *et al*, 2018). A marker gene set consisting of upregulated gene per cluster among the cells, combined with marker genes for all the 256 cell types in the atlas, was used in the comparison. The 256 atlas cell types were grouped into main clusters at Taxonomy rank 4 (39 groups), and mean expression per group was calculated using the marker gene set. These were compared to the mean expression in our clusters using Spearman correlation. Based on clusters annotation, clusters 0, 1, 2, 5, and 6, 7 were manually merged as excitatory

© 2021 The Authors

and inhibitory neurons, respectively. For each cell type, differential expression between knockout and control samples was carried out using Seurat function FindMarkers (Wilcox test, P adjusted < 0.01).

The expression of transposable elements was analyzed by extracting cell barcodes for all clusters using Seurat function WhichCells, and the original.bam files obtained from the cellranger pipeline were used to subset aligned files for each cluster (subset-bam tool provided by 10×). Each.bam file was then converted back to clusters' fastq files using bamtofastq tool from 10×.

The resulting fastq files were mapped using default parameters in STAR using gencode mouse annotation GRCm38.p6 vM19 as a guide. The resulting bam files were used to quantify reads mapping to genes with featureCounts (forward strandness). The output matrix was then used to calculate sizeFactors with DESeq2 that would later be used to normalize TE counts.

The cluster fastq files were also mapped allowing for 100 loci and 200 anchors, as recommended by TEtranscripts authors. Read quantification was then performed with TEtranscripts in multimode (forward strand) using GRCm38.p6 vM19 for the gene annotation, and a curated GTF file of TEs given by TEtranscripts' authors. For further details, see the RNA-sequencing paragraph.

For the data presented in Fig 5B, the fold change bar plots were made from a DEA performed with DESeq2 of TE subfamilies of each cell type comparing control and knockout samples, for further details see the RNA-sequencing paragraph. The mean plots in the same figure were normalized using the sizeFactors resulting from the gene quantification with the default parameters' mapping.

CUT&RUN

The CUT&RUN were performed according to (Skene & Henikoff, 2017). In brief, 200,000 mouse NPCs were washed, permeabilized, and attached to ConA-coated magnetic beads (Bang Laboratories) before incubated with the H3K9me3 (1:50, ab8898, Abcam) antibody at 4°C overnight. Cells were washed and incubated with pA-MNase fusion protein, and digestion was activated by adding $CaCl_2$ at 0°C. The digestion was stopped after 30 min and the target chromatin released from the insoluble nuclear chromatin before extracting the DNA. Experimental success was evaluated by capillary electrophoresis (Agilent) and the presence of nucleosome ladders for H3K9me3 but not for IgG controls.

The library preparation was performed using the Hyper prep kit (KAPA biosystems) and sequenced on NextSeq500 2 × 75 bp. Mapping of the reads to mm10 was performed with Bowtie2 2.3.4.2 (Langmead & Salzberg, 2012) using default settings for local alignment. Multi-mapper reads were filtered by SAMtools view version 1.4 (Li *et al*, 2009).

Using the ERVK prediction described in the section *RNA-sequencing*, we retrieved full length MMERVK10Cs. An ERVK was considered to be a full length MMERVK10C when an annotated MMERVK10C-int of mm10 RepeatMasker annotation (open-4.0.5—Repeat Library 20140131) would overlap more than 50% into the full length ERVK prediction. The intersection was performed with BEDtools intersect 2.26.0 (-f 0.5) (Quinlan & Hall, 2010). Heatmaps and profile plots were produced using deepTools' plotHeatmap (Ramirez *et al*, 2016) and sorted using maximum expression of the *Trim28*-KO samples or guide 3 for the *in vitro* and *in vivo* CRISPRs. Tracks for genome browser were normalized using RPKM using deepTools' bamCoverage (version 2.4.3) (Ramirez *et al*, 2016).

The H3K9me3 ChIP-seq data from adult mouse cortex were retrieved from (Jiang *et al*, 2017), mapped, and analyzed in the same way as the in-house Cut & Run samples described above.

qRT–PCR

Cortical brain pieces were disrupted in a tissue lyser (2 min, 30 Hz, 4°C) prior to RNA isolation using an RNeasy Mini Kit (Qiagen). cDNA was synthesized by the Maxima First-Strand cDNA Synthesis Kit (Invitrogen) and analyzed with SYBR Green I master (Roche) on a LightCycler 480 (Roche). Data are represented with the ΔΔCt method normalized to the housekeeping genes *Gapdh* and β-*actin*. Primers are listed in Table EV4.

Western blot

Dissected cortical pieces from the *Emx1*-Cre (+/−); *Trim28*-flox (+/− and +/+) animals (Ctl $n = 5$, KO $n = 5$) were put in RIPA buffer (Sigma-Aldrich) containing Protease inhibitor cocktail (PIC, Complete, 1:25) and lysed at 4°C using a TissueLyser LT (Qiagen) on 50 Hz for 2 min, twice, and then kept on ice for 30 min before spun at 10,000 ×g for 10 min at 4°C. Supernatants were collected and transferred to a new tube and stored at −20°C. Each sample was mixed 1:1 (10 µl + 10 µl) with Laemmli buffer (Bio-Rad) and boiled at 95°C for 5 min before loaded onto a 4–12% SDS–PAGE gel and run at 200 V before electrotransferred to a membrane using Transblot-Turbo Transfer system (Bio-Rad). The membrane was then washed 2 × 15 min in TBS with 0.1% Tween20 (TBST), blocked for 1 h in TBST containing 5% non-fat dry milk, and then incubated at 4°C overnight with the primary antibody diluted in TBST with 5% non-fat dry milk (rabbit anti-Trim28, Abcam ab10484, 1:1,000; rabbit anti-CD68, 1:1,000, Abcam ab125212; rabbit anti-IAP-Gag, 1:10,000, a kind gift from Bryan Cullen and described in (Dewannieux *et al*, 2004)). The membrane was washed in TBST 2 × 15 min and incubated for 1 h in room temperature with HRP-conjugated anti-rabbit antibody (Sigma-Aldrich, NA9043, 1:2,500) diluted in TBST with 5% non-fat dry milk. The membrane was washed 2 × 15 min in TBST again and 1 × 15 min in TBS, before the protein expression was revealed by chemiluminescence using Immobilon Western (Millipore) and the signal detected using a ChemiDoc MP system (Bio-Rad). The membrane was stripped by treating it with methanol for 15 s followed 15 min in TBST before incubating it in stripping buffer (100 mM 2-mercaptoethanol, 2% (w/v) SDS, 62.4 mM Tris–HCL, pH 6.8) for 30 min 50°C. The membrane was washed in running water for 15 min, followed by 3 × 15 min in TBST before blocked for 1 h in TBST containing 5% non-fat dry milk. The membrane was then stained and visualized for β-actin (mouse anti-β-actin HRP, Sigma-Aldrich, A3854, 1:50,000) as described above.

Morphological analysis

The morphology of Iba1$^+$ cells in the *Emx1*-Cre/*Trim28*-flox animals (Ctl $n = 3$, KO $n = 2$) was analyzed in 2D through an unbiased, automated process using the Cellomics Array Scan (Array Scan VTI, Thermo Fischer). The scanner took a high number of photographs

© 2021 The Authors

(using a 20× objective) throughout cortex (Ctl n = 361, KO n = 215) and striatum (Ctl n = 104, KO n = 63) and the program "Neuronal profiling" allowed analysis of process length, process area, and branchpoints per cell. 10 photographs of cortex from each animal were randomly selected, and Iba1$^+$ cells were manually counted in a blinded manner and presented as Iba1$^+$ cells per mm^2.

Code availability

The pipeline, configuration files, and downstream analyses are available in the src folder at GitHub (https://github.com/ra7555gas/trim28_Jonsson2020.git). All downstream analysis and visualization were performed in R 3.5.1.

Data availability

There are no restrictions in data availability. All file names are described in Table EV5, and the accession code for the RNA and DNA sequencing data presented in this study is GSE154196.

Expanded View for this article is available online.

Acknowledgements

We would like to thank Molly Gale Hammell, Magdalena Götz, Sten Linnarsson, Chris Douse, Florence Cammas and Bryan Cullen for providing valuable reagents and input on the manuscript. We also thank, M. Persson Vejgården, U. Jarl, and A. Hammarberg for technical assistance. We are grateful to all members of the Jakobsson laboratory. The work was supported by grants from the Swedish Research Council (2018-02694, JJ & 2018-03017, PJ), the Swedish Brain Foundation (FO2019-0098, JJ), Cancerfonden (190326, JJ), Barncancerfonden (PR2017-0053, JJ), Formas (2018-01008, PJ) and the Swedish Government Initiative for Strategic Research Areas (MultiPark & StemTherapy).

Author contributions

All authors took part in designing the study as well as interpreting the data. MEJ and JJ conceived and designed the study. MEJ, RP, JGJ, PAJ, DAMA, KP, SM, DY, RR performed experimental research and RG, PJ and YS performed bioinformatical analyses, JL ES, JH contributed resources. MEJ, RG, and JJ wrote the manuscript and all authors reviewed the final version.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Ahmad S, Mu X, Yang F, Greenwald E, Park JW, Jacob E, Zhang CZ, Hur S (2018) Breaching self-tolerance to alu duplex RNA underlies MDA5-mediated inflammation. *Cell* 172: 797 – 810.e13

Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11: R106

Andrews WD, Tuke PW, Al-Chalabi A, Gaudin P, Ijaz S, Parton MJ, Garson JA (2000) Detection of reverse transcriptase activity in the serum of patients with motor neurone disease. *J Med Virol* 61: 527 – 532

Antony JM, van Marle G, Opii W, Butterfield DA, Mallet F, Yong VW, Wallace JL, Deacon RM, Warren K, Power C (2004) Human endogenous retrovirus glycoprotein-mediated induction of redox reactants causes oligodendrocyte death and demyelination. *Nat Neurosci* 7: 1088 – 1095

Bale TL, Baram TZ, Brown AS, Goldstein JM, Insel TR, McCarthy MM, Nemeroff CB, Reyes TM, Simerly RB, Susser ES *et al* (2010) Early life programming and neurodevelopmental disorders. *Biol Psychiatry* 68: 314 – 319

Bauer ME, Teixeira AL (2019) Inflammation in psychiatric disorders: what comes first? *Ann N Y Acad Sci* 1437: 57 – 67

Berko ER, Cho MT, Eng C, Shao Y, Sweetser DA, Waxler J, Robin NH, Brewer F, Donkervoort S, Mohassel P *et al* (2017) De novo missense variants in HECW2 are associated with neurodevelopmental delay and hypotonia. *J Med Genet* 54: 84 – 86

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 Rep* 18: 1 – 11

Bunch H, Zheng X, Burkholder A, Dillon ST, Motola S, Birrane G, Ebmeier CC, Levine S, Fargo D, Hu G *et al* (2014) TRIM28 regulates RNA polymerase II promoter-proximal pausing and pause release. *Nat Struct Mol Biol* 21: 876 – 883

Cammas F, Mark M, Dolle P, Dierich A, Chambon P, Losson R (2000) Mice lacking the transcriptional corepressor TIF1beta are defective in early postimplantation development. *Development* 127: 2955 – 2963

Chuong EB, Elde NC, Feschotte C (2017) Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* 18: 71 – 86

Conti L, Pollard SM, Gorba T, Reitano E, Toselli M, Biella G, Sun Y, Sanzone S, Ying QL, Cattaneo E *et al* (2005) Niche-independent symmetrical self-renewal of a mammalian tissue stem cell. *PLoS Biol* 3: e283

De Cecco M, Ito T, Petrashen AP, Elias AE, Skvir NJ, Criscione SW, Caligiana A, Brocculi G, Adney EM, Boeke JD *et al* (2019) L1 drives IFN in senescent cells and promotes age-associated inflammation. *Nature* 566: 73 – 78

Dembny P, Newman AG, Singh M, Hinz M, Szczepek M, Kruger C, Adalbert R, Dzaye O, Trimbuch T, Wallach T *et al* (2020) Human endogenous retrovirus HERV-K(HML-2) RNA causes neurodegeneration through Toll-like receptors. *JCI Insight* 5: e131093

Dewannieux M, Dupressoir A, Harper F, Pierron G, Heidmann T (2004) Identification of autonomous IAP LTR retrotransposons mobile in mammalian cells. *Nat Genet* 36: 534 – 539

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: 15 – 21

Douville R, Liu J, Rothstein J, Nath A (2011) Identification of active loci of a human endogenous retrovirus in neurons of patients with amyotrophic lateral sclerosis. *Ann Neurol* 69: 141 – 151

Fasching L, Kapopoulou A, Sachdeva R, Petri R, Jonsson ME, Manne C, Turelli P, Jern P, Cammas F, Trono D *et al* (2015) TRIM28 represses transcription of endogenous retroviruses in neural progenitor cells. *Cell Rep* 10: 20 – 28

Garson JA, Tuke PW, Giraud P, Paranhos-Baccala G, Perron H (1998) Detection of virion-associated MSRV-RNA in serum of patients with multiple sclerosis. *Lancet* 351: 33

Guissart C, Latypova X, Rollier P, Khan TN, Stamberger H, McWalter K, Cho MT, Kjaergaard S, Weckhuysen S, Lesca G *et al* (2018) Dual molecular effects of dominant RORA mutations cause two variants of syndromic intellectual disability with either autism or cerebellar ataxia. *Am J Hum Genet* 102: 744 – 759

Guo C, Jeong HH, Hsieh YC, Klein HU, Bennett DA, De Jager PL, Liu Z, Shulman JM (2018) Tau activates transposable elements in Alzheimer's disease. *Cell Rep* 23: 2874 – 2880

© 2021 The Authors

Horwitz T, Lam K, Chen Y, Xia Y, Liu C (2019) A decade in psychiatric GWAS research. *Mol Psychiatry* 24: 378 – 389

Hurst TP, Magiorkinis G (2015) Activation of the innate immune response by endogenous retroviruses. *J Gen Virol* 96: 1207 – 1218

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

Ishak CA, Classon M, De Carvalho DD (2018) Deregulation of retroelements as an emerging therapeutic opportunity in cancer. *Trends Cancer* 4: 583 – 597

Iwasato T, Datwani A, Wolf AM, Nishiyama H, Taguchi Y, Tonegawa S, Knopfel T, Erzurumlu RS, Itohara S (2000) Cortex-restricted disruption of NMDAR1 impairs neuronal patterns in the barrel cortex. *Nature* 406: 726 – 731

Jakobsson J, Cordero MI, Bisaz R, Groner AC, Busskamp V, Bensadoun JC, Cammas F, Losson R, Mansuy IM, Sandi C *et al* (2008) KAP1-mediated epigenetic repression in the forebrain modulates behavioral vulnerability to stress. *Neuron* 60: 818 – 831

Jern P, Coffin JM (2008) Effects of retroviruses on host genome function. *Annu Rev Genet* 42: 709 – 732

Jiang Y, Loh YE, Rajarajan P, Hirayama T, Liao W, Kassim BS, Javidfar B, Hartley BJ, Kleofas L, Park RB *et al* (2017) The methyltransferase SETDB1 regulates a large neuron-specific topological chromatin domain. *Nat Genet* 49: 1239 – 1250

Jin Y, Tam OH, Paniagua E, Hammell M (2015) TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* 31: 3593 – 3599

Jonsson ME, Garza R, Johansson PA, Jakobsson J (2020) Transposable elements: a common feature of neurodevelopmental and neurodegenerative disorders. *Trends Genet* 36: 610 – 623

Jonsson ME, Ludvik Brattas P, Gustafsson C, Petri R, Yudovich D, Pircs K, Verschuere S, Madsen S, Hansson J, Larsson J *et al* (2019) Activation of neuronal genes via LINE-1 elements upon global DNA demethylation in human neural progenitors. *Nat Commun* 10: 3182

Kao WT, Wang Y, Kleinman JE, Lipska BK, Hyde TM, Weinberger DR, Law AJ (2010) Common genetic variation in Neuregulin 3 (NRG3) influences risk for schizophrenia and impacts NRG3 expression in human brain. *Proc Natl Acad Sci USA* 107: 15619 – 15624

Karlsson H, Bachmann S, Schroder J, McArthur J, Torrey EF, Yolken RH (2001) Retroviral RNA identified in the cerebrospinal fluids and brains of individuals with schizophrenia. *Proc Natl Acad Sci USA* 98: 4634 – 4639

Khashan AS, Abel KM, McNamee R, Pedersen MG, Webb RT, Baker PN, Kenny LC, Mortensen PB (2008) Higher risk of offspring schizophrenia following antenatal maternal exposure to severe adverse life events. *Arch Gen Psychiatry* 65: 146 – 152

Kremer D, Gruchot J, Weyers V, Oldemeier L, Gottle P, Healy L, Ho Jang J, Kang TXY, Volsko C, Dutta R *et al* (2019) pHERV-W envelope protein fuels microglial cell-dependent damage of myelinated axons in multiple sclerosis. *Proc Natl Acad Sci USA* 116: 15216 – 15225

Krug L, Chatterjee N, Borges-Monroy R, Hearn S, Liao WW, Morrill K, Prazak L, Rozhkov N, Theodorou D, Hammell M *et al* (2017) Retrotransposon activation contributes to neurodegeneration in a *Drosophila* TDP-43 model of ALS. *PLoS Genet* 13: e1006635

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9: 357 – 359

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078 – 2079

Li W, Jin Y, Prazak L, Hammell M, Dubnau J (2012) Transposable elements in TDP-43-mediated neurodegenerative disorders. *PLoS One* 7: e44099

Li W, Lee MH, Henderson L, Tyagi R, Bachani M, Steiner J, Campanac E, Hoffman DA, von Geldern G, Johnson K *et al* (2015) Human endogenous retrovirus-K contributes to motor neuron disease. *Sci Transl Med* 7: 307ra153

Li W, Prazak L, Chatterjee N, Gruninger S, Krug L, Theodorou D, Dubnau J (2013) Activation of transposable elements during aging and neuronal decline in *Drosophila*. *Nat Neurosci* 16: 529 – 531

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

Lim YW, Sanz LA, Xu X, Hartono SR, Chedin F (2015) Genome-wide DNA hypomethylation and RNA:DNA hybrid accumulation in Aicardi-Goutieres syndrome. *Elife* 4: e08007

Liu N, Lee CH, Swigut T, Grow E, Gu B, Bassik MC, Wysocka J (2018) Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators. *Nature* 553: 228 – 232

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

MacGowan DJ, Scelsa SN, Imperato TE, Liu KN, Baron P, Polsky B (2007) A controlled study of reverse transcriptase in serum and CSF of HIV-negative patients with ALS. *Neurology* 68: 1944 – 1946

Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD (2019) PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res* 47: D419 – D426

Muhoberac BB, Vidal R (2019) Iron, ferritin, hereditary ferritinopathy, and neurodegeneration. *Front Neurosci* 13: 1195

Naseer MI, Chaudhary AG, Rasool M, Kalamegam G, Ashgan FT, Assidi M, Ahmed F, Ansari SA, Zaidi SK, Jan MM *et al* (2016) Copy number variations in Saudi family with intellectual disability and epilepsy. *BMC Genom* 17: 757

Oksenberg N, Ahituv N (2013) The role of AUTS2 in neurodevelopment and human evolution. *Trends Genet* 29: 600 – 608

Perron H, Garson JA, Bedin F, Beseme F, Paranhos-Baccala G, Komurian-Pradel F, Mallet F, Tuke PW, Voisset C, Blond JL *et al* (1997) Molecular identification of a novel retrovirus repeatedly isolated from patients with multiple sclerosis. The Collaborative Research Group on Multiple Sclerosis. *Proc Natl Acad Sci U S A* 94: 7583 – 7588

Perron H, Mekaoui L, Bernard C, Veas F, Stefas I, Leboyer M (2008) Endogenous retrovirus type W GAG and envelope protein antigenemia in serum of schizophrenic patients. *Biol Psychiatry* 64: 1019 – 1023

Piovani G, Savio G, Traversa M, Pilotta A, De Petro G, Barlati S, Magri C (2014) De novo 1Mb interstitial deletion of 8p22 in a patient with slight mental retardation and speech delay. *Mol Cytogenet* 7: 25

Platt RJ, Chen S, Zhou Y, Yim MJ, Swiech L, Kempton HR, Dahlman JE, Parnas O, Eisenhaure TM, Jovanovic M *et al* (2014) CRISPR-Cas9 knockin mice for genome editing and cancer modeling. *Cell* 159: 440 – 455

Quenneville S, Verde G, Corsinotti A, Kapopoulou A, Jakobsson J, Offner S, Baglivo I, Pedone PV, Grimaldi G, Riccio A *et al* (2011) In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. *Mol Cell* 44: 361 – 372

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841 – 842

Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dundar F, Manke T (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* 44: W160 – 165

© 2021 The Authors

Ross CA, Poirier MA (2004) Protein aggregation and neurodegenerative disease. *Nat Med* 10(Suppl): S10 – 17

Roulois D, Loo Yau H, Singhania R, Wang Y, Danesh A, Shen SY, Han H, Liang G, Jones PA, Pugh TJ *et al* (2015) DNA-Demethylating agents target colorectal cancer cells by inducing viral mimicry by endogenous transcripts. *Cell* 162: 961 – 973

Rowe HM, Jakobsson J, Mesnard D, Rougemont J, Reynard S, Aktas T, Maillard PV, Layard-Liesching H, Verp S, Marquis J *et al* (2010) KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature* 463: 237 – 240

Sachdeva R, Jonsson ME, Nelander J, Kirkeby A, Guibentif C, Gentner B, Naldini L, Bjorklund A, Parmar M, Jakobsson J (2010) Tracking differentiating neural progenitors in pluripotent cultures using microRNA-regulated lentiviral vectors. *Proc Natl Acad Sci U S A* 107: 11602 – 11607

Saleh A, Macia A, Muotri AR (2019) Transposable elements, inflammation, and neurological disease. *Front Neurol* 10: 894

Sankowski R, Strohl JJ, Huerta TS, Nasiri E, Mazzarello AN, D'Abramo C, Cheng KF, Staszewski O, Prinz M, Huerta PT *et al* (2019) Endogenous retroviruses are associated with hippocampus-based memory impairment. *Proc Natl Acad Sci U S A* 116: 25982 – 25990

Schizophrenia Psychiatric Genome-Wide Association Study, C (2011) Genome-wide association study identifies five new schizophrenia loci. *Nat Genet* 43: 969 – 976

Skene PJ, Henikoff S (2017) An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* 6: e21856

Sodersten E, Toskas K, Rraklli V, Tiklova K, Bjorklund AK, Ringner M, Perlmann T, Holmberg J (2018) Author Correction: a comprehensive map coupling histone modifications with gene regulation in adult dopaminergic and serotonergic neurons. *Nat Commun* 9: 4639

Sperber GO, Airola T, Jern P, Blomberg J (2007) Automated recognition of retroviral sequences in genomic data–RetroTector. *Nucleic Acids Res* 35: 4964 – 4976

Sripathy SP, Stevens J, Schultz DC (2006) The KAP1 corepressor functions to coordinate the assembly of de novo HP1-demarcated microenvironments of heterochromatin required for KRAB zinc finger protein-mediated transcriptional repression. *Mol Cell Biol* 26: 8623 – 8638

Steele AJ, Al-Chalabi A, Ferrante K, Cudkowicz ME, Brown Jr RH, Garson JA (2005) Detection of serum reverse transcriptase activity in patients with ALS and unaffected blood relatives. *Neurology* 64: 454 – 458

Sun W, Samimi H, Gamez M, Zare H, Frost B (2018) Pathogenic tau-induced piRNA depletion promotes neuronal death through transposable element dysregulation in neurodegenerative tauopathies. *Nat Neurosci* 21: 1038 – 1048

Susser E, St Clair D, He L (2008) Latent effects of prenatal malnutrition on adult health: the example of schizophrenia. *Ann N Y Acad Sci* 1136: 185 – 192

Tam OH, Ostrow LW, Gale Hammell M (2019a) Diseases of the nERVous system: retrotransposon activity in neurodegenerative disease. *Mob DNA* 10: 32

Tam OH, Rozhkov NV, Shaw R, Kim D, Hubbard I, Fennessey S, Propp N, Consortium NA, Fagegaltier D, Harris BT *et al* (2019b). Postmortem cortex samples identify distinct molecular subtypes of ALS: retrotransposon activation, oxidative stress, and activated glia. *Cell Rep* 29: 1164 – 1177 e1165

Thomas CA, Tejwani L, Trujillo CA, Negraes PD, Herai RH, Mesci P, Macia A, Crow YJ, Muotri AR (2017) Modeling of TREX1-dependent autoimmune disease using human stem cells highlights L1 accumulation as a source of neuroinflammation. *Cell Stem Cell* 21: 319 – 331.e8

Ulusoy A, Sahin G, Bjorklund T, Aebischer P, Kirik D (2009) Dose optimization for long-term rAAV-mediated RNA interference in the nigrostriatal projection neurons. *Mol Ther* 17: 1574 – 1584

Van Meter M, Kashyap M, Rezazadeh S, Geneva AJ, Morello TD, Seluanov A, Gorbunova V (2014) SIRT6 represses LINE1 retrotransposons by ribosylating KAP1 but this repression fails with stress and age. *Nat Commun* 5: 5011

Whitelaw NC, Chong S, Morgan DK, Nestor C, Bruxner TJ, Ashe A, Lambley E, Meehan R, Whitelaw E (2010) Reduced levels of two modifiers of epigenetic gene silencing, Dnmt3a and Trim28, cause increased phenotypic noise. *Genome Biol* 11: R111

Wiznerowicz M, Jakobsson J, Szulc J, Liao S, Quazzola A, Beermann F, Aebischer P, Trono D (2007) The Kruppel-associated box repressor domain can trigger de novo promoter methylation during mouse early embryogenesis. *J Biol Chem* 282: 34535 – 34541

Xiao X, Zheng F, Chang H, Ma Y, Yao YG, Luo XJ, Li M (2018) The gene encoding protocadherin 9 (PCDH9), a novel risk factor for major depressive disorder. *Neuropsychopharmacology* 43: 1128 – 1137

Yoder JA, Walsh CP, Bestor TH (1997) Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet* 13: 335 – 340

Young GR, Eksmond U, Salcedo R, Alexopoulou L, Stoye JP, Kassiotis G (2012) Resurrection of endogenous retroviruses in antibody-deficient mice. *Nature* 491: 774 – 778

Yu P, Lubben W, Slomka H, Gebler J, Konert M, Cai C, Neubrandt L, Prazeres da Costa O, Paul S, Dehnert S *et al* (2012) Nucleic acid-sensing Toll-like receptors are essential for the control of endogenous retrovirus viremia and ERV-induced tumors. *Immunity* 37: 867 – 879

Zeisel A, Hochgerner H, Lonnerberg P, Johnsson A, Memic F, van der Zwan J, Haring M, Braun E, Borm LE, La Manno G *et al* (2018) Molecular architecture of the mouse nervous system. *Cell* 174: 999 – 1014.e22

Ziv Y, Bielopolski D, Galanty Y, Lukas C, Taya Y, Schultz DC, Lukas J, Bekker-Jensen S, Bartek J, Shiloh Y (2006) Chromatin relaxation in response to DNA double-strand breaks is modulated by a novel ATM- and KAP-1 dependent pathway. *Nat Cell Biol* 8: 870 – 876

Zufferey R, Nagy D, Mandel RJ, Naldini L, Trono D (1997) Multiply attenuated lentiviral vector achieves efficient gene delivery *in vivo*. *Nat Biotechnol* 15: 871 – 875

License: This is an open access article under the terms of the Creative Commons Attribution 4.0 License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

PAPER II

NEUROSCIENCE

LINE-1 retrotransposons drive human neuronal transcriptome complexity and functional diversification

Raquel Garza[1,2]†, Diahann A. M. Atacho[1,2]†, Anita Adami[1,2], Patricia Gerdes[1], Meghna Vinod[1], PingHsun Hsieh[3,4], Ofelia Karlsson[1], Vivien Horvath[1], Pia A. Johansson[1], Ninoslav Pandiloski[1,5], Jon Matas-Fuentes[5], Annelies Quaegebeur[2,6], Antonina Kouli[7], Yogita Sharma[1], Marie E. Jönsson[1], Emanuela Monni[8], Elisabet Englund[9], Evan E. Eichler[3,10], Molly Gale Hammell[2,11,12], Roger A. Barker[2,7], Zaal Kokaia[8], Christopher H. Douse[5], Johan Jakobsson[1,2]*

Copyright © 2023 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution License 4.0 (CC BY).

The genetic mechanisms underlying the expansion in size and complexity of the human brain remain poorly understood. Long interspersed nuclear element–1 (L1) retrotransposons are a source of divergent genetic information in hominoid genomes, but their importance in physiological functions and their contribution to human brain evolution are largely unknown. Using multiomics profiling, we here demonstrate that L1 promoters are dynamically active in the developing and the adult human brain. L1s generate hundreds of developmentally regulated and cell type–specific transcripts, many that are co-opted as chimeric transcripts or regulatory RNAs. One L1-derived long noncoding RNA, *LINC01876*, is a human-specific transcript expressed exclusively during brain development. CRISPR interference silencing of *LINC01876* results in reduced size of cerebral organoids and premature differentiation of neural progenitors, implicating L1s in human-specific developmental processes. In summary, our results demonstrate that L1-derived transcripts provide a previously undescribed layer of primate- and human-specific transcriptome complexity that contributes to the functional diversification of the human brain.

INTRODUCTION

During evolution, primate brains have expanded in size and complexity resulting in a unique level of cognitive functions. The genetic alterations responsible for this enhancement remain poorly understood (*1*–*4*). Our closest living relative, the chimpanzee, shares more than 98% of protein-coding sequences with humans, making it unlikely that species-specific protein-coding variants are the sole evolutionary drivers of brain complexity (*5*, *6*). Rather, a substantial fraction of the genetic basis for the differences in nonhuman primate and human brains likely resides in the noncoding part of the genome.

Transposable elements (TEs) make up at least 50% of the human genome (*7*). Since TEs have populated the genome through mobilization, this has resulted in major interspecies and interindividual differences in their genomic composition. Hundreds of thousands of TEs are primate specific, and several thousand of them are human specific (*8*, *9*). TEs pose a threat to genomic integrity—as their activation may result in retrotransposition events that cause deleterious mutations (*10*, *11*)—and the host has therefore evolved numerous mechanisms to prevent mobilization (*12*, *13*). In somatic human tissues such as the brain, it is thought that the vast majority of TEs is transcriptionally repressed, which correlates with the presence of DNA CpG methylation (*14*, *15*). However, TEs have the potential to be exapted, providing a benefit for the host as a source of gene regulatory elements and co-opted RNAs and peptides (*16*). For example, TEs are largely responsible for the emergence of species-specific long noncoding RNAs (lncRNAs) (*17*), which are untranslated transcripts of more than 200 nucleotides that have been implicated to control a wide variety of cellular processes (*18*).

The most abundant and only autonomously mobilizing TE family in humans is long interspersed nuclear element–1 (L1) (*19*). The human genome holds around half a million individual L1 copies, occupying ~17% of genomic DNA, including ancient fragments and evolutionarily younger full-length copies (*7*, *20*). Since L1s have colonized the human genome via a copy-and-paste mechanism in different waves, it is possible to approximate the evolutionary age of each individual L1 copy and assign them to chronologically ordered subfamilies (*21*). Only L1s with an intact 5′ untranslated region (UTR) allows for element-derived expression. However, most L1s are inactivated because of 5′ truncations and the accumulation of inactivating deletions and mutations. Full-length L1s are transcribed from an internal 5′ RNA polymerase II promoter as a bicistronic mRNA encoding two proteins, ORF1p and ORF2p, which are essential for L1

[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]Aligning Science Across Parkinson's (ASAP) Collaborative Research Network, Chevy Chase, MD, 20815, USA. [3]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, 98195, USA. [4]Department of Genetics, Cell Biology, and Development, University of Minnesota Medical School, Minneapolis, MN, 55455, USA. [5]Epigenetics and Chromatin Dynamics, Department of Experimental Medical Science, Wallenberg Neuroscience Center and Lund Stem Cell Center, BMC B11, Lund University, 221 84 Lund, Sweden. [6]Department of Clinical Neurosciences, University of Cambridge and Department of Pathology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [7]Department of Clinical Neuroscience and Wellcome-MRC Cambridge Stem Cell Institute, University of Cambridge, John van Geest Centre for Brain Repair, Cambridge CB2 0PY, UK. [8]Laboratory of Stem Cells and Restorative Neurology, Lund Stem Cell Center, Lund University, SE-22184 Lund, Sweden. [9]Department of Clinical Sciences Lund, Division of Pathology, Lund University, Lund, Sweden. [10]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA. [11]Institute for Systems Genetics, Department of Neuroscience and Physiology, NYU Langone Health, New York, NY 10016, USA. [12]Neuroscience Institute, NYU Grossman School of Medicine, New York, NY 10016, USA.
*Corresponding author. Email: johan.jakobsson@med.lu.se
†These authors contributed equally to this work.

Downloaded from https://www.science.org at Lund University on November 15, 2023

mobilization (*22*–*24*). Notably, the L1 promoter is bidirectional, and in evolutionarily young L1s, the antisense transcript encodes a small peptide, ORF0, with poorly characterized function (*25*, *26*). L1-antisense transcripts can also give rise to chimeric transcripts and act as alternative promoters for protein-coding genes (*14*, *26*).

Over the past two decades, L1 activity has been implicated in the functional regulation of the human brain, primarily based on the observation of somatic L1 retrotransposition events in the neural lineage leading to genomic mosaicism (*27*–*33*). However, it has been challenging to determine the functional impact of these events. Given their abundance and repetitive nature, L1s are difficult to study using standard molecular biology techniques. For

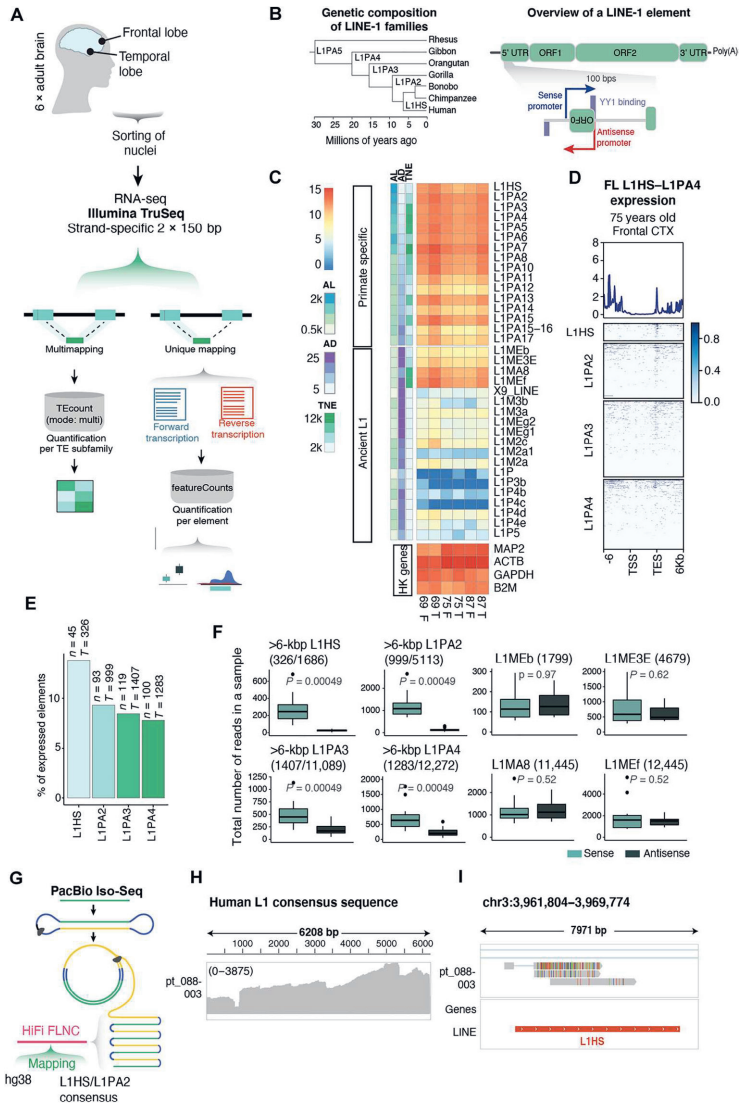


Fig. 1. L1-derived transcripts are abundant in the adult human brain. (**A**) Schematic illustrating sample collection, sequencing strategy, and bioinformatics approach. (**B**) Left: Phylogenetic tree showing the evolutionary age of young L1 subfamilies. Right: Structure of a L1 element with a zoom-in to its 5′UTR. Arrows indicate promoters in sense (blue) and antisense (red). YY1 binding sites indicated in purple boxes (sense on top and antisense on bottom). (**C**) Expression of primate-specific L1 subfamilies compared to ancient L1 subfamilies and selected housekeeping (HK) genes as reference. Row annotation showing average length (AL), average percentage of divergence from consensus (AD), and the total number of elements (TNE) (information extracted from RepeatMasker open-4.0.5). (**D**) Expression [reads per kilobase per million mapped reads (RPKM)] over full-length (>6 kbp) L1HS, L1PA2, L1PA3, and L1PA4 plus 6-kbp flanking regions. (**E**) Percentage of expressed full-length (>6 kbp) elements (mean normalized counts, >10; see Methods) among young L1 subfamilies (*n* = number of expressed elements; *T* = total number of full-length elements). (**F**) Read counts in sense (light teal) and antisense (dark teal) per sample. First four showing full-length elements in young L1 subfamilies and last four showing ancient L1 subfamilies with a comparable number of copies. (**G**) PacBio Iso-Seq schematic and mapping approach. (**H**) Coverage of PacBio Iso-Seq library mapped to L1HS and L1PA2 consensus sequence. (**I**) Genome browser tracks showing PacBio Iso-Seq reads over the promoter region of a full-length L1HS.

example, estimation of L1-derived RNA expression using quantitative polymerase chain reaction (PCR)–based techniques or standard short-read RNA sequencing (RNA-seq) approaches, whether bulk or single cell, often fails to separate L1 expression originating from the L1 promoter from that of bystander transcripts that are the result of readthrough transcription (*34*). Therefore, it is still debated whether and in which cell type L1 expression occurs in the developing and adult human brain and the impact of L1s on the physiology of the human brain remains unresolved.

In this study, we have used a combination of bulk short-read, long-read, and single-nucleus RNA-seq (snRNA-seq) coupled with cleavage under targets and release using nuclease (CUT&RUN) epigenomic profiling, together with tailored bioinformatics approaches (*35*, *36*) to demonstrate that L1-derived transcripts are highly expressed in the developing and the adult human brain. We found that the bidirectional L1 promoter is dynamically active, resulting in the generation of hundreds of L1-derived transcripts that display developmental regulation and cell type specificity. We provide evidence for the expression of full-length L1s and L1s that are co-opted as regulatory RNAs or alternative promoters. One human-specific L1-derived lncRNA (L1-lncRNA), *LINC01876*, is exclusively expressed during human brain development. CRISPR interference (CRISPRi)–based silencing of *LINC01876* results in reduced size of cerebral organoids and premature differentiation of neural progenitor cells (NPCs) and neurons, suggesting that it has an important role in brain development. Together, these results demonstrate that L1-derived transcripts are abundant in the human brain where they provide an additional layer of primate- and human-specific transcriptome complexity that may have contributed to the evolution of the human brain.

RESULTS

L1-derived transcripts are abundant in the adult human brain

To investigate the expression of L1s in the adult human brain, we obtained cortical tissue biopsies (temporal and frontal lobe) from three non-neurological deaths in people aged 69, 75, and 87 years (table S1). We sorted cell nuclei from the biopsies, extracted RNA, and used an in-house 2 × 150–base pair (bp), polyadenylate [poly (A)]-enriched stranded library preparation for bulk RNA-seq using a reduced fragmentation step to optimize library insert size for L1 analysis. These reads can be mapped uniquely and assigned to individual L1 loci, except for reads originating from a few of the youngest L1s and polymorphic L1 alleles that are not in the hg38 reference genome. We obtained ~30 million reads per sample. To quantify L1 expression, we used two different bioinformatics methodologies (Fig. 1A). First, we allowed reads to map to different locations (multimapping) and used the TEtranscripts software (*35*) in multimode to best assign these reads (fig. S1A). Second, we discarded all ambiguously mapping reads and only quantified those that map uniquely to a single location (unique mapping).

We found that L1s expressed in the adult human brain primarily belonged to primate-specific families, including both hominoid-specific (L1PA2 to L1PA4) and human-specific elements (L1HS) (Fig. 1B) (*21*). The total expression level of these subfamilies, as quantified with TEtranscripts (*35*), corresponds to expression levels of housekeeping genes (Fig. 1C). Using unique mapping, we were able to detect expression coming from hundreds of evolutionarily young L1s (Fig. 1D), including 138 full-length L1HS or L1PA2 elements (Fig. 1E). The RNA-seq signal over the full-length L1s was highly enriched at the 3′ end, which not only reflects the presence of degraded RNA in human postmortem samples and L1-mappability issues in the central part of the element but also indicates that the transcription of L1s terminates in the internal L1 polyadenylation signal (*37*). When comparing the number of reads transcribed in the same orientation as the L1s (in sense) to those in the opposite direction (in antisense), we found that most of the transcription in these regions was in sense to the L1s (Fig. 1F and fig. S1B). This suggests that most L1 transcripts originate from the L1 promoter and are not a consequence of readthrough or bystander transcription. In a few cases, we also found clear evidence of activity of the antisense L1 promoter (*26*), resulting in transcription extending out into the upstream flanking genome (fig. S1C).

To complement this analysis, we performed long-read PacBio Iso-Seq on a cortical biopsy from a deceased 84-year-old man (Fig. 1G). This allows for the identification of L1-derived transcripts that can be accurately mapped to full-length L1s and enables the identification of transcription starting sites (TSSs) and splicing events. We mapped reads [mean read length, 2.9 kilo base pairs (kbp)] to the L1HS and L1PA2 consensus sequence to which 11,120 reads mapped (of a total of 2 million reads in the library). The density of the mapped reads throughout the sequence reflected the common 5′ truncation that is present in most L1 copies in the human genome (*20*, *38*), but 1714 reads still mapped to the 5′UTR (Fig. 1H). Notably, we found several clear examples of long reads mapping to the promoter region of young full-length L1s providing further support to L1 promoter–driven expression in the adult human brain (Fig. 1I).

L1 expression is enriched in neurons in the adult human brain

To investigate the expression of L1s at cell type resolution, we performed snRNA-seq analysis using the 3′ 10x Chromium Platform and five of the adult cortical samples that we sequenced in bulk RNA-seq (Fig. 2A). In total, we sequenced 8089 high-quality nuclei with a mean of 3042 genes detected per cell. Unbiased clustering using Seurat resulted in 22 clusters (Fig. 2B), and on the basis of the expression of canonical gene markers, we identified excitatory neurons, inhibitory neurons, astrocytes, oligodendrocytes, oligodendrocyte precursors (OPC), and microglia at expected ratios (Fig. 2, C and D, and fig. S2A).

Quantification of L1 expression is challenging using single-cell technologies, as the number of mapped reads in a single cell falls short of accurate quantification, regardless of the mapping technique. To circumvent this limitation, we used an in-house bioinformatics pipeline allowing the analysis of L1 expression from the snRNA-seq dataset (Fig. 2A). This method uses the cell clusters determined on the basis of gene expression. Then, by back-tracing the reads from cells forming each cluster, it is possible to analyze the expression of L1s, using the TEtranscripts software (*35*) or with unique mapping, in distinct cell populations. This pseudo-bulk approach greatly increases the sensitivity of the TE analysis and enables quantitative estimation of L1 expression at single–cell type resolution (*36*).

We found clear evidence of L1 expression in the snRNA-seq data. Notably, L1 expression was higher in neurons, including

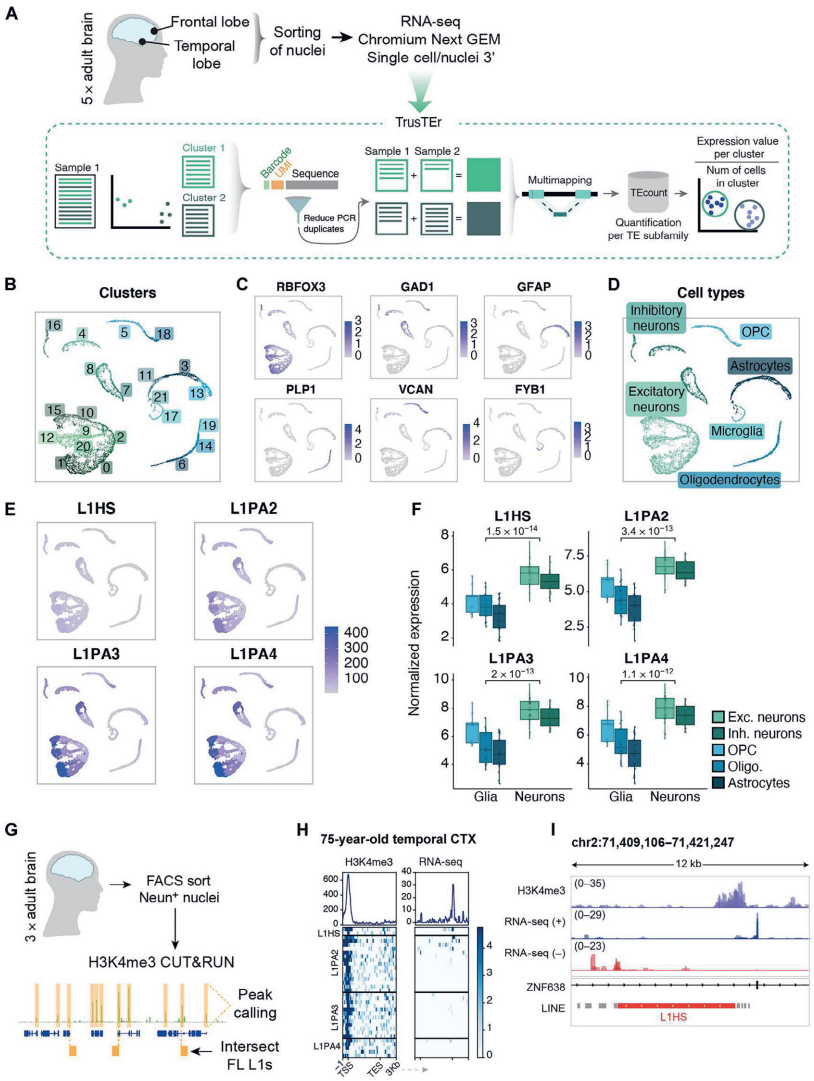


Fig. 2. L1 expression in neurons in the adult human brain. (**A**) Schematic of sample collection, sequencing approach, and analytical bioinformatics pipeline for TE expression in single-nucleus data. (**B**) snRNA-seq: Uniform Manifold Approximation and Projection (UMAP) colored by defined clusters. (**C**) Expression of selected markers for different cell types. (**D**) UMAP colored by characterized cell types. (**E**) Pseudo-bulk cluster expression of young L1 subfamilies on UMAP. OPC, oligodendrocyte precursor cell. (**F**) Comparison of glia versus neuronal clusters per L1 family (each data point corresponds to a particular cluster expression value in a sample) (*P* value as per Wilcoxon test). (**G**) Schematic of NeuN$^+$ H3K4me3 CUT&RUN in adult human brain samples and bioinformatics approach. (**H**) H3K4me3 peaks (left heatmap) over full-length L1 subfamilies (L1HS to L1PA4) and their RNA-seq signal (right heatmap). Profile plots showing sum of signal. CTX, cortex. (**I**) Genome browser tracks showing the expression of a full-length L1HS with an H3K4me3 peak on its promoter and RNA-seq signal (RPKM) split by direction of transcription (blue, forward; red, reverse).

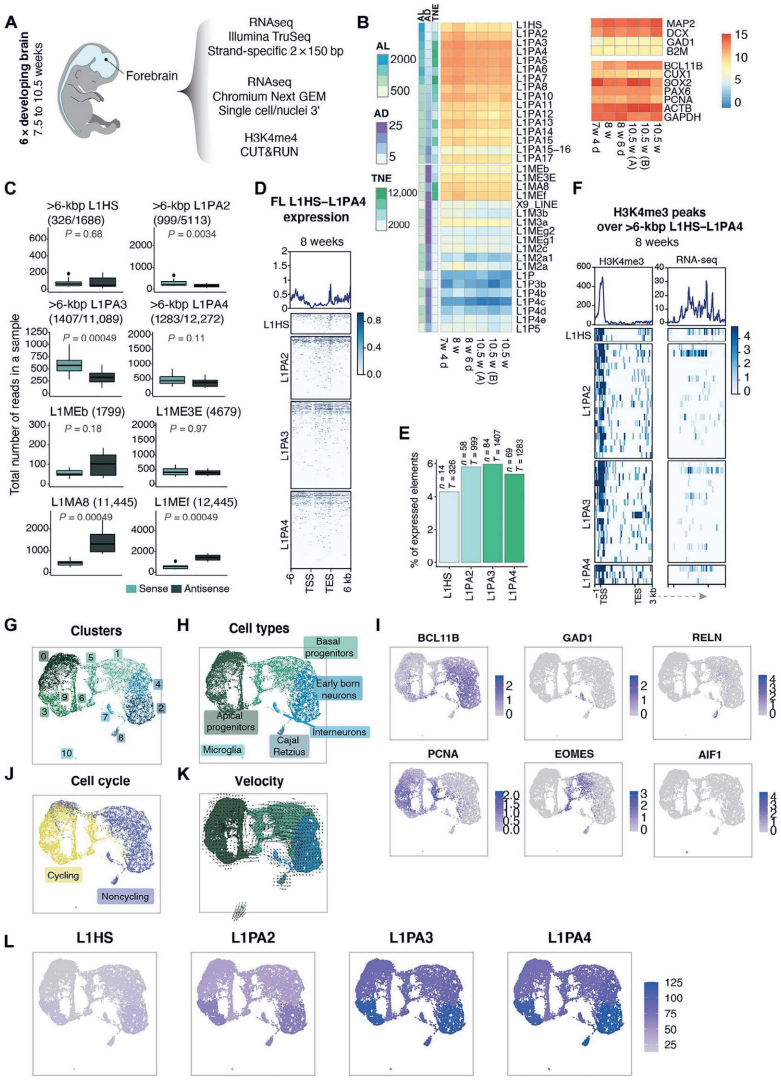


Fig. 3. L1s are expressed in human brain development. (**A**) Schematic of sequencing strategy of fetal human forebrain samples. (**B**) Expression of primate-specific L1 subfamilies compared to ancient L1 subfamilies and selected housekeeping and development-related genes as reference. Row annotation showing average length, average percentage of divergence from consensus, and the total number of elements (information extracted from RepeatMasker open-4.0.5). (**C**) Read count in sense (light teal) and antisense (dark teal) per sample. First four boxplots showing full-length elements in young L1 subfamilies and last four showing ancient L1 subfamilies with a comparable number of copies. (**D**) Expression (RPKM) over full-length (>6 kbp) L1HS, L1PA2, L1PA3, and L1PA4 plus 6-kbp flanking regions. (**E**) Percentage of expressed full-length (>6 kbp) elements (mean normalized counts, >10; see Methods) among young L1 subfamilies (*n* = number of expressed elements; *T* = total number of full-length elements). (**F**) Detected H3K4me3 peaks (left heatmap) over full-length L1 subfamilies (L1HS to L1PA4) and RNA-seq signal (right heatmap). Profile plots showing sum of signal. (**G**) Fetal human forebrain snRNA-seq UMAP colored by cluster. (**H**) UMAP colored by cell types. (**I**) Expression of selected biomarkers for different cell types. (**J**) UMAP colored by cell cycle state (based on CellCycleScoring from Seurat). (**K**) Velocity plot colored by cell type. (**L**) Pseudo-bulk cluster expression of young L1 subfamilies on UMAP.

(Fig. 4D). For example, L1s expressed uniquely during development were often located in introns of genes with a developmental specific expression pattern (Fig. 4D). Thus, this analysis indicates that the expression of individual L1 loci is governed by their integration site and the transcriptional activity of the nearby genome.

L1-derived transcripts contribute to transcriptome complexity in human neurons

The activity of the L1 promoter in the human brain suggests that L1s are a rich potential source of primate-specific and human-specific transcripts, which, in turn, may be co-opted and contribute to transcriptome complexity and speciation. When searching our dataset,

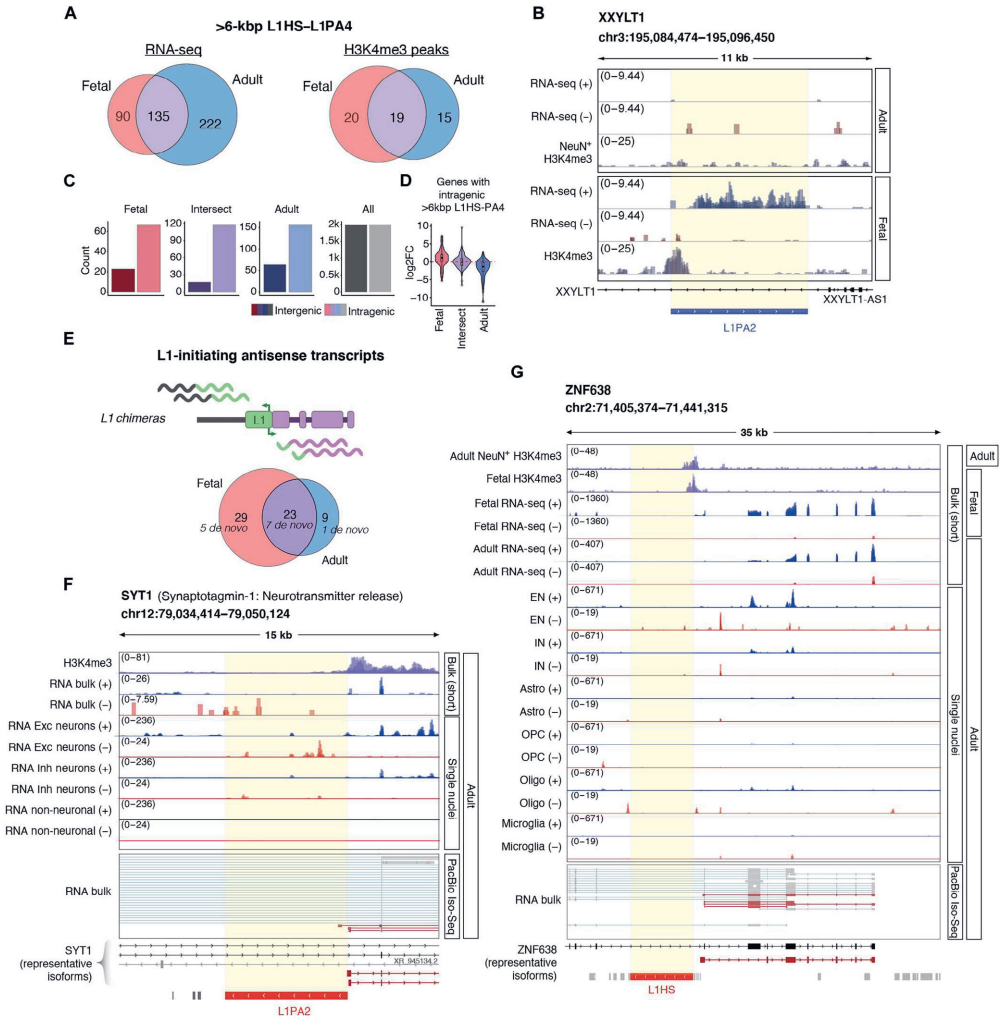


Fig. 4. L1s are dynamically expressed in the developing and the adult human brain. (**A**) Left: Number of expressed L1HS-L1PA4 (>6 kbp) in fetal (red) and adult samples (blue) (mean normalized counts, >10; see Methods) and the number of elements found to be expressed in both datasets (intersection; purple). Right: Number of H3K4me3 peaks over L1HS-L1PA4 (>6 kbp) in fetal (red) and adult samples (blue) and the intersection between datasets (purple). (**B**) Genome browser track showing the expression of a development-specific full-length L1PA2 with an H3K4me3 peak at its promoter. (**C**) Number of intragenic (light) or intergenic (dark) L1HS to L1PA4 (>6 kbp) in fetal (red), adult (blue), or those expressed in both datasets (purple). (**D**) $Log_2FoldChange$ (log_2FC) of the genes with an intragenic L1HS to L1PA4 (>6 kbp) in fetal (red), adult (blue), and the intersection (purple) [fetal versus adult (ref); DESeq2]. (**E**) L1s initiating antisense transcripts. Top: Schematic definition of L1 chimeras. Bottom: Total number of L1 chimeras expressed in fetal and adult samples. Number of the subset de novo annotated transcripts (not present in GENCODE hg38 version 38) in italics. (**F** and **G**) Genome browser tracks showing (from top to bottom): H3K4me3 CUT&RUN (samples overlayed in purple), short-read bulk RNA-seq (overlayed) split by strand (blue, forward; red, reverse), overlayed cluster expression (adult snRNA-seq) per cell type (or group of cell types), and PacBio Iso-Seq reads validating the presence of the transcript (supporting reads are highlighted in red). Annotation to the right showing data type and dataset (adult/fetal). (F) SYT1 with an antisense full-length L1PA2 at the beginning of one of its isoforms (L1 chimera). snRNA-seq tracks showing excitatory neurons (EN), inhibitory neurons (IN), and non-neuronal cell types overlayed [astrocytes, oligodendrocyte precursor cells (OPC), oligodendrocytes (Oligo), and microglia]. (G) ZNF638 with an antisense full-length L1HS as an alternative promoter (L1 chimera).

we found several such examples of co-option where L1s appear to have integrated into and modified the human transcriptome.

To investigate the presence of L1-derived transcripts, we performed de novo transcriptome assembly from the short-read bulk RNA-seq data from the fetal and adult samples. This analysis resulted in the identification of more than 60 chimeric transcripts originating from L1 promoters of which 13 represent transcripts previously not annotated in GENCODE (version 38) (table S3). When using the de novo transcript assembly in combination with our long-read RNA-seq data, we were able to validate L1 chimeras that create an alternative start site for several genes (table S3). For example, we found an L1PA2 that provides an alternative promoter for an isoform of *SYT1*. This transcript variant was supported by H3K4me3 and long-read bulk RNA-seq and was exclusively expressed in neurons as monitored by snRNA-seq (Fig. 4F). Another example was *ZNF638*, in which an L1HS serves as its alternative promoter (Fig. 4G). This isoform was supported by long-read RNA-seq, is expressed mostly in neurons, and hosts an H3K4me3 peak in both fetal and adult samples. Thus, our multiomics approach revealed several previously uncharacterized examples where L1s are integrated into the gene regulatory landscape of the developing and the mature human brain. Notably, all these L1s represent hominoid- or human-specific insertions.

To investigate the potential role of L1s in contributing to human brain functions, we focused on a transcriptionally active full-length L1PA2 element on chromosome 2 (6013 bp long). The L1 antisense promoter (*14*, *26*) serves as the TSS of an lncRNA: *LINC01876*. RNA-seq, snRNA-seq, and H3K4me3-CUT&RUN supported that the L1PA2 acts as an antisense promoter for this L1-lncRNA in human brain development (Fig. 5A). Notably, this expression appears to be limited to development since no *LINC01876* expression was found in the adult brain (Fig. 5A).

L1-lncRNA LINC01876 is a human-specific transcript

L1-derived RNAs have the potential to contribute to primate and human speciation since they originate from the integration of new DNA sequences into our genome. To investigate the evolutionary conservation of the L1-lncRNA *LINC01876*, we analyzed our previously published dataset from induced pluripotent stem cells (iPSCs) derived human and chimpanzee forebrain NPCs (fbNPCs) (Fig. 5B) (*46*). We found the L1-lncRNA was highly expressed in human fbNPCs, as supported by both RNA-seq and H3K4me3 CUT&RUN data (Fig. 5C). We were not able to detect L1-lncRNA expression in chimpanzee fbNPCs. We verified the human-specific expression of this L1-lncRNA in previously published human, chimpanzee, bonobo, gorilla, and macaque RNA-seq data from NPCs and immature neurons (*47*) and snRNA-seq from human, chimpanzee, and macaque cerebral organoids (*48*) (Fig. 5, D and E). The L1-lncRNA was consistently expressed in human NPCs, immature neurons, and organoids but not in cultures obtained from other primates. Thus, the L1-lncRNA *LINC01876* appears to be a human-specific transcript that is expressed during brain development.

We performed a multiple sequence alignment of the genomic region to investigate the evolutionary time point in which the L1PA2 was inserted into the ancestral primate genome. We found that the L1PA2 insertion site is present—and identical—in human, chimpanzee, bonobo, and gorilla, but not in orangutan, macaque, or other lower species (Fig. 5F) (*49*). Thus, this L1PA2 insertion can be estimated to have occurred around 10 to 20 million years ago. To explain how the L1PA2 element drives the expression of L1-lncRNA in humans, but not in other species, we focused on its promoter region. In intact young L1s, the antisense promoter drives the expression of a small L1 peptide, ORF0 (*25*) (Fig. 1G). When comparing the antisense promoter sequences of the L1PA2 insertion between humans, chimpanzees, bonobos, and gorillas, we noticed a missense mutation (A451G) in the Kozak sequence of the ORF0 in humans (Fig. 5F). This mutation was located at the start codon resulting in a methionine to threonine (M1T) change disabling translation of the ORF0 in humans (*25*). The ORF0 was still intact in chimpanzees, bonobos, and gorillas. Denisova and Neanderthal genomes both displayed the human variant, suggesting that the nucleotide change occurred before the split of archaic human species (Fig. 5F) (*49*). This analysis indicated that it is possible that the L1-lncRNA promoter may be silenced by DNA methylation or other repressive factors in nonhuman primates due to the expression and translation of an ORF0-fusion-transcript. The L1-lncRNA *LINC01876* might escape silencing in humans as ORF0 is not translated, although the underlying mechanisms remain to be investigated.

L1-lncRNA CRISPRi reveals an important role in neural differentiation

To investigate the functional relevance of the L1-lncRNA *LINC01876*, we set up a CRISPRi strategy to silence *LINC01876* expression. We designed two distinct guide RNAs (gRNAs) to target unique genomic locations in the vicinity of the TSS and coexpressed these with a Krüppel-associated box (KRAB) transcriptional repressor domain fused to catalytically dead Cas9 (KRAB-dCas9) (Fig. 6A and fig. S4A). Lentiviral transduction of human iPSCs resulted in efficient, almost complete silencing of *LINC01876* upon differentiation to fbNPCs (Fig. 6B and fig. S4B), but there was no difference in differentiation capacity or expression of cell fate markers compared to controls (Fig. 6C and fig. S4C). We also found no evidence that the expression of other L1 loci was affected by the CRISPRi approach demonstrating the specificity of the silencing to the *LINC01876* locus (fig. S4, D and E). The subsequently obtained results using the two different gRNAs were indistinguishable, and thus results were pooled.

We performed RNA-seq on *LINC01876*-CRISPRi fbNPCs and analyzed the transcriptome for alterations in gene expression. We found 41 significantly up-regulated genes and 10 down-regulated genes (DESeq2; P_{adj} <0.05, $\log_2$FoldChange > 1) (Fig. 6D). As lncRNAs can act in cis or trans (*18*), we scrutinized chromosome 2 to determine whether the differentially expressed genes were located near to the lncRNA, which would indicate a cis function. We found no obvious evidence suggesting that genes in the vicinity of the L1-lncRNA on chromosome 2 were affected by the CRISPRi, indicating that the L1-lncRNA may act in trans (fig. S5F).

We noted that many of the differentially expressed genes when comparing L1-lncRNA-fbNPCs to control fbNPCs were also differentially expressed when comparing human and chimpanzee fbNPCs (*46*). Twenty-seven of the 41 up-regulated genes upon L1-lncRNA CRISPRi were more highly expressed in chimpanzee fbNPCs upon L1-lncRNA CRISPRi, and 8 of the 10 down-regulated genes after L1-lncRNA CRISPRi were expressed at lower levels in chimpanzee fbNPCs (Fig. 6E). Thus, the L1-lncRNA appeared to influence the expression of several genes that distinguish the

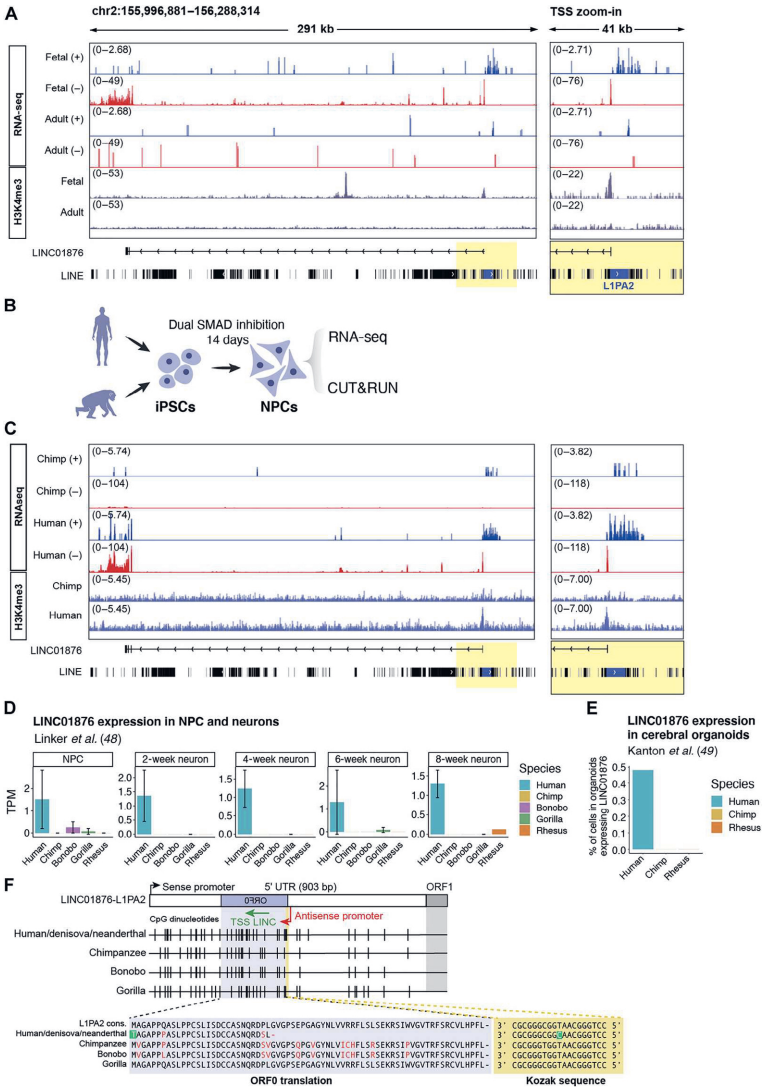


Fig. 5. The L1-lncRNA LINC01876 is a human-specific transcript. (**A**) Genome browser tracks showing RNA-seq and H3K4me3 signal (bottom) (in purple) over L1-lncRNA in fetal and adult samples. RNA-seq signal (RPKM) split by strand (blue, forward; red, reverse). Right: A zoom-in into the TSS (highlighted in yellow). (**B**) Experimental approach for fbNPCs human and chimpanzee comparison. (**C**) Genome browser tracks showing RNA-seq and H3K4me3 signal (bottom) (in purple) over L1-lncRNA in human and chimpanzee fbNPCs. RNA-seq signal (RPKM) split by strand (blue, forward; red, reverse). Right: A zoom-in into the TSS (highlighted in yellow). (**D**) LINC01876 (L1-lncRNA) expression [transcripts per million (TPM)] from bulk RNA-seq of human, chimpanzee, bonobo, gorilla, and macaque rhesus NPCs from Linker *et al.* (*47*). (**E**) Percentage of cells expressing LINC01876 (L1-lncRNA) in human, chimpanzee, and macaque rhesus cerebral organoids from Kanton *et al.* (*48*). (**F**) Multiple sequence alignment of the L1-lncRNA L1PA2 ORF0 (highlighted in purple) in different primates and their Kozak sequence (highlighted in yellow). The TSS of the L1-lncRNA is indicated in orange.

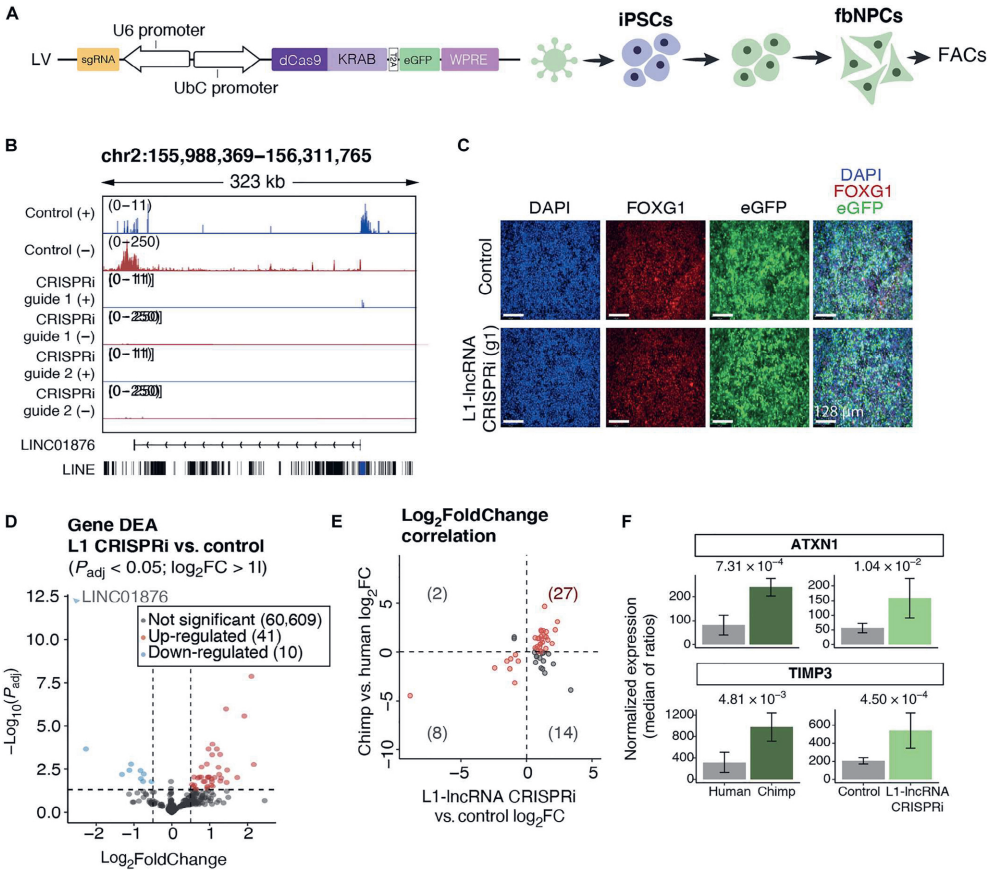


Fig. 6. CRISPRi-silencing of the L1-lncRNA in human fbNPCs. (**A**) CRISPRi construct and schematic of the L1-lncRNA CRISPRi in fbNPCs. (**B**) Genome browser tracks showing the expression over L1-lncRNA in control (LacZ) and L1-lncRNA CRISPRi. RNA-seq signal (RPKM) split by strand (blue, forward; red, reverse). (**C**) Immunohistochemistry of forebrain (red, FOXG1) and nuclear [blue, 4′,6-diamidino-2-phenylindole (DAPI)] markers. Enhanced green fluorescent protein (eGFP) showing transfected cells (green). Scale bars, 128 µm. (**D**) Volcano plot showing differential gene expression results (DESeq2). Significantly up-regulated and down-regulated genes are highlighted in red and blue, respectively (log2FoldChange > 1, P_{adj} < 0.05). (**E**) Log2FoldChange of the significantly up-regulated or down-regulated genes upon L1-lncRNA CRISPRi [as highlighted in (D) in the two datasets (L1-lncRNA CRISPRi versus control and human versus chimp). Genes up-regulated or down-regulated in both datasets are highlighted in red (first and third quadrants). (**F**) Normalized expression (median of ratios; DESeq2) of two example genes up-regulated in both datasets.

human and chimpanzee transcriptome in neural progenitors. Notably, some of these differentially expressed genes play important roles in the human brain such as Ataxin1 (*ATXN1*), which is mutated in spinocerebellar ataxia (*50*), and tissue inhibitor of metalloproteinases 3 (*TIMP3*), which is an inhibitor of the matrix metalloproteinases that have been linked to neurodegenerative disorders (Fig. 6F) (*51*).

L1-lncRNA LINC01876 contributes to developmental timing in cerebral organoids

To investigate the functional role of the L1-lncRNA in human brain development, we generated L1-lncRNA-CRISPRi cerebral organoids. This model allows for the study of human-specific developmental processes in three-dimensional (3D) (Fig. 7A) (*52*). We found that L1-lncRNA-CRISPRi silencing did not impair the organoid formation and the resulting organoids displayed characteristic neural rosettes after 30 days of growth, as visualized with Pax6/ZO1 staining (Fig. 7B). Quantification of organoid size throughout

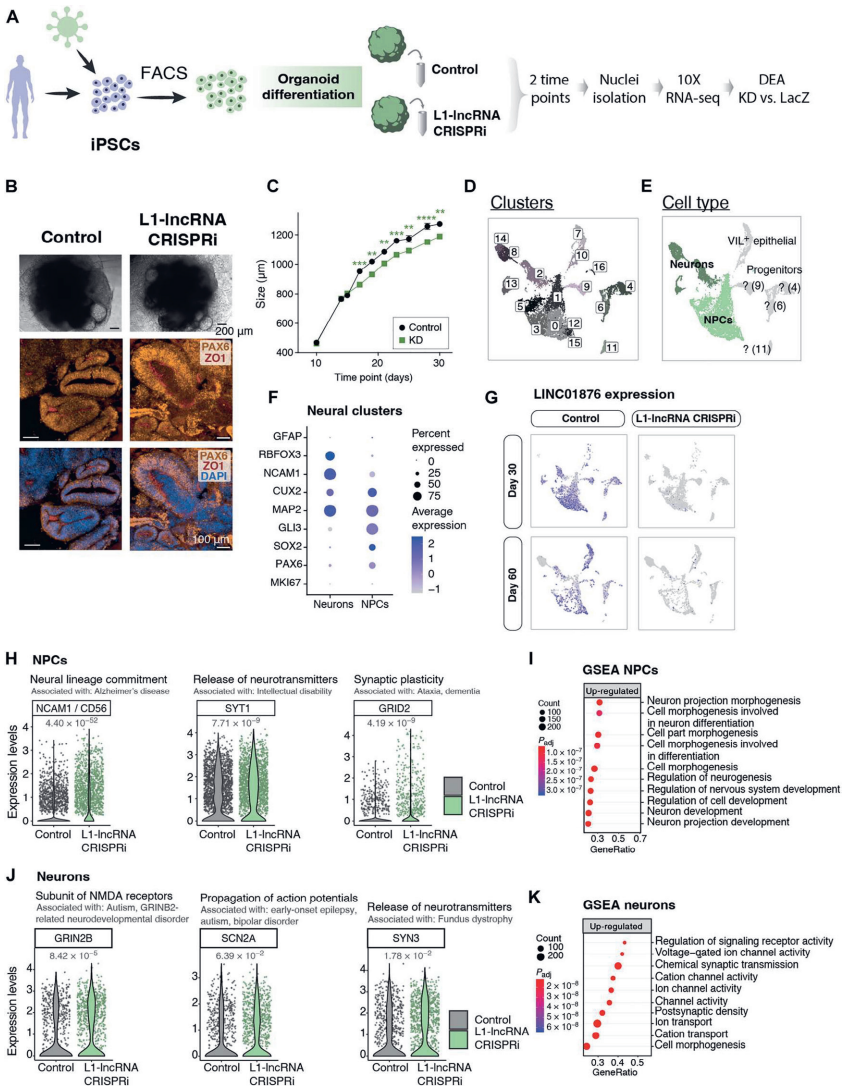


Fig. 7. Silencing of L1-lncRNA in cerebral organoids indicates it has a role in developmental timing. (**A**) Schematic of experimental design for organoid differentiation, L1-lncRNA CRISPRi, and sequencing. DEA, Differential Expression Analysis; KD, knock down. (**B**) Bright-field pictures of iPSC-derived cerebral organoids (top). Black scale bars, 200 µm. Immunohistochemistry of PAX6 (orange), ZO1 (red), and DAPI (blue) (bottom). White scale bars, 100 µm. (**C**) Quantification of organoid diameter between days 10 and 30 ($n = 20$ to 30 organoids per time point) (mixed-effects analysis and a Sidak correction for multiple comparisons). (**D**) snRNA-seq: UMAP colored by cluster. (**E**) UMAP colored by identified cell types. Neuronal-like clusters colored in two shades of green and uncharacterized clusters or progenitor-like cells colored in grey. VIL+ , *Villin 1* positive cells. (**F**) Dot plot showing expression of neuronal and neuronal progenitor markers in the NPC and neuronal clusters. (**G**) UMAP showing expression of L1-lncRNA. (**H**) Selected examples of significantly up-regulated genes in L1-lncRNA CRISPRi NPCs (FindMarkers from Seurat; $P_{adj} < 0.05$). (**I**) Selected up-regulated terms of the gene set enrichment analysis (GSEA) over NPCs (gseGO; $P_{adj} < 0.05$). (**J**) Selected examples of significantly up-regulated genes in L1-lncRNA CRISPRi neurons (FindMarkers from Seurat; $P_{adj} < 0.05$). (**K**) Selected up-regulated terms of GSEA over neurons (gseGO; $P_{adj} < 0.05$).

differentiation revealed that L1-lncRNA-CRISPRi organoids were reproducibly smaller than control organoids (Fig. 7C, table S2, and fig. S5A). This difference appeared after 2 weeks of growth and was sustained up until 1 month, which was the last time point quantified (Fig. 7C and table S2). The results were reproduced from three independent experiments using two different gRNAs (table S2).

To further evaluate the long-term molecular consequences of L1-lncRNA inhibition on human cerebral organoids, we analyzed organoids at 1 and 2 months of growth using snRNA-seq. High-quality data were generated from a total of 11,669 cells, including 6099 from L1-lncRNA-CRISPRi organoids (two gRNAs, in total 45 organoids) and 5570 from control organoids (*lacZ*-gRNA, in total 25 organoids). We performed an unbiased clustering analysis to identify and quantify the different cell types present in the organoids. Seventeen separate clusters were identified (Fig. 7D), including cerebral cells of different stages of maturation, such as NPCs and newborn neurons (Fig. 7, E and F). All the clusters contained cells from both 1 and 2 months, and we found no apparent difference in the contribution to the different clusters by L1-lncRNA-CRISPRi organoids, suggesting that the L1-lncRNA *LINC01876* does not influence developmental fate in cerebral organoids (fig. S5B).

Next, we analyzed the transcriptional difference between control and L1-lncRNA-CRISPRi organoids. We confirmed the transcriptional silencing of L1-lncRNA in all cell populations at both time points (Fig. 7G). Notably, in control (ctrl) organoids, the L1-lncRNA was expressed in NPCs but not in neurons, demonstrating that the 3D system is able to replicate an appropriate developmentally regulated expression pattern of this L1-derived transcript (Fig. 7G). We found that in the NPC population, genes linked to neuronal differentiation, such as *NCAM1*, *SYT1*, and *GRID2*, were up-regulated in L1-lncRNA-CRISPRi organoids (Fig. 7H). An unbiased gene set enrichment analysis (GSEA) of the up-regulated genes in NPCs was significantly enriched in gene ontology (GO) terms linked to neuronal differentiation (Fig. 7I). In line with this, we found that in newborn neurons, genes linked to mature neuronal functions, such as *GRIN2B*, *SCN2A*, and *SYN3*, were up-regulated in L1-lncRNA-CRISPRi organoids (Fig. 7J), and GSEA confirmed enrichment of up-regulated genes linked to neuronal maturation (Fig. 7K). These results indicate that NPCs and neurons present in organoids that lack the L1-lncRNA *LINC01876* display a more mature transcriptional profile than those found in control cerebral organoids.

Together, these results demonstrate that silencing of the L1-lncRNA *LINC01876* results in organoids that contain the same cell types as control organoids, suggesting that the L1-lncRNA does not influence developmental fate. However, we found that the L1-lncRNA organoids were smaller during early differentiation and displayed transcriptome changes in line with more mature NPCs and neurons. These observations are in line with a role for the L1-lncRNA in developmental timing since L1-lncRNA-CRISPRi organoids appear to differentiate quicker.

DISCUSSION

L1 mobilization represents a threat to human genomic integrity, and it has therefore been assumed that L1 expression is silenced in somatic human tissues. However, the abundance and repetitive nature of L1s make their transcription difficult to precisely estimate (*34*). Previous studies have, on the basis of retrotransposition activity, indirectly indicated that L1s may be expressed in the brain (*27*–*33*). In addition, observations based on quantitative real-time PCR (qRT-PCR), Northern blots, and Cap analysis of gene expression sequencing support that full-length (defined as >6 kbp) L1 transcripts are expressed in the human brain (*53*, *54*), but these approaches lack in precision, and it has been difficult to pinpoint the expression to individual loci. Therefore, several open questions remain as follows: Which L1-loci are expressed in the human brain and in what cell types? Are L1s developmentally regulated? Do L1-derived transcripts contribute to brain functions? In this study, we resolve many of these issues through the use of a careful multiomics analysis of human tissue, combined with a customized bioinformatics pipelines. We found that L1s are highly expressed in the developing human brain and in neurons in the adult human brain.

Our data demonstrate that the expression of L1s in the developing and adult human brain is largely limited to evolutionarily young, primate specific L1s, primarily subclasses found only in hominoids. The lack of expression of more ancient L1s is likely explained by the higher burden of deletions, mutations, and genomic rearrangements of old TEs that reduce their capacity to be transcribed. A strand-specific analysis of full-length elements that contain an intact 5′ promoter revealed that the RNA-seq signal was present in sense to the L1s. We thereby confirmed that hundreds of different L1 loci are expressed and that the L1 signal is not transcriptional noise but rather that the L1 promoter drives expression. This strongly suggests that the signal is not the result of passive expression in which the L1 sequence is incorporated into another transcript (*34*). We confirmed this with two orthogonal approaches: by performing long-read RNA-seq analysis to identify L1 transcripts that initiate in the L1 5′UTR and by H3K4me3 profiling to identify L1 promoters active in the human brain, benefiting from the fact that the signal of this histone modification spreads to the flanking (and thus unique) genomic context. We thus found bona fide evidence that full-length L1s are expressed in both the developing and the adult human brain. However, we acknowledge that with our approach, we miss the expression of polymorphic L1s not present in the reference genome. Future studies using individual-matched RNA-seq and long-read genome data will be crucial to investigate whether L1s individualize the neuronal transcriptome.

From our analysis, it is evident that not all L1 loci are expressed in the brain, but rather a small subset. Our data also indicate that the L1 integration site is important and that the presence of highly active nearby gene promoters or other regulatory elements is key for L1 expression. Thus, the activity of the surrounding genome is one parameter that is important for how this subset of L1s escapes silencing. In this respect, our results are similar to what have previously been found in cancer cell lines (*45*). In addition, single-nucleotide variants or small deletions in regulatory regions of individual L1 integrant could result in the avoidance of recruiting silencing factors. A previous study indicated that a subset of young L1s that have lost a Yin Yang 1 (YY1) binding site in the promoter avoids silencing in the brain in a DNA methylation–dependent manner (*32*). However, in our dataset, we found L1s both with and without the YY1 binding site to be expressed (fig. S6, A and B). Thus, we do not fully understand the mechanism by which these L1s escape silencing. However, it is worth noting that the adult brain tissue samples used for this study came from individuals aged between 69 and 87 years old at the time of death. It is well

established that DNA methylation patterns change with age and there are emerging studies linking age-related epigenetic changes to activation of TE expression (*55*–*57*). This raises the possibility that some of the L1 transcripts we detect in adult neurons may be aging dependent. Future studies investigating the link between human brain aging and TE expression are needed to resolve this question. Another interesting aspect of our data is that L1HS elements appear to be globally silenced in brain development. This indicates that L1HS elements are controlled by unique, specialized mechanisms during brain development, likely to avoid abundant retrotransposition events in proliferating cell populations. The nature of this mechanism remains unknown, but it will be interesting to investigate further to understand how the human brain avoids waves of retrotransposition events during early development and what the consequences are if this mechanism fails.

The fact that many L1 promoters are active in the human brain demonstrates that L1s are a rich source of genetic sequences that provides a primate-specific layer of transcriptional complexity. Our data indicate that L1s influence the expression of protein-coding genes and noncoding transcripts in the human brain through several mechanisms, including acting as alternative promoters or by altering 5′UTR and 3′UTR. In addition, there is the possibility that L1-derived peptides or fusion peptides play important functional roles (*58*). One example of an L1-derived noncoding transcript that we identified is *LINC01876*, an L1-lncRNA that exploits the antisense promoter of an L1PA2 element that is transcriptionally active in human brain development. In the *LINC01876* promoter, the first amino acid of ORF0 is specifically mutated in humans, and the subsequent loss of ORF0 coding capacity correlates with the appearance of the L1-lncRNA. It is possible that this single-nucleotide variant, at a key position for the L1-life cycle, enables the escape of DNA methylation–mediated silencing resulting in transcription of the lncRNA.

Our loss-of-function studies of the L1-lncRNA *LINC01876* in cerebral organoids suggest that it may play an important role in regulating developmental timing during human brain development. *LINC01876* is a previously uncharacterized lncRNA, but we have noted that there is a T > C single-nucleotide polymorphism in the L1-derived promoter region of *LINC01876* that has been linked to major depressive disorders in a genome-wide association meta-analysis (*59*). Our data demonstrate that organoids in which *LINC01876* expression was silenced were smaller in size and displayed NPCs and neurons with a more mature transcriptome than control counterparts. These findings are reminiscent of previously observed differences when comparing human cerebral organoids to those derived from nonhuman great apes (*48*, *60*, *61*). Thus, our data provide experimental evidence as to how an L1 insertion may have contributed to the evolution of the human brain and provide a potential link between L1s and the genetics of neuropsychiatric disorders that will be interesting to study in more detail in the future.

In summary, our results illustrate how L1s provide a layer of transcriptional complexity in the brain and provide evidence for L1s as genetic elements with relevance in human brain function. It has been estimated that a new L1 germline insertion occurs in every 50 to 200 human births (*9*, *40*). This extensive L1 mobilization in the human population has resulted in hundreds of unfixed polymorphic L1 insertions in each human genome (*9*, *62*). Since L1s are highly polymorphic within the human population, the prevalence of certain L1 copies or single-nucleotide polymorphisms and structural variants in fixed L1s in the genome is therefore likely to influence the etiology of brain disorders. Thus, L1s represent a set of genetic materials that are implicated in the evolution of our brain and may contribute to important gene regulatory and transcriptional networks in the human brain. L1s should no longer be neglected, and these sequences need to be included in future investigations of the underlying genetic causes of human brain disorders.

METHODS

Human tissue

Human fetal forebrain tissue was obtained from material available following elective termination of pregnancy at the University Hospital in Malmö, Sweden, in accordance with the national ethical permit (Dnr 6.1.8-2887/2017). Postmortem cortical tissue was obtained in accordance with the national ethical permit (Dnr 2019-06582, beslut 2020-02-12). Written informed consent was obtained from all donors.

Induced pluripotent stem cells

Human iPSC line generated by mRNA transfection was used: RBRC-HPS0328 606A1, hereafter referred to as HS1 (Riken, RRID:CVCL_DQ11). The iPSC line was maintained as previously described (*46*, *63*, *64*). Briefly, the iPSC lines were maintained on LN521 (0.7 μg/cm^2; BioLamina)–coated Nunc multidishes in iPSC medium (StemMACS iPSC-Brew XF and 0.5% penicillin/streptomycin; Gibco) and were passaged 1:2 to 1:6 every 2 to 4 days once 70 to 90% confluency was reached. The medium was changed daily, and 10 μM Y27632 (Rock inhibitor, Miltenyi) was added when cells were passaged.

Forebrain neural progenitor cells

iPSCs were differentiated into fbNPCs as previously described (*46*, *63*). Upon dissociation at 70 to 90% confluency, the cells were plated on LN111 (1.14 μg/cm^2; BioLamina)–coated Nunc multidishes at a density of 10,000 cells/cm^2 and grown in N2 medium [1:1 Dulbecco's modified Eagle's medium (DMEM)/F12 (21331020, Gibco) and Neurobasal (21103049, Gibco) supplemented with 1% N2 (Gibco), 2 mM L-glutamine (Gibco), and 0.2% penicillin/streptomycin]. SB431542 (10 μM; Axon) and noggin (100 ng/ml; Miltenyi) were given for dual SMAD inhibition. The medium was changed every 2 to 3 days. On day 9, N2 medium without dual SMAD inhibitors was used. On day 11, cells were dissociated and replated on LN111-coated Nunc multidishes at a density of 800,000 cells/cm^2 in B27 medium [Neurobasal supplemented with 1% B27 without vitamin A (Gibco), 2 mM L-glutamine, and 0.2% penicillin/streptomycin Y27632 (10 μM), brain-derived neurotrophic factor (BDNF; 20 ng/ml; R&D), and L-ascorbic acid (0.2 mM; Sigma-Aldrich)]. Cells were kept in the same medium until day 14 when cells were harvested for downstream analysis.

CRISPR interference

To silence the expression of *LINC01876* in iPSCs, we adapted a previously described protocol (*46*). Single-guide sequences were designed to recognize DNA regions near the TSS according to the GPP Portal (Broad Institute). The guide sequences were inserted into a dCas9-KRAB-T2A-GFP lentiviral backbone and pLV hU6-sgRNA hUbC-dCas9-KRAB-T2a-GFP, a gift from C. Gersbach

(Addgene plasmid #71237, RRID:Addgene 71237), using annealed oligodendrocytes and the Bsm BI cloning site. Lentivirus was produced as described below, and iPSCs were transfected with multiplicity of infection of 10 of *LacZ* and *LINC01876*-targeting gRNA. Guide efficiency was validated using standard qRT-PCR techniques: *LINC01876* guide 1, ACGAGATTATAAGCCGCACC; *LINC01876* guide 2, AGGGGCGCCCGCCGTTGCCC; *LacZ*, TGCGAATACGCCCACGCGAT.

Green fluorescent protein–positive cell isolation of fbNPCs

At day 14, cells were detached with Accutase, resuspended in B27 medium containing RY27632 (10 μM) and Draq7 (1:1000; BD Biosciences), and strained with a 70-μm (BD Biosciences) filter. Gating parameters were determined by side and forward scatter to eliminate debris and aggregated cells. The green fluorescent protein (GFP)−positive gates were set using untransduced fbNPCs. The sorting gates and strategies were validated via reanalysis of sorted cells (>95% purity cutoff). A total of 200,000 GFP-positive/Draq7-negative cells were collected per sample, spun down at 400*g* for 5 min, and snap-frozen on dry ice. Cell pellets were kept at −80°C until RNA was isolated.

GFP-positive cell isolation of transduced iPSCs

Seven days after transduction, cells were detached with Accutase, resuspended in iPSC medium containing RY27632 (10 μM) and Draq7 (1:1000), and strained with a 70-μm filter. Gating parameters were determined by side and forward scatter to eliminate debris and aggregated cells. The GFP-positive gates were set using untransduced iPSCs. The sorting gates and strategies were validated via reanalysis of sorted cells (>95% purity cutoff). A total of 200.000 GFP-positive/Draq7-negative cells were collected per sample, spun down at 400*g* for 5 min and resuspended in iPSC medium containing RY27632 (10 μM) and expanded as described above and frozen down for further use. Detailed protocol can be found at DOI: dx.doi.org/10.17504/protocols.io.yxmvm25n9g3p/v1.

Lentiviral production

Lentiviral vectors were produced according to Zufferey *et al.* (*65*) and were titered by qRT-PCR. Briefly, human embryonic kidney–293T cells (RRID:CVCL_0063) were grown to a confluency of 70 to 90% for lentiviral production. Third-generation packaging and envelope vectors [pMDL (#12251, Addgene), psRev (#12253, Addgene), and pMD2G (#12259, Addgene)] together with polyethyleneimine (PN 23966, PEI Polysciences) in Dulbecco's phosphate-buffered saline (DPBS; Gibco) were used in conjunction with the lentiviral plasmids previously generated. The lentivirus was harvested 2 days after transfection. The medium was collected, filtered, and centrifuged at 25,000*g* for 1.5 hours at 4°C. The supernatant was removed from the tubes, and the virus was resuspended in DPBS and left at 4°C. The resulting lentivirus was aliquoted and stored at −80°C.

Quantitative real-time polymerase chain reaction

Total RNA was first extracted using the miniRNeasy kit (QIAGEN). Complementary DNA (cDNA) was generated using the Maxima First Strand cDNA Synthesis Kit (Thermo Fisher Scientific). Quantitative PCR was performed using SYBR Green I master (Roche) on a LightCycler 480 (Roche). The $2^{-\Delta\Delta Ct}$ method was used to normalize expression to control, relative to glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and B-ACTIN as described previously (*66*).The gene primers used are as follows: *LINC01876*, 5′-AATCCGTGCCAGCAGTAAGT-3′ (forward) and 5′-GGACCTCTTCAAGTCCCAGG-3′ (reverse); *ACTB*, 5′-CCTTGCACATGCCGGAG-3′ (forward) and 5′-GCACAGAGCCTCGCCTT-3′; *GAPDH*, 5′-TTGAGGTCAARGAAGGGGTC-3′ (forward) and 5′-GAAGGTGAAGGTCGGAGTCA-3′ (reverse).

Human cerebral organoid culture

To generate the human cerebral-like organoids, we followed the protocol detailed in (*46*). We used three HS1-derived lines obtained by transduction and FACS sorting as described above: one control line (guide against *LacZ*) and two *LINC01876* CRISPRi lines (guide 1 and guide 2). Briefly, 8000 cells per well were plated in a 96-well plate (Costar, ultra low attachment, round bottom; REF 7007) with 250 μl of mTeSR1 (STEMCELL Technologies Inc.) and 10 μM RY27632. This is considered day −5 of the differentiation of the iPSC-derived human forebrain organoids. On days −3 and −1, the medium was changed (150 and 200 μl of mTeSR1, respectively). At day 0, the cells are fed with neural induction medium [DMEM/F12 medium, N2 supplement (1:100), L-glutamine (2 mM), penicillin/streptomycin (1:500), nonessential amino acids (1:100), and heparin (2 μg/ml)] enriched with 3% knockout replacement serum (#10828010, Gibco). On days 2, 4, and 6, the organoids were fed with neural induction medium with no added knockout replacement serum.

On day 8, the organoids were embedded in 30 to 50 μl of Matrigel (Corning) and incubated at 37°C for 25 min to allow the Matrigel to solidify. The organoids were then transferred in Corning (REF 3471) six-well plates with flat bottoms containing 4 ml per well of cortical differentiation medium [F12 medium (−glut) (48.5%), Neurobasal (48.5%), N2 supplement (1:200), B27 supplement (−vitamin A; 1:100), L-glutamine (2 mM), penicillin/streptomycin (1:500), nonessential amino acids (1:200), β-mercaptoethanol (50 μM), and insulin (2.5 μg/ml)].

On days 10 and 12 of the differentiation, the medium was changed exchanging 3 ml per well for 3 ml of fresh cortical differentiation medium. On days 15, 17, 19, 21, and 23, ~4 ml of the medium was replaced with 4 ml of improved differentiation medium + A [F12 media (−glut) (48.5%), Neurobasal (48.5%), N2 supplement (1:200), B27 supplement (+vitamin A; 1:50), L-glutamine (2 mM), penicillin/streptomycin (1:500), nonessential amino acids (1:200), β-mercaptoethanol (50 μM), insulin (2.5 μg/ml), and ascorbic acid (400 μM)]. From day 25, the medium was changed every 3 days with 3 to 4 ml of cortical terminal differentiation medium [F12 media (−glut) (48.5%), Neurobasal (48.5%), N2 supplement (1:200), 800 μL of B27 supplement (+vitamin A; 1:50), L-glutamine (2 mM), penicillin/streptomycin (1:500), nonessential amino acids (1:200), β-mercaptoethanol (50 μM), insulin (2.5 μg/ml), ascorbic acid (400 μM), BDNF (10 ng/μl), adenosine 3′,5′-monophosphate (200 μM), and glial cell line–derived neurotrophic factor (10 ng/μl)].

We performed three independent replicates of the CRISPRi experiment in cerebral organoids (three batches). We measured the size of 10 organoids per time point and condition in each batch. All the diameter measurements of the organoids were taken with the measure tool from ImageJ (RRID:SCR_003070). The chosen measuring unit was micrometers.

The statistical analysis to test the difference in organoid growth upon L1-lncRNA CRISPRi per guide was performed using a two-

way analysis of variance (ANOVA), adjusting for multiple comparison using a Dunnett correction. The statistical analysis to test the difference in organoid growth pooling both gRNAs for the L1-lncRNA CRISPRi was performed using a mixed-effects analysis and a Sidak correction for multiple comparisons. Detailed can be found at DOI: dx.doi.org/10.17504/protocols.io.e6nvwjo27lmk/v1.

Immunocytochemistry

The cells were washed three times with DPBS and fixed for 10 min with 4% paraformaldehyde (Merck Millipore), followed by three more rinses with DPBS. The fixed cells were then blocked for 60 min in a blocking solution of KPBS with 0.25% Triton X-100 (Thermo Fisher Scientific) and 5% donkey serum at room temperature.

The primary antibody [rabbit anti-FOXG1 (Abcam, RRID: AB_732415); 1:50] was added to the blocking solution and incubated overnight at room temperature. Subsequently, the cells were washed three times with KPBS. The secondary antibody [donkey anti-rabbit Cy3 (catalog no. 711165152, Jackson ImmunoResearch, RRID:AB_2307443); 1:200] was added with 4′,6-diamidino-2-phenylindole (DAPI) (Sigma-Aldrich; 1:1000) to the blocking solution and incubated at room temperature for 1 hour, followed by two to three rinses with KPBS. The cells were visualized on a Leica microscope (model DMI6000 B). Detail protocol can be found at DOI: dx.doi.org/10.17504/protocols.io.5qpvor7pdv4o/v1.

Immunohistochemistry

Organoids were fixed in 4% paraformaldehyde for 2 hours at room temperature. They were subsequently washed three times with KPBS and left in a 1:1 30% sucrose solution and OCT (catalog no. 45830, HistoLab) mixture overnight at 4°C. Organoids were then transferred to a cryomold containing OCT, frozen on dry ice, and stored at −80°C in freezer bags.

Before staining, organoids were sectioned on a cryostat at −20°C at a thickness of 20 μm and placed onto Superfrost plus microscope slides. They were then washed three times with KPBS for 5 min and subsequently blocked and permeabilized in 0.1% Triton X-100 and 5% normal donkey serum in KPBS for 1 hour at room temperature. The primary antibody [rabbit anti-PAX6 (catalog no. 901301, BioLegend, RRID:AB_2565003), 1:300 dilution; and rat anti-ZO1 (catalog no. NB110-68140, Novus, RRID:AB_1111431), 1:300 dilution] was added to the blocking solution and incubated overnight at room temperature. Subsequently, the sections were washed three times with KPBS. The secondary antibody [donkey anti-rabbit Cy3 (catalog no. 711165152, Jackson ImmunoResearch, RRID: AB_2307443), 1:200; and donkey anti-rat Cy5 (catalog no. 712175153, Jackson ImmunoResearch, RRID: AB_2340672), 1:200] was added with DAPI (Sigma-Aldrich; 1:1000) to the blocking solution and incubated at room temperature for 1 hour, followed by two to three rinses with KPBS. Sections were imaged using Operetta CLS (PerkinElmer). Detail protocol can be found at DOI: dx.doi.org/10.17504/protocols.io.n92ldp22nl5b/v1.

Single-nucleus isolation

The nucleus isolation from the embryonic brain tissue and organoids was performed as described previously (*36*). Briefly, the tissue and organoids were thawed and dissociated in ice-cold lysis buffer [0.32 M sucrose, 5 mM $CaCl_2$, 3 mM MgAc, 0.1 mM Na_2 EDTA, 10 mM tris-HCl (pH 8.0), and 1 mM dithiothreitol] using a 1-ml tissue douncer (Wheaton). The homogenate was carefully layered on top of a sucrose cushion [1.8 M sucrose, 3 mM MgAc, 10 mM tris-HCl (pH 8.0), and 1 mM dithiothreitol] before centrifugation at 30,000*g* for 2 hours and 15 min. Pelleted nuclei were softened for 10 min in 100 μl of nuclear storage buffer [15% sucrose, 10 mM tris-HCl (pH 7.2), 70 mM KCl, and 2 mM $MgCl_2$] before being resuspended in 300 μl of dilution buffer [10 mM tris-HCl (pH 7.2), 70 mM KCl, and 2 mM $MgCl_2$] and run through a cell strainer (70 μm). Cells were run through the FACS (FACS Aria, BD Biosciences) at 4°C at a low flow rate using a 100-μm nozzle (reanalysis showed >99% purity). Nuclei intended for bulk RNA-seq were pelleted at 1300*g* for 15 min. Detail protocol can be found DOI: dx.doi.org/10.17504/protocols.io.5jyl8j678g2w/v1.

3′ and 5′ single-nucleus sequencing

Nuclei or cells intended for single-cell/nucleus RNA-seq (8500 nuclei/cells per sample) were directly loaded onto the Chromium Next GEM Chip G or Chromium Next GEM Chip K Single Cell Kit along with the reverse transcription mastermix following the manufacturer′s protocol for the Chromium Next GEM single cell 3′ kit (PN-1000268, 10x Genomics) or Chromium Next GEM Single Cell 5′ Kit (PN-1000263, 10x Genomics), respectively, to generate single-cell gel beads in emulsion. cDNA amplification was done as per the guidelines from 10x Genomics using 13 cycles of amplification for 3′ and 15 cycles of amplification for 5′ libraries. Sequencing libraries were generated with unique dual indices (TT set A) and pooled for sequencing on a Novaseq6000 using a 100-cycle kit and 28-10-10-90 reads.

Single-cell/nucleus RNA-seq analysis

Gene quantification. The raw base calls were demultiplexed and converted to sample-specific fastq files using 10x Genomics Cell Ranger mkfastq (version 3.1.0; RRID:SCR_017344) (*67*). Cell Ranger count was run with default settings, using an mRNA reference for single-cell samples and a pre-mRNA reference (generated using 10x Genomics Cell Ranger 3.1.0 guidelines) for single-nucleus samples.

To produce velocity plots, loom files were generated using velocyto (*43*) (version 0.17.17; RRID:SCR_018167) run 10× in default parameters, masking for TEs [same general feature format (GTF) file as input for TEtranscripts; see the "TE subfamily quantification" section] and GENCODE version 36 as guide for features. Plots were generated using velocyto.R (see GitHub under src/analysis/fetal_velocity.Rmd).

Clustering. Samples were analyzed using Seurat (version 3.1.5; RRID:SCR_007322) (*68*). For each sample, cells were filtered out if the percentage of mitochondrial content was over 10% (perc_mitochondrial). For adult samples, cells were discarded if the number of detected features (nFeature_RNA) was higher than two SDs over the mean in the sample (to avoid keeping doublets) or lower than an SD below the mean in the sample (to avoid low quality cells). For fetal samples, cells were discarded if the number of detected features was higher than two SDs over the mean in the sample or lower than 2000 features detected. Counts were normalized using the centered log ratio transformation (Seurat::NormalizeData), and clusters were found with a resolution of 0.5 (Seurat::FindClusters).

Gene differential expression analysis. We used Seurat′s FindMarkers grouped by cell types and on default parameters as for version 4.3.0 to identify differentially expressed genes (Wilcoxon test). A gene was considered to be differentially expressed on a

cell type if its adjusted *P* value was below 0.05 and its average $\log_2$FoldChange is over 0.25 (default).

TE quantification. We used an in-house pseudo-bulk approach to processing snRNA-seq data to quantify TE expression per cluster, similar to what has been previously described (*36*). All clustering, normalization and merging of samples were performed using the contained scripts of get_clusters.R [get_custers() from the Sample class] and merge_samples.R [merge_samples() from the Experiment class] of trusTEr (version 0.1.1; doi:10.5281/zenodo.7589548). Documentation of the pipeline can be found at https://raquelgarza.github.io/truster/.

The main functionality of trusTEr is to create collections of reads to remap and quantify TE subfamilies or elements per group of cells. The function tsv_to_bam() backtraces cells barcodes to Cell Ranger's output binary alignment map (BAM) file. tsv_to_bam() runs using subset-bam from 10x Genomics version 1.0 (RRID:SCR_023216). As the next step of the pipeline, the function filter_UMIs() filters potential PCR duplicates in the BAM files; this step uses Pysam version 0.15.1 (RRID:SCR_021017). Next, to convert BAM to fastq files, we used bamtofastq from 10x Genomics (version 1.2.0; RRID: SCR_023215). The remapping of the clusters was performed using STAR aligner (version 2.7.8a; RRID:SCR_004463). Quantification of TE subfamilies was done using TEcount (version 2.0.3; RRID:SCR_023208), and individual elements were quantified using featureCounts (subread version 1.6.3; RRID:SCR_012919). The normalization step of trusTEr, to integrate with Seurat and normalize TE subfamilies' expression, was performed using Seurat version 3.1.5 (RRID:SCR_007322).

For the purposes of this paper, we combined the samples from the same condition (all embryonic samples and all adult samples). The quantification was run twice: with all samples together and per sample in the combined clustering. For the fetal samples, we also ran trusTEr grouping clusters per cell cycle state, for which we prepared a directory with tsv files containing the barcodes of the cells in each of the clusters of interest (e.g., cluster0_cycling.tsv, cluster0_noncycling.tsv, ...) and ran the set_merge_samples_outdir function from the Experiment class to register these as cluster objects.

Bulk RNA-seq

Total RNA was isolated from nuclei, cell culture samples, or tissue using the RNeasy Mini Kit (QIAGEN). Libraries were generated using Illumina TruSeq Stranded mRNA library prep kit [poly(A) selection] and sequenced on a NextSeq500 (PE, 2 × 150 bp). Protocol can be found at DOI: https://dx.doi.org/10.17504/protocols.io.36wgqjqbkvk5/v1.

Bulk RNA-seq analysis

TE subfamily quantification. For the quantification of TE subfamilies, the reads were mapped using STAR aligner (version 2.6.0c; RRID:SCR_004463) (*69*) with an hg38 index and GENCODE version 36 as the guide GTF (--sjdbGTFfile), allowing for a maximum of 100 multimapping loci (--outFilterMultimapNmax 100) and 200 anchors (--winAnchorMultimapNmax). The rest of the parameters affecting the mapping was left in default as for version 2.6.0c.

The TE subfamily quantification was performed using TEcount from the TEToolkit (version 2.0.3; RRID:SCR_023208) in mode multi (--mode). GENCODE annotation v36 was used as the input gene GTF (--GTF), and the provided hg38 GTF file from the author's web server was used as the TE GTF (--TE) (*35*).

TE quantification. Reads were mapped using STAR aligner (version 2.6.0c; RRID:SCR_004463) (*69*) with an hg38 index and GENCODE version 30 (adult data) and 36 (fetal data) as the guide GTF (--sjdbGTFfile). To quantify only confident alignments, we allowed a single mapping locus (--outFilterMultimapNmax 1) and a ratio of mismatches to the mapped length of 0.03 (--outFilterMismatchNoverLmax).

To measure the antisense transcription over a feature, we divided the resulting BAM file into two, containing the forward and reverse transcription, respectively. We used SAMtools view (version 1.9; RRID:SCR_002105) (*70*) to keep only the alignments in forward transcription, we separated alignments of the second pair mate if they mapped to the forward strand (-f 128 -F 16) and alignments of the first pair mate if they map to the reverse strand (-f 80). To keep the reverse transcription, we kept alignments of the second pair mate if they mapped to the reverse strand (-f 144) and alignments of the first pair mate if they mapped to the forward strand (-f 64 -F 16).

Both BAM files were then quantified using featureCounts from the subread package (version 1.6.3; RRID:SCR_012919) (*71*) forcing strandness to the features being quantified (-s 2). For consistency (and to avoid quantifying over simple repeats, small RNAs, and low-complexity regions), we input the same curated hg38 GTF file provided by the TEtranscripts authors (*35*).

Gene quantification. Reads were mapped using STAR aligner (version 2.6.0c; RRID:SCR_004463) (*69*) with an hg38 index and GENCODE version 36 as the guide GTF (--sjdbGTFfile), and no other parameters were modified (default values for --outFilterMultimapNmax, --outFilterMismatchNoverLmax, and --winAnchorMultimapNmax). Genes were quantified using featureCounts from the subread package (version 1.6.3; RRID:SCR_012919) (*71*) forcing strandness (-s 2) to quantify by gene_id (-g) from the GTF of GENCODE version 36.

Differential gene expression analysis. We performed differential expression analysis using DESeq2 (version 1.28.1; RRID:SCR_015687) (*72*) with the read count matrix from featureCounts (subread version 1.6.3; RRID:SCR_012919) as input. Fold changes were shrunk using DESeq2:: lfcShrink.

For the produced heatmaps, counts were normalized by median of ratios as described by Love *et al.* (*72*), summed with a pseudocount of 0.5 and $\log_2$-transformed.

For further detail, please refer to the Rmarkdown on the GitHub.

Transcript assembly and quantification. Transcript assembly for each of the short-read bulk RNA-seq samples was performed using StringTie (version 1.3.3b; RRID:SCR_016323) (*73*) with GENCODE hg38 version 38 as guide (-G). Output assemblies were merged by StringTie -merge, using the same GENCODE annotation as guide (-G). Transcript assemblies were then performed for each sample using the resulting GTF output from StringTie merge as the guide reference annotation (-G); the resulting GTFs from this step will hereon be referred as the samples' transcript assembly GTF. Read count tables were then generated using the accessory script from StringTie prepDE.py (https://github.com/gpertea/stringtie/blob/master/prepDE.py).

To identify L1 chimeras (table S3), we concatenated the samples' transcript assembly GTFs. We kept only unique transcript features with over 1 kbp of length. We created an auxiliary GTF file keeping

only the TSS of each transcript plus 100-bp windows in both directions—this file will hereon be referred to as the transcripts' TSS GTF. Using BEDTools intersect and forcing for opposing strands (-S), we intersected the transcripts' TSS GTF to full-length (>6 kbp) L1PAs using the RepeatMasker's annotation (open-4.0.5).

Transcripts' read count matrices were normalized using the DESeq2 (version 1.28.1; RRID:SCR_015687) (*72*) sizeFactors as calculated using the gene count matrix (see the "Differential gene expression analysis" section). A transcript was considered to be expressed on a sample if its normalized expression value exceeded that of 20. These transcripts were considered for Venn diagrams shown in Fig. 4E.

Differential TE subfamilies expression analysis. We performed differential expression analysis using DESeq2 (version 1.28.1; RRID:SCR_015687) (*72*) with the read count matrix from TEcount (version 2.0.3; RRID:SCR_023208) (*35*) using only the TE subfamilies entries. Fold changes were shrunk using DESeq2:: lfcShrink.

Using the gene DESeq2 object (see section above), we normalized the TE subfamily counts by dividing the read count matrix by the sample distances (sizeFactor) as calculated by DESeq2 with the quantification of genes without multimapping reads (see the "Bulk RNA-seq analysis: Gene quantification" section). For heatmap visualization, a pseudo-count of 0.5 was added and $\log_2$-transformed.

Comparison between sense and antisense transcription over TEs. To normalize uniquely mapped read counts per strand (see the "Bulk RNA-seq analysis: TE quantification" section), we divided the read count matrix by the sample distances (sizeFactor) as calculated by DESeq2 (version 1.28.1; RRID:SCR_015687) with the quantification of genes without multimapping reads (see the "Bulk RNA-seq analysis: Gene quantification" section).

Each point in the boxplot (Figs. 1E and 4E) refers to a sample. "Antisense" refers to counts of reverse transcription in forward features and counts from forward transcription in reverse features. "Sense" refers to counts of reverse transcription in features annotated in the reverse strand and forward counts in features annotated in the forward strand. Boxplots were produced by summing counts of the same subfamily and strand, per sample, per the direction of transcription (e.g., all L1PA2s in the reverse strand were summed using only the counts from the reverse strand).

Comparing the ratio of detected elements of all L1s. Once normalized for the counts of individual elements by the gene sizeFactors (see the "Comparison between sense and antisense transcription over TEs" section; Figs. 1F and 3C), we defined a "detected" element as an element with a mean of >10 normalized counts in the group of samples of interest. The total number of elements is the number of elements from a particular subfamily annotated in the GTF file that was input to featureCounts (version 1.6.3; RRID:SCR_012919).

Transcription over evolutionary young L1 elements in bulk datasets. The browser extensible data (BED) file version of TEcount's GTF file was used to create BED files containing all L1HS, L1PA2, L1PA3, and L1PA4 elements longer than 6 kbp (full length). These BED files were then split by the strand of the element.

Using the bigwig files of the uniquely mapped BAM files, we created four matrices per dataset using the DeepTools' (version 2.5.4; RRID:SCR_016366) computeMatrix function (*74*)—one for elements annotated in the positive strand using only the bigwig files with forward transcription (transcription in sense of the element), another one for elements annotated in the reverse strand using only bigwig files with reverse transcription (transcription in sense of the element), and another two with the antisense transcription being used (e.g., elements annotated in the positive strand using reverse transcription bigwig files). We then concatenate the matrices of transcription in sense of the elements together using rbind from computeMatrixOperations (*74*). The same operation was performed for the antisense matrices. Heatmaps were plotted using plotHeatmap (*74*), setting missing values to white (--missingDataColor white), and colorMap to blues (sense) or reds (antisense).

To investigate whether the expressed elements contained an intact YY1 binding site, we extracted the relevant sequences using getfasta from BEDTools (version 2.30.0; RRID:SCR_006646) (*75*) using GRCh38.p13 as input fasta (-fi) and forcing strandness (-s). We quantified the number of elements with an exact match to the YY1 binding motif (CAAGATGGCCG) (*76*) in the first 100 bp of the element (see GitHub under src/analysis/yy1_present.py).

PacBio Iso-Seq sample preparation

Total RNA was obtained from tissue samples using miRNA Easy Mini Kit (QIAGEN). RNA samples were subsequently put on dry ice and shipped to the National Genomics Infrastructure of Sweden. There, input quality control of samples was performed on the Agilent Bioanalyzer instrument, using the Eukaryote Total RNA Nano kit (Agilent) to evaluate RNA Integrity Number (RIN) and concentration. The sample libraries were prepared as described in "Procedure & Checklist—Iso-Seq Express Template Preparation for Sequel and Sequel II Systems" (PN-101763800, PacBio, version 02; October 2019) using the NEBNext Single Cell/Low Input cDNA Synthesis & Amplification Module (catalog nos. E6421S for 24 reactions and E6421L for 96 reactions, New England Biolabs), the Iso-Seq Express Oligo Kit (catalog no. PN-101737500, PacBio), ProNex beads [catalog nos. NG2001 (10 ml), NG2002 (125 ml), and NG2003 (500 ml), Promega] and the SMRTbell Express Template Prep Kit 2.0 (catalog no. PN-100938900, PacBio). Total RNA (300 ng) was used for cDNA synthesis, followed by 12 + 3 cycles of cDNA amplification. In the purification step of amplified cDNA, the standard workflow was applied (sample is composed primarily of transcripts centered around 2 kb). After purification, the amplified cDNA went into the SMRTbell library construction. Quality control of the SMRTbell libraries was performed with the Qubit dsDNA HS kit (catalog no. Q32851, Invitrogen) and the Agilent Bioanalyzer High Sensitivity Kit. Primer annealing and polymerase binding were performed using the Sequel II binding kit 2.0 (catalog no. PN-101789500, PacBio). Last, the samples were sequenced on Sequel II and Sequel IIe System using Sequel II Sequencing Plate 2.0, with an on-plate loading concentration of 110 pM, a movie time of 24 hours, and a pre-extension time of 2 hours.

Detail protocol can be found at DOI: dx.doi.org/10.17504/protocols.io.yxmvm25j6g3p/v1. For additional information, please contact the National Genomics Infrastructure of Sweden.

Iso-Seq mapping to L1HS/PA2 consensus sequence

A L1HS and L1PA2 consensus sequence was used to create a minimap2 (version 2.24; RRID:SCR_018550) (*77*) index (minimap2 -d L1consensus.mmi L1consensus.fa) to map full-length nonconcatemer reads (HiFi reads). The density of mapped reads was visualized in the Integrative Genomics Viewer (version 2.12.3; RRID:SCR_011793) (*78*). The number of mapped reads in

the L1s 5′UTR was retrieved using SAMtools view (-c) (version 1.9; RRID:SCR_002105), specifying the first 900 bp of the consensus sequence as the coordinates of interest.

Isolation of NeuN$^+$ cells

Nuclei were isolated from frozen tissue as described above. Before FACSing, nuclei were incubated with recombinant Alexa Fluor 488 anti-NeuN antibody [EPR12763]–neuronal marker (catalog no. ab190195, Abcam, RRID:AB_2716282) at a concentration of 1:500 for 30 min on ice as previously described (*79*). The nuclei were run through the FACS at 4°C with a low flow rate using a 100-mm nozzle, and 300,000 Alexa Fluor 488–positive nuclei were sorted. The sorted nuclei were pelleted at 1300*g* for 15 min and resuspended in 1 ml of ice-cold nuclear wash buffer (20 mM Hepes, 150 mM NaCl, 0.5 mM spermidine, 1× cOmplete protease inhibitors, 0.1% bovine serum albumin) and 10 μl per antibody treatment of ConA-coated magnetic beads (Epicypher) added with gentle vortexing (pipette tips for transferring nuclei were precoated with 1% bovine serum albumin). Protocol can be found at DOI: dx.doi.org/10.17504/protocols.io.4r3l27pejg1y/v1.

CUT&RUN

We followed the protocol detailed by the Henikoff laboratory (*41*). Briefly, 100,000 sorted nuclei were washed twice [20 mM Hepes (pH 7.5), 150 mM NaCl, 0.5 mM spermidine, and 1× Roche cOmplete protease inhibitors] and attached to 10 ConA-coated magnetic beads (Bangs Laboratories) that had been preactivated in binding buffer [20 mM Hepes (pH 7.9), 10 mM KCl, 1 mM $CaCl_2$, and 1 mM $MnCl_2$]. Bead-bound cells were resuspended in 50 μl of buffer [20 mM Hepes (pH 7.5), 0.15 M NaCl, 0.5 mM spermidine, 1× Roche cOmplete protease inhibitors, 0.02% (w/v) digitonin, and 2 mM EDTA] containing primary antibody (rabbit anti-H3K4me3: 39159, Active Motif, RRID:AB_2615077; or goat anti-rabbit immunoglobulin G: ab97047, Abcam, RRID:AB_10681025) at 1:50 dilution and incubated at 4°C overnight with gentle shaking. Beads were washed thoroughly with digitonin buffer [20 mM Hepes (pH 7.5), 150 mM NaCl, 0.5 mM spermidine, 1× Roche cOmplete protease inhibitors, and 0.02% digitonin]. After the final wash, pA-MNase (a gift from S. Henikoff) was added to the digitonin buffer and incubated with the cells at 4°C for 1 hour. Bead-bound cells were washed twice, resuspended in 100 μl of digitonin buffer, and chilled to 0° to 2°C. Genome cleavage was stimulated by the addition of 2 mM $CaCl_2$ at 0°C for 30 min. The reaction was quenched by the addition of 100 μl of 2× stop buffer [0.35 M NaCl, 20 mM EDTA, 4 mM EGTA, 0.02% digitonin, glycogen (50 ng/μl), ribonuclease A (50 ng/μl), yeast spike-in DNA (10 fg/μl; a gift from S. Henikoff)] and vortexing. After 10 min of incubation at 37°C to release genomic fragments, cells and beads were pelleted by centrifugation (16,000*g* for 5 min at 4°C), and fragments from the supernatant were purified. Illumina sequencing libraries were prepared using the HyperPrep Kit (KAPA) (catalog no. 7962347001, Roche) with unique dual-indexed adapters (KAPA) (catalog no. 8278555702, Roche), pooled, and sequenced on a NextSeq500 instrument (Illumina). Detail protocol can be found at DOI: dx.doi.org/10.17504/protocols.io.j8nlkwb8dl5r/v1.

CUT&RUN analysis

Paired-end reads (2 × 75) were aligned to the human genome (hg38) using bowtie2 (version 2.3.4.2; RRID:SCR_016368) (*80*) (–local –very-sensitive-local –no-mixed –no-discordant –phred33 -I 10 -X 700), converted to BAM files with SAMtools (version 1.4; RRID: SCR_002105) and sorted (SAMtools version 1.9; RRID: SCR_002105). Reads per kilobase per million mapped reads (RPKM) normalized bigwig coverage tracks were made with bamCoverage (DeepTools, version 2.5.4; RRID:SCR_016366) (*74*).

Tag directories were created using Homer (version 4.10; RRID: SCR_010881) (*81*) makeTagDirectory on default parameters. Peak calling was performed using findPeaks (Homer), using the option histone as style (-style). The rest of the parameters were left on default options. Peaks were then annotated using the script annotatePeaks.pl (Homer; http://homer.ucsd.edu/homer/ngs/annotation.html) and intersected (BEDtools, version 2.30.0; RRID: SCR_006646) to bed files containing coordinates of >6-kbp L1HS, L1PA2, L1PA3, or L1PA4. Matrices for heatmaps were created (computeMatrix, DeepTools, version 2.5.4; RRID:SCR_016366) using the peaks with an overlap on these elements (only peaks that were called in all samples of a dataset) and visualized using plotHeatmap (DeepTools).

Supplementary Materials

This PDF file includes:

Figs. S1 to S6
Tables S1 to S3
References

REFERENCES AND NOTES

1. R. S. Hill, C. A. Walsh, Molecular insights into human brain evolution. *Nature* **437**, 64–67 (2005).
2. J. H. Lui, D. V. Hansen, A. R. Kriegstein, Development and evolution of the human neocortex. *Cell* **146**, 18–36 (2011).
3. P. Rakic, Evolution of the neocortex: A perspective from developmental biology. *Nat. Rev. Neurosci.* **10**, 724–735 (2009).
4. A. M. M. Sousa, K. A. Meyer, G. Santpere, F. O. Gulden, N. Sestan, Evolution of the human nervous system function, structure, and development. *Cell* **170**, 226–247 (2017).
5. Chimpanzee Sequencing and Analysis Consortium, Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
6. Z. N. Kronenberg, I. T. Fiddes, D. Gordon, S. Murali, S. Cantsilieris, O. S. Meyerson, J. G. Underwood, B. J. Nelson, M. J. P. Chaisson, M. L. Dougherty, K. M. Munson, A. R. Hastie, M. Diekhans, F. Hormozdiari, N. Lorusso, K. Hoekzema, R. Qiu, K. Clark, A. Raja, A. E. Welch, M. Sorensen, C. Baker, R. S. Fulton, J. Armstrong, T. A. Graves-Lindsay, A. M. Denli, E. R. Hoppe, P. Hsieh, C. M. Hill, A. W. C. Pang, J. Lee, E. T. Lam, S. K. Dutcher, F. H. Gage, W. C. Warren, J. Shendure, D. Haussler, V. A. Schneider, H. Cao, M. Ventura, R. K. Wilson, B. Paten, A. Pollen, E. E. Eichler, High-resolution comparative analysis of great ape genomes. *Science* **360**, eaar6343 (2018).
7. S. J. Hoyt, J. M. Storer, G. A. Hartley, P. G. Grady, A. Gershman, L. G. de Lima, C. Limouse, R. Halabian, L. Wojenski, M. Rodriguez, N. Altemose, A. Rhie, L. J. Core, J. L. Gerton, W. Makalowski, D. Olson, J. Rosen, A. F. A. Smit, A. F. Straight, M. R. Vollger, T. J. Wheeler, M. C. Schatz, E. E. Eichler, A. M. Phillippy, W. Timp, K. H. Miga, R. J. O'Neill, From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science* **376**, eabk3112 (2022).
8. R. Cordaux, M. A. Batzer, The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **10**, 691–703 (2009).
9. A. D. Ewing, H. H. Kazazian Jr., High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res.* **20**, 1262–1270 (2010).
10. M. E. Jonsson, R. Garza, P. A. Johansson, J. Jakobsson, Transposable elements: A common feature of neurodevelopmental and neurodegenerative disorders. *Trends Genet.* **36**, 610–623 (2020).
11. H. H. Kazazian Jr., J. V. Moran, Mobile DNA in health and disease. *N. Engl. J. Med.* **377**, 361–370 (2017).
12. O. Deniz, J. M. Frost, M. R. Branco, Regulation of transposable elements by DNA modifications. *Nat. Rev. Genet.* **20**, 417–431 (2019).
13. J. L. Goodier, Restricting retrotransposons: A review. *Mob. DNA* **7**, 16 (2016).

14. M. E. Jönsson, P. Ludvik Brattas, C. Gustafsson, R. Petri, D. Yudovich, K. Pircs, S. Verschuere, S. Madsen, J. Hansson, J. Larsson, R. Mansson, A. Meissner, J. Jakobsson, Activation of neuronal genes via LINE-1 elements upon global DNA demethylation in human neural progenitors. *Nat. Commun.* **10**, 3182 (2019).
15. C. P. Walsh, J. R. Chaillet, T. H. Bestor, Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat. Genet.* **20**, 116–117 (1998).
16. E. B. Chuong, N. C. Elde, C. Feschotte, Regulatory activities of transposable elements: From conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).
17. A. Kapusta, Z. Kronenberg, V. J. Lynch, X. Zhuo, L. Ramsay, G. Bourque, M. Yandell, C. Feschotte, Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLOS Genet.* **9**, e1003470 (2013).
18. J. L. Rinn, H. Y. Chang, Long noncoding RNAs: Molecular modalities to organismal functions. *Annu. Rev. Biochem.* **89**, 283–308 (2020).
19. C. R. Beck, J. L. Garcia-Perez, R. M. Badge, J. V. Moran, LINE-1 elements in structural variation and disease. *Annu. Rev. Genomics Hum. Genet.* **12**, 187–215 (2011).
20. E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki; International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
21. H. Khan, A. Smit, S. Boissinot, Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res.* **16**, 78–87 (2006).
22. Q. Feng, J. V. Moran, H. H. Kazazian Jr., J. D. Boeke, Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**, 905–916 (1996).
23. S. L. Mathias, A. F. Scott, H. H. Kazazian Jr., J. D. Boeke, A. Gabriel, Reverse transcriptase encoded by a human transposable element. *Science* **254**, 1808–1810 (1991).
24. G. D. Swergold, Identification, characterization, and cell specificity of a human LINE-1 promoter. *Mol. Cell. Biol.* **10**, 6718–6729 (1990).
25. A. M. Denli, I. Narvaiza, B. E. Kerman, M. Pena, C. Benner, M. C. Marchetto, J. K. Diedrich, A. Aslanian, J. Ma, J. J. Moresco, L. Moore, T. Hunter, A. Saghatelian, F. H. Gage, Primate-specific ORF0 contributes to retrotransposon-mediated diversity. *Cell* **163**, 583–593 (2015).
26. M. Speek, Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Mol. Cell. Biol.* **21**, 1973–1985 (2001).
27. N. G. Coufal, J. L. Garcia-Perez, G. E. Peng, G. W. Yeo, Y. Mu, M. T. Lovci, M. Morell, K. S. O'Shea, J. V. Moran, F. H. Gage, L1 retrotransposition in human neural progenitor cells. *Nature* **460**, 1127–1131 (2009).
28. J. A. Erwin, A. C. Paquola, T. Singer, I. Gallina, M. Novotny, C. Quayle, T. A. Bedrosian, F. I. Alves, C. R. Butcher, J. R. Herdy, A. Sarkar, R. S. Lasken, A. R. Muotri, F. H. Gage, L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nat. Neurosci.* **19**, 1583–1591 (2016).
29. G. D. Evrony, X. Cai, E. Lee, L. B. Hills, P. C. Elhosary, H. S. Lehmann, J. J. Parker, K. D. Atabay, E. C. Gilmore, A. Poduri, P. J. Park, C. A. Walsh, Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483–496 (2012).
30. G. D. Evrony, E. Lee, B. K. Mehta, Y. Benjamini, R. M. Johnson, X. Cai, L. Yang, P. Haseley, H. S. Lehmann, P. J. Park, C. A. Walsh, Cell lineage analysis in human brain using endogenous retroelements. *Neuron* **85**, 49–59 (2015).
31. A. R. Muotri, V. T. Chu, M. C. N. Marchetto, W. Deng, J. V. Moran, F. H. Gage, Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**, 903–910 (2005).
32. F. J. Sanchez-Luque, M.-J. H. C. Kempen, P. Gerdes, D. B. Vargas-Landin, S. R. Richardson, R.-L. Troskie, J. S. Jesuadian, S. W. Cheetham, P. E. Carreira, C. Salvador-Palomeque, M. Garcia-Cañadas, M. Muñoz-Lopez, L. Sanchez, M. Lundberg, A. Macia, S. R. Heras, P. M. Brennan, R. Lister, J. L. Garcia-Perez, A. D. Ewing, G. J. Faulkner, LINE-1 evasion of epigenetic repression in humans. *Mol. Cell* **75**, 590–604.e12 (2019).
33. K. R. Upton, D. J. Gerhardt, J. S. Jesuadian, S. R. Richardson, F. J. Sanchez-Luque, G. O. Bodea, A. D. Ewing, C. Salvador-Palomeque, M. S. van der Knaap, P. M. Brennan, A. Vanderver, G. J. Faulkner, Ubiquitous L1 mosaicism in hippocampal neurons. *Cell* **161**, 228–239 (2015).
34. S. Lanciano, G. Cristofari, Measuring and interpreting transposable element expression. *Nat. Rev. Genet.* **21**, 721–736 (2020).
35. Y. Jin, O. H. Tam, E. Paniagua, M. Hammell, *TEtranscripts*: A package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**, 3593–3599 (2015).
36. M. E. Jonsson, R. Garza, Y. Sharma, R. Petri, E. Sodersten, J. G. Johansson, P. A. Johansson, D. A. Atacho, K. Pircs, S. Madsen, D. Yudovich, R. Ramakrishnan, J. Holmberg, J. Larsson, P. Jern, J. Jakobsson, Activation of endogenous retroviruses during brain development causes an inflammatory response. *EMBO J.* **40**, e106423 (2021).
37. V. P. Belancio, M. Whelton, P. Deininger, Requirements for polyadenylation at the 3′ end of LINE-1 elements. *Gene* **390**, 98–107 (2007).
38. E. M. Ostertag, H. H. Kazazian Jr., Twin priming: A proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res.* **11**, 2059–2065 (2001).
39. E. K. Gustavsson, S. Sethi, Y. Gao, J. W. Brenton, S. García-Ruiz, D. Zhang, R. Garza, R. H. Reynolds, J. R. Evans, Z. Chen, M. Grant-Peters, H. Macpherson, K. Montgomery, R. Dore, A. I. Wernick, C. Arber, S. Wray, S. Gandhi, J. Esselborn, C. Blauwendraat, C. H. Douse, A. Adami, D. A. M. Atacho, A. Kouli, A. Quaegebeur, R. A. Barker, E. Englund, F. Platt, J. Jakobsson, N. W. Wood, H. Houlden, H. Saini, C. F. Bento, J. Hardy, M. Ryten, The annotation and function of the Parkinson's and Gaucher disease-linked gene GBA1 has been concealed by its protein-coding pseudogene GBAP1 (Cold Spring Harbor Laboratory, 2022).
40. J. Feusier, W. S. Watkins, J. Thomas, A. Farrell, D. J. Witherspoon, L. Baird, H. Ha, J. Xing, L. B. Jorde, Pedigree-based estimation of human mobile element retrotransposition rates. *Genome Res.* **29**, 1567–1577 (2019).
41. P. J. Skene, S. Henikoff, An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife* **6**, e21856 (2017).
42. B. Brouha, J. Schustak, R. M. Badge, S. Lutz-Prigge, A. H. Farley, J. V. Moran, H. H. Kazazian Jr., Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 5280–5285 (2003).
43. G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lonnerberg, A. Furlan, J. Fan, L. E. Borm, Z. Liu, D. van Bruggen, J. Guo, X. He, R. Barker, E. Sundstrom, G. Castelo-Branco, P. Cramer, I. Adameyko, S. Linnarsson, P. V. Kharchenko, RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
44. M. Astick, P. Vanderhaeghen, From human pluripotent stem cells to cortical circuits. *Curr. Top. Dev. Biol.* **129**, 67–98 (2018).
45. C. Philippe, D. B. Vargas-Landin, A. J. Doucet, D. van Essen, J. Vera-Otarola, M. Kuciak, A. Corbin, P. Nigumann, G. Cristofari, Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. *eLife* **5**, e13926 (2016).
46. P. A. Johansson, P. L. Brattas, C. H. Douse, P. Hsieh, A. Adami, J. Pontis, D. Grassi, R. Garza, E. Sozzi, R. Cataldo, M. E. Jonsson, D. A. M. Atacho, K. Pircs, F. Eren, Y. Sharma, J. Johansson, A. Fiorenzano, M. Parmar, M. Fex, D. Trono, E. E. Eichler, J. Jakobsson, A *cis*-acting structural variation at the ZNF558 locus controls a gene regulatory network in human brain development. *Cell Stem Cell* **29**, 52–69.e8 (2022).
47. S. B. Linker, I. Narvaiza, J. Y. Hsu, M. Wang, F. Qiu, A. P. D. Mendes, R. Oefner, K. Kottilil, A. Sharma, L. Randolph-Moore, E. Mejia, R. Santos, M. C. Marchetto, F. H. Gage, Human-specific regulation of neural maturation identified by cross-primate transcriptomics. *Curr. Biol.* **32**, 4797–4807.e5 (2022).

48. S. Kanton, M. J. Boyle, Z. He, M. Santel, A. Weigert, F. Sanchis-Calleja, P. Guijarro, L. Sidow, J. S. Fleck, D. Han, Z. Qian, M. Heide, W. B. Huttner, P. Khaitovich, S. Paabo, B. Treutlein, J. G. Camp, Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* **574**, 418–422 (2019).
49. Y. Mao, W. T. Harvey, D. Porubsky, K. M. Munson, K. Hoekzema, A. P. Lewis, P. A. Audano, A. Rozanski, X. Yang, S. Zhang, D. S. Gordon, X. Wei, G. A. Logsdon, M. Haukness, P. C. Dishuck, H. Jeong, R. Del Rosario, V. L. Bauer, W. T. Fattor, G. K. Wilkerson, Q. Lu, B. Paten, G. Feng, S. L. Sawyer, W. C. Warren, L. Carbone, E. E. Eichler, Structurally divergent and recurrently mutated regions of primate genomes (Cold Spring Harbor Laboratory, 2023).
50. S. Banfi, A. Servadio, M. Y. Chung, T. J. Kwiatkowski Jr., A. E. McCall, L. A. Duvick, Y. Shen, E. J. Roth, H. T. Orr, H. Y. Zoghbi, Identification and characterization of the gene causing type 1 spinocerebellar ataxia. *Nat. Genet.* **7**, 513–520 (1994).
51. J. M. Dewing, R. O. Carare, A. J. Lotery, J. A. Ratnayaka, The diverse roles of TIMP-3: Insights into degenerative diseases of the senescent retina and brain. *Cells* **9**, 39 (2019).
52. M. A. Lancaster, M. Renner, C. A. Martin, D. Wenzel, L. S. Bicknell, M. E. Hurles, T. Homfray, J. M. Penninger, A. P. Jackson, J. A. Knoblich, Cerebral organoids model human brain development and microcephaly. *Nature* **501**, 373–379 (2013).
53. V. P. Belancio, A. M. Roy-Engel, R. R. Pochampally, P. Deininger, Somatic expression of LINE-1 elements in human tissues. *Nucleic Acids Res.* **38**, 3909–3922 (2010).
54. G. J. Faulkner, Y. Kimura, C. O. Daub, S. Wani, C. Plessy, K. M. Irvine, K. Schroder, N. Cloonan, A. L. Steptoe, T. Lassmann, K. Waki, N. Hornig, T. Arakawa, H. Takahashi, J. Kawai, A. R. Forrest, H. Suzuki, Y. Hayashizaki, D. A. Hume, V. Orlando, S. M. Grimmond, P. Carninci, The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.* **41**, 563–571 (2009).
55. S. Horvath, DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).
56. M. De Cecco, T. Ito, A. P. Petrashen, A. E. Elias, N. J. Skvir, S. W. Criscione, A. Caligiana, G. Brocculi, E. M. Adney, J. D. Boeke, O. Le, C. Beausejour, J. Ambati, K. Ambati, M. Simon, A. Seluanov, V. Gorbunova, P. E. Slagboom, S. L. Helfand, N. Neretti, J. M. Sedivy, L1 drives IFN in senescent cells and promotes age-associated inflammation. *Nature* **566**, 73–78 (2019).
57. M. Van Meter, M. Kashyap, S. Rezazadeh, A. J. Geneva, T. D. Morello, A. Seluanov, V. Gorbunova, SIRT6 represses LINE1 retrotransposons by ribosylating KAP1 but this repression fails with stress and age. *Nat. Commun.* **5**, 5011 (2014).
58. D. Ardeljan, J. P. Steranka, C. Liu, Z. Li, M. S. Taylor, L. M. Payer, M. Gorbounov, J. S. Sarnecki, V. Deshpande, R. H. Hruban, J. D. Boeke, D. Fenyo, P. H. Wu, A. Smogorzewska, A. J. Holland, K. H. Burns, Cell fitness screens reveal a conflict between LINE-1 retrotransposition and DNA replication. *Nat. Struct. Mol. Biol.* **27**, 168–178 (2020).
59. N. R. Wray, S. Ripke, M. Mattheisen, M. Trzaskowski, E. M. Byrne, A. Abdellaoui, M. J. Adams, E. Agerbo, T. M. Air, T. M. F. Andlauer, S.-A. Bacanu, M. Bækvad-Hansen, A. F. T. Beekman, T. B. Bigdeli, E. B. Binder, D. R. H. Blackwood, J. Bryois, H. N. Buttenschøn, J. Bybjerg-Grauholm, N. Cai, E. Castelao, J. H. Christensen, T.-K. Clarke, J. I. R. Coleman, L. Colodro-Conde, B. Couvy-Duchesne, N. Craddock, G. E. Crawford, C. A. Crowley, H. S. Dashti, G. Davies, I. J. Deary, F. Degenhardt, E. M. Derks, N. Direk, C. V. Dolan, E. C. Dunn, T. C. Eley, N. Eriksson, V. Escott-Price, F. H. F. Kiadeh, H. K. Finucane, A. J. Forstner, J. Frank, H. A. Gaspar, M. Gill, P. Giusti-Rodríguez, F. S. Goes, S. D. Gordon, J. Grove, L. S. Hall, E. Hannon, C. S. Hansen, T. F. Hansen, S. Herms, I. B. Hickie, P. Hoffmann, G. Homuth, C. Horn, J.-J. Hottenga, D. M. Hougaard, M. Hu, C. L. Hyde, M. Ising, R. Jansen, F. Jin, E. Jorgenson, J. A. Knowles, I. S. Kohane, J. Kraft, W. W. Kretzschmar, J. Krogh, Z. Kutalik, J. M. Lane, Y. Li, Y. Li, P. A. Lind, X. Liu, L. Lu, D. J. MacIntyre, D. F. MacKinnon, R. M. Maier, W. Maier, J. Marchini, H. Mbarek, P. M. Grath, P. M. Guffin, S. E. Medland, D. Mehta, C. M. Middeldorp, E. Mihailov, Y. Milaneschi, L. Milani, J. Mill, F. M. Mondimore, G. W. Montgomery, S. Mostafavi, N. Mullins, M. Nauck, B. Ng, M. G. Nivard, D. R. Nyholt, P. F. O'Reilly, H. Oskarsson, M. J. Owen, J. N. Painter, C. B. Pedersen, M. G. Pedersen, R. E. Peterson, E. Pettersson, W. J. Peyrot, G. Pistis, D. Posthuma, S. M. Purcell, J. A. Quiroz, P. Qvist, J. P. Rice, B. P. Riley, M. Rivera, S. S. Mirza, R. Saxena, R. Schoevers, E. C. Schulte, L. Shen, J. Shi, S. I. Shyn, E. Sigurdsson, G. B. C. Sinnamon, J. H. Smit, D. J. Smith, H. Stefansson, S. Steinberg, C. A. Stockmeier, F. Streit, J. Strohmaier, K. E. Tansey, H. Teismann, A. Teumer, W. Thompson, P. A. Thomson, T. E. Thorgeirsson, C. Tian, M. Traylor, J. Treutlein, V. Trubetskoy, A. G. Uitterlinden, D. Umbricht, S. Van der Auwera, A. M. van Hemert, A. Viktorin, P. M. Visscher, Y. Wang, B. T. Webb, S. M. Weinsheimer, J. Wellmann, G. Willemsen, S. H. Witt, Y. Wu, H. S. Xi, J. Yang, F. Zhang; eQTLGen; 23andMe, V. Arolt, B. T. Baune, K. Berger, D. I. Boomsma, S. Cichon, U. Dannlowski, E. C. J. de Geus, J. R. De Paulo, E. Domenici, K. Domschke, T. Esko, H. J. Grabe, S. P. Hamilton, C. Hayward, A. C. Heath, D. A. Hinds, K. S. Kendler, S. Kloiber, G. Lewis, Q. S. Li, S. Lucae, P. F. A. Madden, P. K. Magnusson, N. G. Martin, A. M. McIntosh, A. Metspalu, O. Mors, P. B. Mortensen, B. Müller-Myhsok, M. Nordentoft, M. M. Nöthen, M. C. O'Donovan, S. A. Paciga, N. L. Pedersen, B. W. J. H. Penninx, R. H. Perlis, D. J. Porteous, J. B. Potash, M. Preisig, M. Rietschel, C. Schaefer, T. G. Schulze, J. W. Smoller, K. Stefansson, H. Tiemeier, R. Uher, H. Völzke, M. M. Weissman, T. Werge, A. R. Winslow, C. M. Lewis, D. F. Levinson, G. Breen, A. D. Børglum, P. F. Sullivan; Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).
60. S. Benito-Kwiecinski, S. L. Giandomenico, M. Sutcliffe, E. S. Riis, P. Freire-Pritchett, I. Kelava, S. Wunderlich, U. Martin, G. A. Wray, K. McDole, M. A. Lancaster, An early cell shape transition drives evolutionary expansion of the human forebrain. *Cell* **184**, 2084–2102. e19 (2021).
61. M. C. Marchetto, B. Hrvoj-Mihic, B. E. Kerman, D. X. Yu, K. C. Vadodaria, S. B. Linker, I. Narvaiza, R. Santos, A. M. Denli, A. P. Mendes, R. Oefner, J. Cook, L. McHenry, J. M. Grasmick, K. Heard, C. Fredlender, L. Randolph-Moore, R. Kshirsagar, R. Xenitopoulos, G. Chou, N. Hah, A. R. Muotri, K. Padmanabhan, K. Semendeferi, F. H. Gage, Species-specific maturation profiles of human, chimpanzee and bonobo neural cells. *eLife* **8**, e37527 (2019).
62. P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H. Fritz, M. K. Konkel, A. Malhotra, A. M. Stutz, X. Shi, F. P. Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. J. P. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Y. K. Lam, X. J. Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. M. Kidd, Y. Kong, E. W. Lameijer, S. McCarthy, P. Flicek, R. A. Gibbs, G. Marth, C. E. Mason, A. Menelaou, D. M. Muzny, B. J. Nelson, A. Noor, N. F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E. E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalin, A. Untergasser, J. A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M. A. Batzer, S. A. McCarroll; 1000 Genomes Project Consortium, R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler, J. O. Korbel, An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
63. D. A. Grassi, P. L. Brattas, M. E. Jönsson, D. Atacho, O. Karlsson, S. Nolbrant, M. Parmar, J. Jakobsson, Profiling of lincRNAs in human pluripotent stem cell derived forebrain neural progenitor cells. *Heliyon* **6**, e03067 (2020).
64. S. Nolbrant, A. Heuer, M. Parmar, A. Kirkeby, Generation of high-purity human ventral midbrain dopaminergic progenitors for in vitro maturation and intracerebral transplantation. *Nat. Protoc.* **12**, 1962–1979 (2017).
65. R. Zufferey, D. Nagy, R. J. Mandel, L. Naldini, D. Trono, Multiply attenuated lentiviral vector achieves efficient gene delivery in vivo. *Nat. Biotechnol.* **15**, 871–875 (1997).
66. K. J. Livak, T. D. Schmittgen, Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods* **25**, 402–408 (2001).
67. G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. Mcdermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. Mcfarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, J. H. Bielas, Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
68. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
69. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
70. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
71. Y. Liao, G. K. Smyth, W. Shi, featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
72. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with *DESeq2*. *Genome Biol.* **15**, 550 (2014).
73. M. Pertea, G. M. Pertea, C. M. Antonescu, T. C. Chang, J. T. Mendell, S. L. Salzberg, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
74. F. Ramírez, F. Dündar, S. Diehl, B. A. Grüning, T. Manke, DeepTools: A flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
75. A. Quinlan, I. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
76. J. N. Athanikar, R. M. Badge, J. V. Moran, A YY1-binding site is required for accurate human LINE-1 transcription initiation. *Nucleic Acids Res.* **32**, 3846–3855 (2004).
77. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
78. J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, J. P. Mesirov, Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

79. K. L. Spalding, O. Bergmann, K. Alkass, S. Bernard, M. Salehpour, H. B. Huttner, E. Bostrom, I. Westerlund, C. Vial, B. A. Buchholz, G. Possnert, D. C. Mash, H. Druid, J. Frisen, Dynamics of hippocampal neurogenesis in adult humans. *Cell* **153**, 1219–1227 (2013).
80. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
81. S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, C. K. Glass, Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
82. R. Garza, Y. Sharma, D. Atacho, S. A. Hamdeh, M. Jönsson, M. Ingelsson, P. Jern, M. G. Hammell, E. Englund, J. Jakobsson, N. Marklund, Single-cell transcriptomics of resected human traumatic brain injury tissues reveals acute activation of endogenous retroviruses in oligodendroglia (Cold Spring Harbor Laboratory, 2022).
83. E. K. Gustavsson, S. Sethi, Y. Gao, J. Brenton, S. García-Ruiz, D. Zhang, R. Garza, R. H. Reynolds, J. R. Evans, Z. Chen, M. Grant-Peters, H. Macpherson, K. Montgomery, R. Dore, A. I. Wernick, C. Arber, S. Wray, S. Gandhi, J. Esselborn, C. Blauwendraat, C. H. Douse, A. Adami, D. A. M. Atacho, A. Kouli, A. Quaegebeur, R. A. Barker, E. Englund, F. Platt, J. Jakobsson, N. W. Wood, H. Houlden, H. Saini, C. F. Bento, J. Hardy, M. Ryten, Pseudogenes limit the identification of novel common transcripts generated by their parent genes (Cold Spring Harbor Laboratory, 2022).
84. M. Karimzadeh, C. Ernst, A. Kundaje, M. M. Hoffman, Umap and Bismap: Quantifying genome and methylome mappability. *Nucleic Acids Res.* **46**, e120 (2018).

Acknowledgments: We would like to thank J. Frisén, D. Trono, F. H. Gage, W. Huttner, S. Pääbo, N. Marklund, and S. Henikoff for comments and support. We also thank U. Jarl, M. P. Vejgården, and A. Hammarberg for technical assistance. We are grateful to all members of the Jakobsson laboratory. For the purpose of open access, we have applied a CC BY public copyright license to all author-accepted manuscripts arising from this submission. **Funding:** The work was supported by grants from the Aligning Science Across Parkinson's (ASAP-000520 to J.J.) through the Michael J. Fox Foundation for Parkinson's Research, the Swedish Research Council (2018-02694 to J.J., 2020-01660 to Z.K, and 2021-03494 to C.H.D.), the Swedish Brain Foundation (FO2019-0098 to J.J. and FO2022-0079 to Z.K.), Cancerfonden (190326 to J.J.), Barncancerfonden (PR2017-0053 to J.J.), NIHR Cambridge Biomedical Research Centre (NIHR203312 to R.A.B), the Swedish Society for Medical Research (S19-0100 to C.H.D.), National Institutes of Health (HG002385 to E.E.E. and K99HG011041 to P.H.), and the Swedish Government Initiative for Strategic Research Areas (MultiPark & StemTherapy). **Author contributions:** Design and interpretation: All authors. Conceptualization: R.G., D.A.M.A., and J.J. Experimental research: D.A.M.A., A.A., P.G., O.K., V.H., P.A.J., M.V., J.M.-F., M.E.J., E.M., and C.H.D. Bioinformatics: R.G., D.A.M.A., N.P., P.H., Y.S., and C.H.D. Material, reagents, and expertise: A.Q., A. K., E.E., E.E.E., M.H., R.A.B., and Z.K. Writing—original draft: R.G., D.A.M.A., and J.J. Writing—review and editing: All authors. **Competing interests:** The authors declare that they have no competing interests. E.E.E. is a member of the scientific advisory board of Variant Bio Inc. and an investigator of the Howard Hughes Medical Institute. M.G.H. is a member of the scientific advisory board of Transposon Therapeutics. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The sequencing data presented in this study have been deposited at GEOs: GSE225081: bulk, short and long read, RNA-seq of adult samples. We also included Cell Ranger's matrices that were used for this paper from the 5′ 10x snRNA-seq (raw data have been previously published at GSE211870); GSE224747: 3′ snRNA-seq, CUT&RUN, and bulk RNA-seq of fetal samples; GSE224659: CRISPRi in fbNPCs and organoids. Additional repositories from publicly available data used in the study: GSE209552: 3′ snRNA-seq of adult samples (*82*); GSE211870: 5′ snRNA-seq of adult samples (*83*); GSE211871, adult $NeuN^+$ CUT&RUN (*83*); GSE182224: chimpanzee and human fbNPCs (*46*); NCBI BioProject accession numbers PRJNA941352-55, PRJNA941358-59, and PRJNA941362-65: long-read sequencing of primate genomes (*49*). Original code has been deposited at GitHub and is publicly available at https://github.com/raquelgarza/truster.git (doi:10.5281/zenodo.7589548) and https://github.com/raquelgarza/L1_transcriptional_complexity_Garza2023.git (doi:https://doi.org/10.5281/zenodo.8116135).

Submitted 24 March 2023
Accepted 29 September 2023
Published 1 November 2023
10.1126/sciadv.adh9543

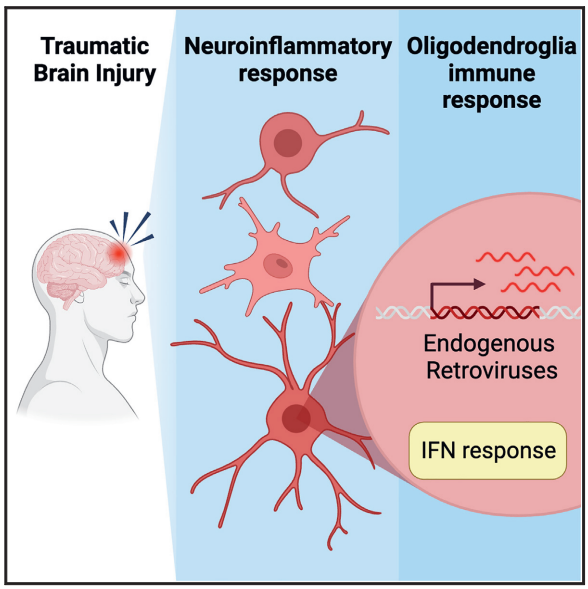
PAPER III

Report

Cell Reports

Single-cell transcriptomics of human traumatic brain injury reveals activation of endogenous retroviruses in oligodendroglia

Graphical abstract



Highlights

- Unique snRNA-seq analysis of human TBI tissue
- Activation of an interferon response in oligodendroglia after TBI
- Expression of endogenous retroviruses in TBI oligodendroglia

Authors

Raquel Garza, Yogita Sharma, Diahann A.M. Atacho, ..., Agnete Kirkeby, Johan Jakobsson, Niklas Marklund

Correspondence

johan.jakobsson@med.lu.se

In brief

From transcriptional profiling of human brain tissue after acute TBI, Garza et al. find an oligodendroglia-specific innate immune response, correlated with transcriptional activation of endogenous retroviruses in the same cell types. The results provide insight into the beginning of human neuroinflammation and implicate endogenous retroviruses in this process.

Garza et al., 2023, Cell Reports ■■, 113395 ■■, 2023 © 2023 The Author(s).
https://doi.org/10.1016/j.celrep.2023.113395

Cell Reports

CellPress
OPEN ACCESS

Report

Single-cell transcriptomics of human traumatic brain injury reveals activation of endogenous retroviruses in oligodendroglia

Raquel Garza,[1] Yogita Sharma,[1] Diahann A.M. Atacho,[1] Arun Thiruvalluvan,[2] Sami Abu Hamdeh,[3] Marie E. Jönsson,[1] Vivien Horvath,[1] Anita Adami,[1] Martin Ingelsson,[4,5,6] Patric Jern,[7] Molly Gale Hammell,[8,9] Elisabet Englund,[10] Agnete Kirkeby,[2,11] Johan Jakobsson,[1,13,14,*] and Niklas Marklund[12,13]

[1]Laboratory of Molecular Neurogenetics, Department of Experimental Medical Science, Wallenberg Neuroscience Center and Lund Stem Cell Center, Lund University, 221 84 Lund, Sweden
[2]Novo Nordisk Foundation Center for Stem Cell Medicine (reNEW) and Department of Neuroscience, University of Copenhagen, Copenhagen, Denmark
[3]Department of Neuroscience, Neurosurgery, Uppsala University, Uppsala, Sweden
[4]Department of Public Health and Caring Sciences, Molecular Geriatrics, Rudbeck Laboratory, Uppsala University, Uppsala, Sweden
[5]Tanz Centre for Research in Neurodegenerative Diseases, Departments of Medicine and Laboratory Medicine & Pathobiology, University of Toronto, Toronto, ON, Canada
[6]Krembil Brain Institute, University Health Network, Toronto, ON, Canada
[7]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden
[8]Institute for Systems Genetics, Department of Neuroscience and Physiology, NYU Langone Health, New York, NY 10016, USA
[9]Neuroscience Institute, NYU Grossman School of Medicine, New York, NY 10003, USA
[10]Department of Clinical Sciences Lund, Division of Pathology, Lund University, Lund, Sweden
[11]Department of Experimental Medical Science, Wallenberg Center for Molecular Medicine and Lund Stem Cell Center, Lund University, 221 84 Lund, Sweden
[12]Department of Clinical Sciences Lund, Neurosurgery, Lund University, Skåne University Hospital, Lund, Sweden
[13]Senior author
[14]Lead contact
*Correspondence: johan.jakobsson@med.lu.se
https://doi.org/10.1016/j.celrep.2023.113395

SUMMARY

Traumatic brain injury (TBI) is a leading cause of chronic brain impairment and results in a robust, but poorly understood, neuroinflammatory response that contributes to the long-term pathology. We used single-nuclei RNA sequencing (snRNA-seq) to study transcriptomic changes in different cell populations in human brain tissue obtained acutely after severe, life-threatening TBI. This revealed a unique transcriptional response in oligodendrocyte precursors and mature oligodendrocytes, including the activation of a robust innate immune response, indicating an important role for oligodendroglia in the initiation of neuroinflammation. The activation of an innate immune response correlated with transcriptional upregulation of endogenous retroviruses in oligodendroglia. This observation was causally linked *in vitro* using human glial progenitors, implicating these ancient viral sequences in human neuroinflammation. In summary, this work provides insight into the initiating events of the neuroinflammatory response in TBI, which has therapeutic implications.

INTRODUCTION

Traumatic brain injury (TBI) caused by, for example, motor vehicle accidents, violence, or falls is a leading cause of mortality and persisting morbidity. The impact to the head results in immediate death to neuronal and glial cells, as well as injury to axons and the microvasculature. This initial primary brain injury is then substantially exacerbated by a poorly understood and progressive cascade of secondary events that may ultimately result in severe long-term consequences. This chronic phase of the injury is often characterized by persistent white matter atrophy, linked to neuroinflammation,[1,2] and an increased risk of neurodegenerative disorders, such as Alzheimer's or Parkinson's disease.[3,4]

Clinical and experimental evidence suggests that a robust neuroinflammatory response plays a key role in the subsequent development of post-traumatic disability and the increased risk of chronic neurodegenerative disorders.[2,5–7] The neuroinflammatory cascade after TBI is complex and involves the activation of resident microglia as well the recruitment of peripheral immune cells. In addition, other resident brain cells such as astrocytes, oligodendrocyte precursor cells (OPCs), and mature oligodendrocytes may be directly involved in modulating or driving the neuroinflammatory response, although their role remains poorly understood.[8–10]

Myelinating oligodendrocytes are essential for saltatory nerve conduction in the central nervous system and are involved in

© 2023 The Author(s).
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

the metabolic support of neurons and the modulation of neuronal excitability.[11] There is also emerging evidence that brain oligodendrocytes are morphologically and transcriptionally heterogeneous and that their functional role may be altered in disease.[12,13] The neuroinflammatory response occurring after TBI seems to be associated with oligodendroglia vulnerability,[9] and persistent neuroinflammation is often found in atrophied white matter tracts of long-term TBI survivors.[2] These observations argue for an important contribution of the neuroinflammatory response to white matter injury. Notably, OPCs and oligodendrocytes have recently been shown to activate the transcription of immune-response genes in certain disease contexts,[14,15] suggesting that these cells may play a direct role in the neuroinflammatory process by adopting an immune-like cell state. Despite these observations, the role of oligodendroglia in the neuroinflammatory response occurring after TBI remains poorly understood.

While the presence of neuroinflammation after TBI, and in other neurodegenerative conditions, has been firmly established, the molecular mechanisms contributing to this sterile inflammatory response remains largely unknown. We and others have recently found that transcriptional activation of endogenous retroviruses (ERVs) or other transposable elements (TEs) are involved in the neuroinflammatory response in mouse models.[16,17] ERVs are remnants of old retrovirus infections that have entered our germline and make up about 8% of our genome. Recent evidence suggests that ERVs can be transcriptionally activated in certain neurological disease states.[16] This aberrant expression of ERVs results in the formation of double-stranded RNAs, reverse-transcribed DNA molecules, and ERV-derived peptides that induce a "viral mimicry," where cells of the central nervous system respond by activating the innate immune system in the form of an interferon response, as if they were infected.[18–20] This event results in a downstream trigger or boost of inflammation that may participate in the respective disease processes. If and how ERVs are activated in human brain disorders and after insults, such as in TBI, remain poorly documented.

In this study, we performed single-nuclei RNA sequencing (snRNA-seq) on fresh-frozen brain tissue, which had been surgically removed due to life-threatening TBI. We found that TBI resulted in a unique transcriptional response in several cell types in the injured human brain tissue, including the activation of cell-cycle-related genes in microglia as well as alterations in synaptic gene expression in excitatory neurons. Notably, we detected clear evidence of an innate immune response in both OPCs and oligodendrocytes, including evidence for an interferon response. This innate immune response was accompanied by the transcriptional activation of major histocompatibility complex (MHC) classes I and II. These results demonstrate that oligodendroglia undergo a transformation to an immune-like cell state after TBI and suggest a key role for these cells in the initiation of neuroinflammation following such an insult. Moreover, the activation of the innate immune response was linked to transcriptional activation of ERVs in OPCs and oligodendrocytes, a phenomenon that could be modeled and replicated in human glial progenitors *in vitro*, implicating a potential role for these ancient viral sequences in human neuroinflammation. These results not only provide insights into the initiating events of a neuroinflammatory response following severe TBI but also open therapeutic avenues.

RESULTS

Samples and experimental design

To investigate cell-type-specific transcriptional responses after severe, acute TBI, we performed snRNA-seq from fresh-frozen human brain tissue (Figure 1A). We recruited 12 patients with severe TBI, defined as having a post-resuscitation Glasgow Coma Scale score ≤ 8. Detailed demographic and clinical characteristics are shown in Table S1. The age (mean $\pm$ standard deviation) of the patients (10 males, 2 females) was 49.5 $\pm$ 18.2 years. In these patients, decompressive surgery was a life-saving measure to remove injured and swollen space-occupying brain tissue causing marked mass effect or increased intracranial pressure (ICP) refractory to conservative, medical neurointensive care treatment. The injured and contused brain regions (typically the injured part of a temporal or frontal lobe) were surgically removed between 4 h and 8 days after injury.[21] As control tissue, we used five fresh-frozen post-mortem samples from the frontal and temporal lobes obtained from three non-neurological deaths of patients aged 69, 75, and 87 years (Table S1).

We isolated nuclei from frozen human brain tissue and performed snRNA-seq using the 10× Genomics pipeline. The tissue obtained from TBI patients contained a high degree of blood and damaged cells, which greatly influenced the quality of sequencing data. Given the quality differences between the control and TBI samples, we used a strict threshold to select cells from both conditions to avoid bias due to differences in sequencing depth: only cells with at least 1,000 detected genes were used for further analysis (Figure S1A). After quality control (QC),[22] we kept high-quality sequencing data from the 12 TBI samples, with a total of 6,806 nuclei and a mean of 2,361 genes per nucleus. From the five control samples, we obtained 8,304 nuclei with a mean of 3,389 genes per nucleus (Figure 1A).

Cell-type composition

Using the snRNA-seq data, we performed an unbiased clustering using Seurat to identify and quantify the different cell types present in the injured brain tissue. We detected 15 different clusters and used canonical marker gene expression[17] to identify the different cell types (Figures 1B–1D; Table S2). We identified three clusters of excitatory neurons and four clusters of inhibitory neurons, as well as clusters with oligodendrocytes, OPCs, astrocytes, endothelial cells, and microglia/macrophages (Figures 1C, 1D, and S1B). All clusters were represented in both TBI and control samples, and the cellular composition was similar between samples obtained from temporal and frontal brain regions (Figures 1D, S1C, and S1D).

The most abundant cell type in the control samples was excitatory neurons, which accounted for almost 50% of the cells, followed by interneurons (20%) and astrocytes (13%) (Figure S1C). Oligodendrocytes made up 9% of the cells, while microglia represented 2%. In the TBI samples, we found a reduction in the proportion of excitatory neurons, which corresponded to 29% of the cells. In contrast, the proportions of both oligodendrocytes and microglia were higher (25% and 7%, respectively). The higher

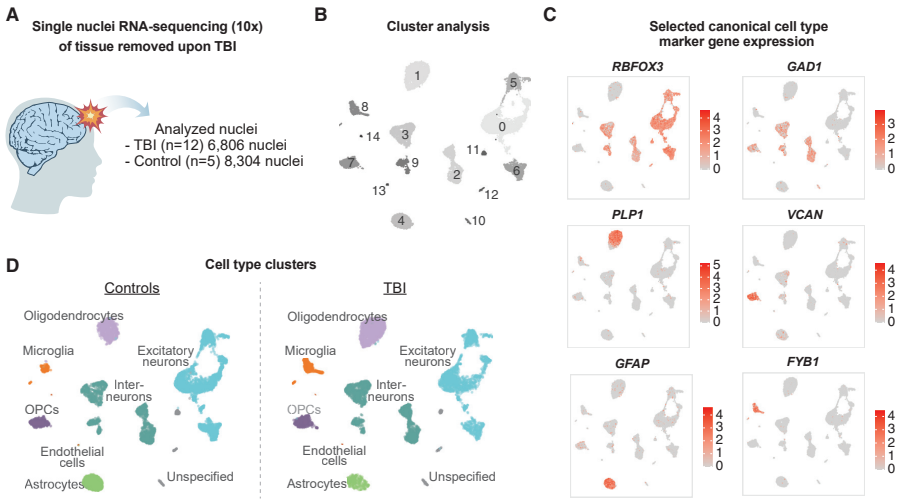


Figure 1. Single-nuclei RNA-seq of human TBI tissue

(A) Schematic of experimental approach. Brain tissue was surgically removed after severe TBI, followed by single-nuclei RNA-seq. Numbers indicate the number of nuclei recovered after quality control per condition.

(B) Uniform manifold approximation and projection (UMAP) labeled with the nuclei clusters identified (clusters 0–14).

(C) Projection of gene expression of canonical gene markers to identify cell types.

(D) UMAP labeled by characterized cell types, split by condition.

percentage of oligodendrocytes is likely explained by an excess of white matter tissue in two of the TBI samples (Figure S1E)—an issue that is difficult to control for in this clinical setting. However, the reduced numbers of excitatory neurons as well as the increase in microglia are likely to be due to the direct, acute consequences of the injury in line with the expected pathological outcome after severe TBI (Figure S1C).

OPCs and oligodendrocytes display an interferon response after acute TBI

The neuroinflammatory response is thought to be a key mechanism in the subsequent pathological processes following TBI. However, how this process starts and what cell types are involved in humans is unknown. To investigate transcriptional changes linked to inflammation, we analyzed cell-type-specific transcriptional alterations after TBI. We found that each cell type displayed a distinct transcriptional response to TBI, while housekeeping genes remained unaltered between conditions (Figures S1F, S1G, and S2A–S2D; Tables S3 and S4). For example, in excitatory neurons, we found that genes linked to synaptic functions, such as NPTX2 (adjusted p value [padj] < 2.2e−308; 2.45 log2 fold change [log2FC]; Wilcoxon rank-sum test [FindMarkers, Seurat]) and HOMER1 (padj = 2.78e−88; 2.1 log2FC), were upregulated: gene set enrichment analysis (GSEA) confirmed that genes linked to cell communication and synaptic signaling were dysregulated in excitatory neurons (Figure S2A). In microglia, we found that genes linked to the cell cycle, such as MKI67 (padj = 5.3e−6; 1.18 log2FC) and TOP2A (padj = 1e−4; 1.02 log2FC), were upregulated, and Gene Ontology (GO) analysis confirmed a transcriptional response linked to cell proliferation (Figure S2B). In line with this observation, a global analysis of cell-cycle-related genes (function "CellCycleScoring," Seurat) confirmed that we detected microglial cells that were in a proliferative state in TBI samples and not in the controls (Figure S2E), which is in line with what is expected after TBI. No evidence of proliferation was detected in any other cell types after TBI. Thus, these results demonstrate that we detected neuronal dysfunction and the initiation of a microglia response upon TBI using the snRNA-seq approach.

When investigating transcriptomic changes in the other cell types, we found that OPCs and mature oligodendrocytes displayed a unique transcriptional response linked to neuroinflammation following severe TBI. In both cell types, we found that many genes involved in innate immunity and an interferon response, including STAT1 and STAT2, were activated (Figures 2A and 2B). GSEA confirmed that upon TBI, genes related to terms such as innate immune response and defense response were significantly enriched among the upregulated genes in both OPCs and mature oligodendrocytes (Figures 2C and S3A). Particularly in mature oligodendrocytes, we also found the activation of terms related to a response to interferon-gamma and cytokine stimulus (Figure 2C) showing clear upregulation of interferon regulatory factor IRF1 (padj = 1.88e−12; 0.64 log2FC), interferon-induced IFI16 (padj = 0.001; 0.53 log2FC), and several PARP genes that respond to DNA damage and have different antiviral properties (PARP9 [padj = 3.02e−39; 1.13 log2FC], PARP12

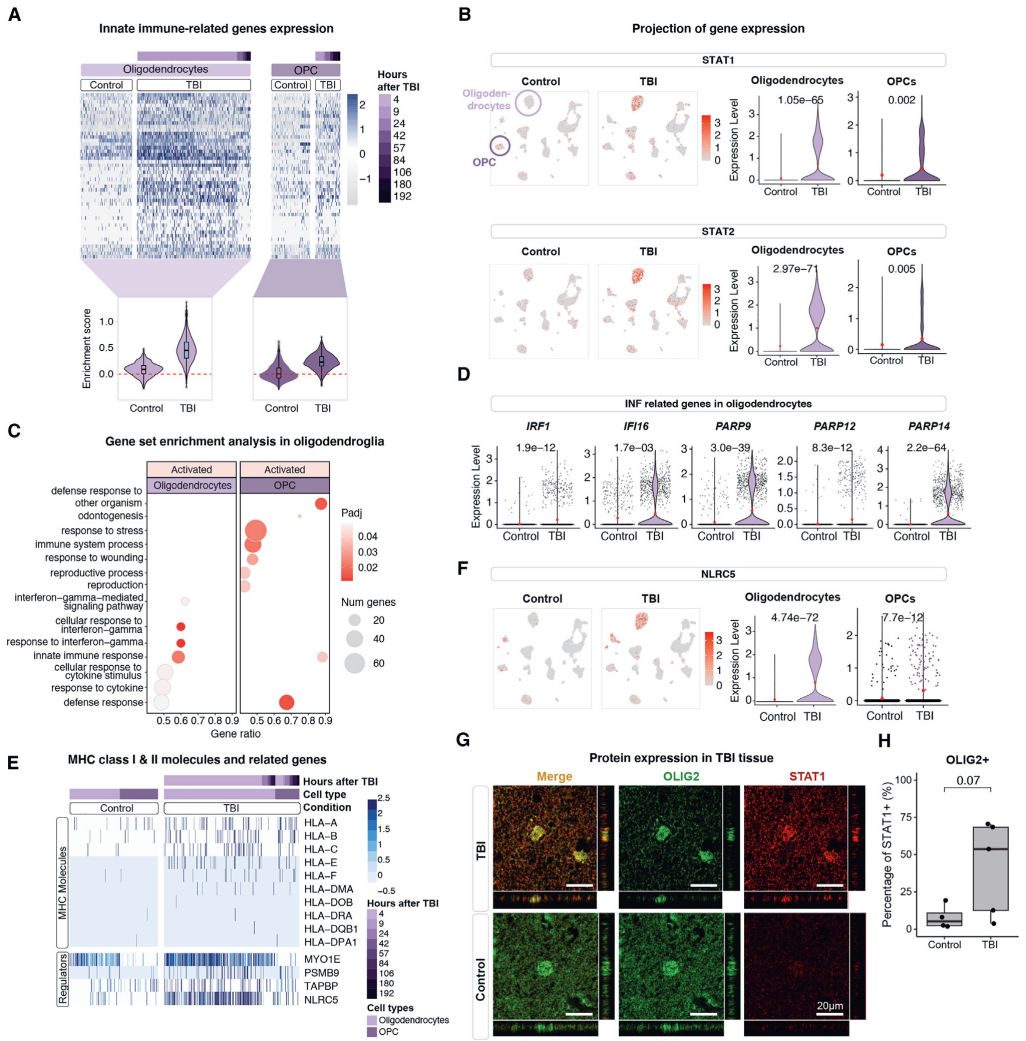


Figure 2. Oligodendroglia display an innate immune response upon TBI

(A) Top: heatmap of oligodendroglia expression of differentially expressed genes (TBI vs. ctrl; padj < 0.01; log2FC > 0.05; n = 12 TBI samples, n = 5 ctrl samples) related to an innate immune response. Bottom: enrichment scores per condition of the genes shown in heatmap (AddModuleScore, Seurat).

(B) Top: STAT1 expression projected onto control and TBI UMAPs (left); violin plots showing expression per condition in oligodendroglia (right; TBI vs. ctrl; Wilcoxon rank-sum test [FindMarkers, Seurat]; n = 12 TBI samples, n = 5 ctrl samples). Bottom: STAT2 expression projected onto control and TBI UMAPs (left); violin plots showing expression per condition in oligodendroglia (right; TBI vs. ctrl; Wilcoxon rank-sum test [FindMarkers, Seurat]; n = 12 TBI samples, n = 5 ctrl samples).

(C) Activated terms from GSEA of differentially expressed genes (TBI vs. ctrl; padj < 0.01, Wilcoxon rank-sum test [FindMarkers, Seurat]; n = 12 TBI samples, n = 5 ctrl samples) in oligodendroglia (biological process ontology).

(D) Violin plots showing the expression of selected genes related to defense, IFN response, or inflammation in oligodendrocytes (TBI vs. ctrl; Wilcoxon rank-sum test [FindMarkers, Seurat]; n = 12 TBI samples, n = 5 ctrl samples).

(E) Expression of MHC classes I and II and related genes found to be significantly upregulated in oligodendroglia.

(legend continued on next page)

[padj = 8.26e−12; 0.44 log2FC], and PARP14 [padj = 2.22e−64; 1.63 log2FC]) (Figure 2D).[23,24]

Many of the genes upregulated in TBI oligodendroglia were MHC class I genes as well as regulators of classes I and II (Figures 2E, 2F, S3B, and S3C). For example, TBI oligodendroglia were found to have a significantly higher expression of NLRC5 (oligodendrocytes padj = 4.74e−72; 1.55 log2FC; OPCs padj = 7.65e−12; 0.6 log2FC) (Figure 2F), the major regulator of genes in MHC class I,[25] as well as expressing other related genes such as PSMB9 (oligodendrocytes padj = 9.65e−09, log2FC 0.49) and TAPBP (oligodendrocytes padj = 5.87e−05, 0.41 log2FC), which process MHC class I molecules and peptides, and MYO1E (oligodendrocytes padj = 1.43e−12; 0.72 log2FC), which regulates antigen presentation of MHC class II molecules[26] (Figure 2E).

When we performed a similar analysis on microglia and astrocytes, we found very limited evidence of an innate or interferon response or the upregulation of MHC molecules (Figures S3A–S3C). The induction of immune genes in oligodendroglia was robust between individuals, with a trend for stronger induction of immune genes in samples collected a short time after injury (up to 4 h) (Figures S3D and S3E). To verify the activation of immune-related genes in oligodendroglia using an alternative quantification strategy for the snRNA-seq data, we used a pseudo-bulk approach. This analysis confirmed that immune-related genes were highly expressed in oligodendroglia in TBI tissue (Figure S3F).

To confirm an activation of the interferon response at the protein level, we performed immunohistochemistry (IHC) in TBI and control tissue for STAT1 in combination with OLIG2, which specifically marks oligodendrocytes (Figure 2G). Confocal microscopy confirmed numerous STAT1+/OLIG2+ cells in TBI tissue. We used automated microscopy analysis to quantify the number of STAT1-expressing OLIG2+ cells. While control tissue exhibited sparse STAT1 expression in oligodendrocytes, we observed an induction of STAT1 expression in oligodendrocytes in TBI tissue, although the signal was variable among individuals due to the technical challenges to perform robust IHC analysis on the severely damaged TBI tissue (Figure 2H, 41% in TBI vs. 8% in control [ctrl], p = 0.073 Student's t test). Taken together, these results suggest that oligodendroglia undergo a unique transcriptional response following TBI, which activates an interferon response and turns on MHC-related genes—thereby transforming to an immune-like cell state.

Activation of ERVs in oligodendrocytes after TBI

The transcriptional response in OPCs and mature oligodendrocytes after TBI is reminiscent to that which occurs after viral infection. In this respect, the transcriptional activation of ERVs (Figure 3A) has been linked to an interferon response. The aberrant expression of ERVs results in the formation of double-stranded RNAs, reverse-transcribed DNA molecules, and ERV-derived peptides that induce a "viral mimicry," where cells respond as though infected and this triggers or boosts inflammation.[18–20]

To investigate if ERVs, or any other TEs, were activated following TBI, we used an in-house bioinformatic pipeline[17] allowing the analysis of ERV and TE expression from our 10× snRNA-seq dataset (Figure 3B). In brief, this method uses the cell clusters determined based on the gene expression. Then, by backtracing the reads from cells forming each cluster, it is possible to analyze the expression of ERVs and other TEs using the specialized software TEtranscripts[27] in distinct cell populations. This approach greatly increases the sensitivity of the analysis and enables quantitative estimation of ERV expression at single-cell-type resolution.

When using this bioinformatic approach, we found that several ERV (long terminal repeat [LTR]) subfamilies were transcriptionally activated in oligodendrocytes and OPCs. This response was especially noticeable in ERV subfamilies such as HERV-K (LTR5-Hs and LTR5B, FCs of 2.34–2.7 in oligodendrocytes and 2–3.7 in OPCs), HERV-W (LTR17, FCs of 3.41 in oligodendrocytes and 2.8 in OPCs), and HERV-H (LTR7, FCs of 1.34 in oligodendrocytes and 2.3 in OPCs), which are evolutionary young ERVs found specifically in primates (Figures 3C, 3D, and S4A). We found no transcriptional activation of other families of TEs such as LINE-1s (Figures 3E and S4B) and no evidence for the activation of ERVs or other TEs in other cell types, except for microglia, which also displayed some evidence of ERV activation (Figures 3D, 3E, and S4A). When quantifying ERV subfamilies while grouping the samples by time after injury (early = 4–9 h; late = >9 h), we found that the ERV upregulation in oligodendroglia occurs early after the injury (Figure S4C), which correlates to the observations regarding immune-related genes in oligodendrocytes (Figure S3E).

The ERV families that were upregulated in TBI oligodendrocytes include provirus insertions with the potential to be transcribed and translated to produce ERV-derived peptides (top, Figure 3A). Such ERVs have previously been linked to viral mimicry and the induction of an interferon response.[19] However, 10× libraries display a 3′ bias, making it impossible to distinguish between ERVs and solo LTR fragments that are present in large numbers in the human genome due to recombination events between the two endogenous provirus LTRs during primate evolution (Figure 3A). Thus, to investigate which unique loci the upregulated ERV expression in TBI oligodendrocytes originates from, we performed deep 2 × 150 bp paired-end, strand-specific bulk RNA-seq of four of the TBI samples with a high composition of oligodendroglia, as well as three control samples (Figures 4A, S1C, and S1E). We discarded all ambiguously mapping reads and only quantified those that map uniquely. We

(F) Expression of the key regulator of MHC class II molecules, NLRC5, projected onto control and TBI UMAPs (left); violin plot showing NLRC5 expression per condition in oligodendroglia (right; TBI vs. ctrl; Wilcoxon rank-sum test [FindMarkers, Seurat]; n = 12 TBI samples, n = 5 ctrl samples).
(G) Immunohistochemical co-labeling of STAT1 (red) and OLIG2 (green). Confocal microscopy analysis revealed the presence of STAT1-expressing OLIG2+ oligodendrocytes in TBI tissue. Scale bar: 20 μm.
(H) Automated microscopy quantification of the percentage of STAT1+ cells among OLIG2+ cells in TBI and control tissue (TBI vs. ctrl; p = 0.073, Student's t test, n = 5 TBI samples, n = 4 ctrl samples). Boxplot points, median line, box limits, and whiskers showing percentage of STAT1+ cells in each sample (points), median among samples (median line), upper and lower quartiles (box), and highest and lowest values (whiskers), respectively.

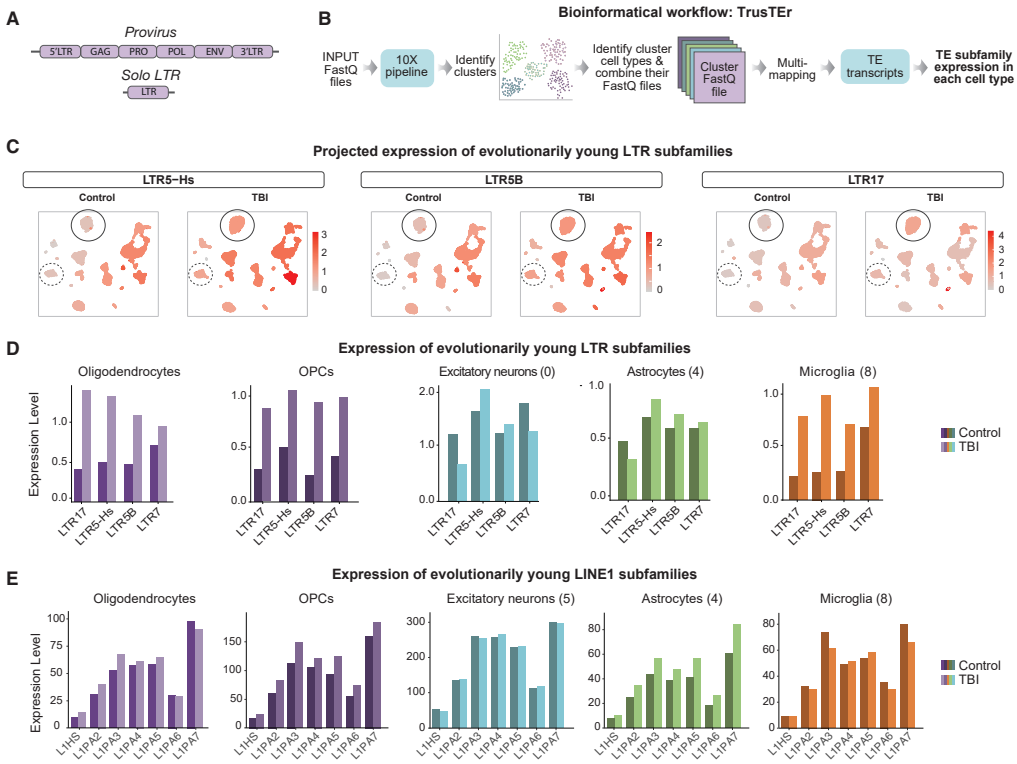


Figure 3. Expression of ERVs in TBI oligodendroglia

(A) Schematic of the structure of an ERV provirus (top) and solo LTR (bottom).

(B) Schematic of bioinformatic approach to quantify TE subfamilies per cell cluster.

(C) Split UMAPs (control and TBI) with projected expression of LTR subfamilies per cluster. Oligodendrocytes are circled, and OPCs are circled with a dotted line.

(D) Expression of evolutionary young LTR subfamilies in oligodendrocytes and OPCs.

(E) Expression of evolutionary young L1 subfamilies in oligodendrocytes and OPCs.

quantified the expression of unique ERV predictions identified using RetroTector[28] (Table S5) and found 13 significantly upregulated ERV loci in the TBI samples (*DESeq2*, padj < 0.2, log2FC > 1) (Figures 4B and S5A; Table S6), which range in length from 6.2 to 13.5 kbp, where transcription was found across the ERV sequence (Figure 4C). Using a unique mapping approach on our snRNA-seq dataset, we were able to verify that some of the highly expressed, upregulated ERV loci were only expressed in oligodendroglia (Figure S5B) and that these were exclusively expressed upon TBI (Figure S5C). These activated ERV loci represent evolutionary young elements present only in primates, including also some human-specific ERVs (Figure 4C). Notably, the transcriptional activation of some of these ERVs resulted in a readthrough transcript extending into the nearby genome (see, e.g., the HERVW loci in Figure 4C). The predicted structures of these elements show that 12 out of the 13 upregulated ERVs still contain at least one open reading frame (ORF) of the ERV genes (*gag*, *pro*, *pol*, and *env*) (Figure 4D; Table S5). These data demonstrate that ERVs, with the potential to induce viral mimicry, are transcriptionally activated in oligodendrocytes within hours after TBI.

Interferon treatment results in transcriptional activation of ERVs in human glial progenitor cells

The induction of an interferon response has been associated with the transcriptional activation of ERVs, which is thought to play a part in boosting the interferon response.[29] To investigate a mechanistic link between interferon activation and ERV expression, we decided to perform *in vitro* experiments in human glial progenitor cells (hGPCs).

We differentiated human embryonic stem cells (hESCs) into hGPCs using a 135 day differentiation protocol (Figure 5A; see STAR Methods for details). The differentiated cells displayed a morphology of immature human glial cells and expressed marker

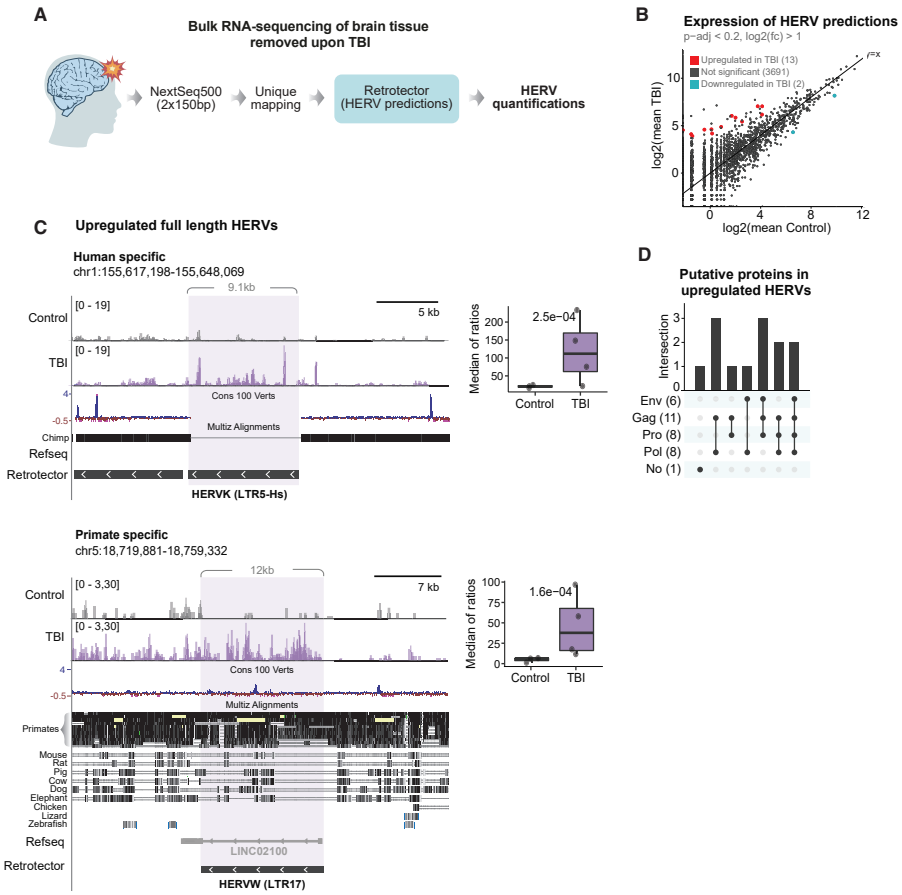


Figure 4. Identification of specific ERV loci expressed after TBI

(A) Schematic for bulk RNA-seq and bioinformatic approach.

(B) Scatterplot showing mean expression per condition and statistical analysis results for differential expression of all ERV predictions (TBI vs. ctrl; DESeq2; padj < 0.2, log2FC > 1 highlighted in red; padj < 0.2 highlighted in blue; n = 4 TBI samples, n = 3 ctrl samples). Reference line y = x in black.

(C) Genome browser tracks per condition with two examples of upregulated HERVs (left – HERVK, right – HERVW) and their sequence conservation. Boxplots on their right showing quantification of the respective element per condition, with the individual sample expression values (points), median expression among samples (median line), upper and lower quartiles (box), and highest and lowest values (whiskers) (normalization by median of ratios and statistical test performed by DESeq2; TBI vs. ctrl; n = 4 TBI samples, n = 3 ctrl samples).

(D) Upset plot showing putative proteins contained in the upregulated ERVs.

genes related to this cell state, including many genes expressed in OPCs *in vivo* (Figures 5B and 5C). The hGPCs were treated with interferon-gamma (IFNγ; 5 ng/mL) for 48 h before being harvested for 2 × 150 bp paired-end, strand-specific bulk RNA-seq analysis.

When investigating changes in gene expression profile upon IFNγ stimulation of hGPCs, we found a distinct and robust transcriptional response. Genes linked to IFN signaling and innate immune activation were highly upregulated in the IFNγ-treated hGPCs (Figures 5D; Table S7). We found a clear correlation between the upregulated genes in IFNγ-treated hGPCs and those observed to be activated in the TBI samples (Figures 5D and 5E). Thus, these observations confirm that hESC-derived GPCs respond to IFNγ treatment by activating downstream targets (IFN-stimulated genes) in a manner that transcriptionally resembles the *in vivo* response following TBI.

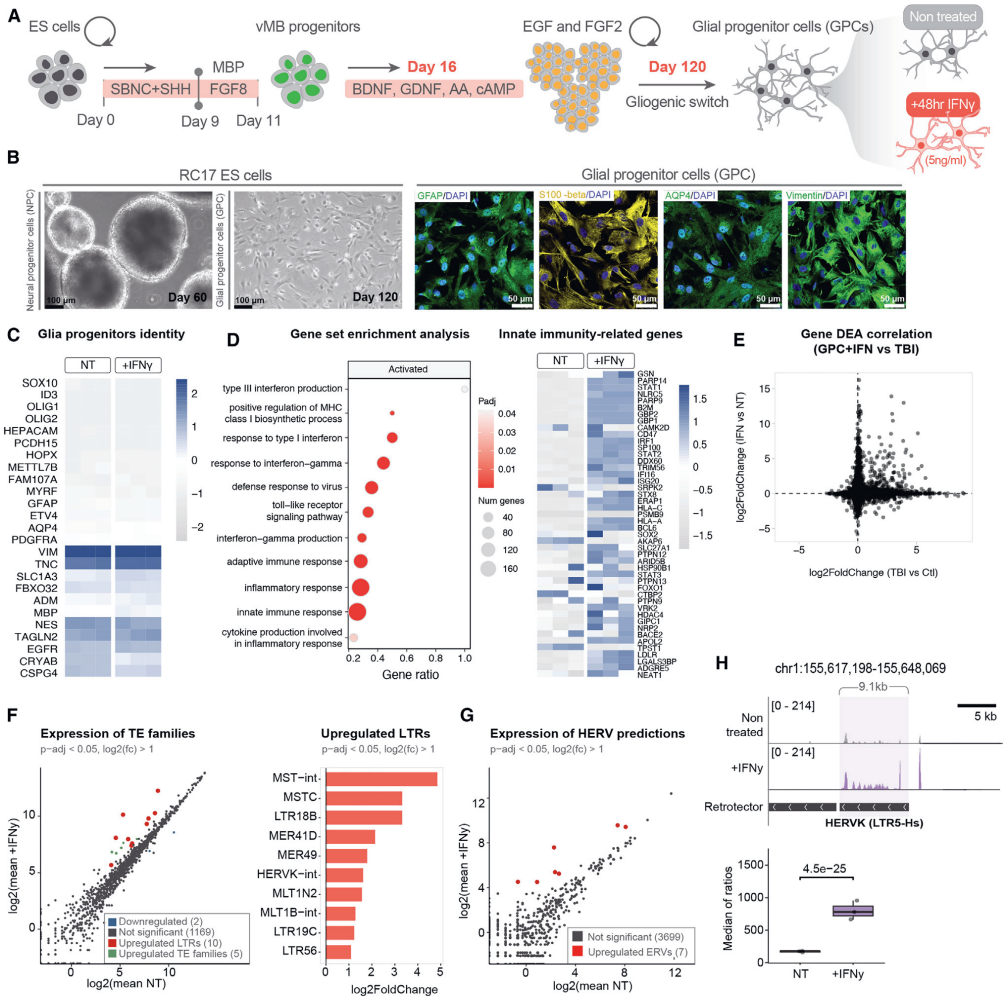


Figure 5. IFNγ treatment triggers transcriptional activation of ERVs in hGPCs

(A) Schematic representation of differentiation protocol for deriving glial progenitor cells from hESCs and experimental approach for IFNγ treatment.

(B) Left: bright-field images of hESC-derived ventral midbrain (vMB) neural progenitor cells at day 60 maintained in a proliferation medium containing EGF and FGF2 (black scale bar: 100 μm). Right: immunocytochemistry of GPCs at day 120 of glia (GFAP, S100-beta, AQP4, Vimentin) and nuclear (DAPI) markers (white scale bar: 50 μm).

(C) Canonical gene markers to validate glia progenitor cell identity.

(D) Left: selection of activated GO terms upon IFNγ treatment (IFNγ vs. non-treated [NT]; gseGO results using genes log2FCs as calculated by DESeq2; n = 3 IFNγ-treated hGPC replicates, n = 3 NT hGPC replicates). Right: expression of innate immune related genes that were found to be upregulated in TBI oligodendroglia (as shown in Figure 2A).

(E) Scatterplot showing all genes log2FC in hGPC IFNγ vs. NT (y axis; IFNγ vs. NT; genes log2FCs as calculated by DESeq2; n = 3 IFNγ-treated hGPC replicates, n = 3 NT hGPC replicates) and TBI vs. control (x axis; TBI vs. ctrl; genes log2FCs as calculated by DESeq2; n = 4 TBI samples, n = 3 ctrl samples).

(F) Left: scatterplot showing mean expression per condition and statistical analysis results for differential expression of TE subfamilies. Upregulated LTR subfamilies are colored in red, and other upregulated TE subfamilies (non-LTRs) are colored in green (IFNγ vs. NT; DESeq2; padj < 0.05; log2FC > 1). Downregulated

(legend continued on next page)

To quantify ERV expression upon IFNγ treatment in hGPCs, we used two different bioinformatic methodologies. First, we allowed reads to map to different locations (multi-mapping) and used the TEtranscripts software[27] in multi-mode to quantify the expression of different ERV subfamilies. Second, we discarded all ambiguously mapping reads and only quantified those that map uniquely and quantified the expression of unique ERV loci. The TEtranscripts approach revealed that several ERV subfamilies were robustly upregulated upon IFNγ treatment (Figures 5F; Table S8). The upregulated ERV families included, for example, HERV-K, which is an evolutionarily young ERV with the capacity to trigger viral mimicry and that we also found to be upregulated in the TBI samples. Notably, we found no activation of other TEs, such as LINE-1s, indicating that the response to IFNγ treatment results in a transcriptional activation response specific to ERVs. The analysis of individual ERV loci identified several proviruses that were upregulated (Figures 5G; Table S8). One of the most robustly upregulated proviral insertions was the same human-specific HERV-K provirus on chromosome 1 that we also found to be upregulated in the TBI samples (compare Figure 5H with the top of Figure 4C). Taken together, this experiment demonstrates that the activation of an IFN response in hGPCs results in transcriptional activation of ERVs, thus providing a mechanistic explanation to our observations of IFN activation and ERV expression in TBI tissue.

DISCUSSION

Neuroinflammation is a hallmark of acute and chronic neurodegenerative states, including TBI, and is therefore a promising route for urgently needed disease-modifying therapies. However, information on how the neuroinflammatory response starts and then transforms to a chronic state is scarce, and the molecular events needed to initiate and maintain this process, and in which cell types, have not been established. This lack of basic understanding hampers the development of treatment strategies. In this study, we used snRNA-seq to perform a detailed analysis of human tissue samples obtained in conjunction with acute surgery for TBI. Our results provide two main insights into the start of a human neuroinflammatory response. First, we found that OPCs and oligodendrocytes may play an unanticipated key role in this process by adopting an immune-like cell state, including evidence for an IFN response. Secondly, we found that the activation of an IFN response in OPCs and oligodendrocytes is mechanistically linked to the transcriptional activation of ERVs. Key to our findings is the use of snRNA-seq analysis, which allows us to look at the quantitative transcriptional responses in discrete cell populations in complex brain tissue samples. This approach solves many of the challenges in this field. For example, quantitative analysis of gene expression in TBI tissue have previously relied on bulk RT-qPCR—which is limited by the heterogeneity of cell composition in the tissue—or *in situ* hybridization and IHC—which are limited in throughput and suffer from severe technical challenges, such as differences in background staining in the injured tissue.

Neuronal cell death is common in acute, severe TBI due to hemorrhages, increased ICP, and energy metabolic failure. In addition, there is a distinct inflammatory response, associated with white matter abnormalities, that persists from the acute post-injury phase for many years post-injury.[2,6,30,31] Prior studies have established that there are numerous contributors to the neuroinflammatory response after TBI, including microglial activation, infiltration of systemic immune cells, and the subsequent release of proinflammatory cytokines.[32,33] Our observations confirm that microglia are likely important players in this process. We observe an increased number of microglia after TBI that are coupled to cell proliferation. However, our results indicate that oligodendroglia also seem to play an important role in the initiation of this process. Oligodendroglia are vulnerable to the TBI-induced neuroinflammation as well as to excitotoxicity and reactive oxygen species formation,[34,35] and loss of mature oligodendrocytes is observed at an early stage following rodent[36–38] and human TBI.[39] Treatment with an anti-inflammatory antibody neutralizing interleukin-1β, a key proinflammatory cytokine, attenuates this post-injury loss of oligodendrocytes in rodent models.[9] The trigger of an IFN response upon TBI has been reported before,[8,40] although the specific cell types involved in this process remained unknown. In the present study, we found that OPCs and oligodendrocytes undergo the activation of IFN response genes as well as genes related to MHC classes I and II. Some of these genes affect the expression, processing, and guidance of MHC molecules, which are crucial to immune cell functions. Our results suggest that, following TBI, oligodendroglia adopt a different cell state characterized by the expression of immune genes. These results are reminiscent of what has been observed in multiple sclerosis, where OPCs and oligodendrocytes initiate a similar transcriptional program.[14,15] The functional outcome of this cellular transformation remains to be explored, but it is possible that oligodendroglia adopt new roles after TBI, including acting as antigen-presenting cells or by triggering an immunologic attack. Mechanistic experiments in animal models are likely needed to resolve this issue.

Almost 10% of the human genome is made up of ERVs, a consequence of their colonization of our germline throughout evolution.[41] In adult tissues, including the brain, ERVs are normally transcriptionally silenced via epigenetic mechanisms, including DNA and histone methylation.[16] However, there is emerging evidence, both from animal models and the analysis of human material, indicating that ERVs can be transcriptionally activated in the diseased brain and that this correlates with

subfamilies are colored in blue (IFNγ vs. NT; DESeq2; padj < 0.05; log2FC < −1). Right: bar plot showing log2FC of upregulated LTR subfamilies (IFNγ vs. NT; DESeq2; padj < 0.05; log2FC > 1) (statistics performed using n = 3 IFNγ-treated hGPC replicates, n = 3 NT hGPC replicates).

(G) Scatterplot showing mean expression per condition and statistical analysis results for differential expression of all ERV predictions. IFNγ vs. NT; DESeq2; p value < 0.05, log2FC > 1 highlighted in red; p value < 0.05 highlighted in blue (n = 3 IFNγ-treated hGPC replicates, n = 3 NT hGPC replicates).

(H) Tracks per condition showing an upregulated HERV-K (as shown in Figure 4C). Boxplot showing quantification of the element per condition, with the individual sample expression values (points), median expression among samples (median line), upper and lower quartiles (box), and highest and lowest values (whiskers) (normalization by median of ratios and statistical test performed by DESeq2; IFNγ vs. NT; n = 3 IFNγ-treated hGPC replicates, n = 3 NT hGPC replicates).

neuroinflammation. ERV expression has been found to be elevated in the cerebrospinal fluid and in post-mortem brain biopsies from patients with multiple sclerosis, amyotrophic lateral sclerosis, and Alzheimer's or Parkinson's disease.[16,42–50] In experimental drosophila or mouse models, there is causative evidence linking upregulation of ERVs and other TEs to neuroinflammation and neurodegeneration.[17,51] However, accurate estimation of ERV expression is challenging: their repetitive nature complicates the bioinformatic analysis, and many of these observations remain controversial. Furthermore, most of the clinical studies use end-stage post-mortem material, where the inflammatory process has been ongoing for decades, or are limited to the study of biofluids such as cerebrospinal fluid or plasma. These limitations have made it challenging to conclusively link ERV expression to the initiation of the neuroinflammatory response. Our results now provide direct evidence that ERV proviruses are transcriptionally activated at the start of human neuroinflammation. Notably, we find that ERVs are specifically activated in oligodendrocytes, OPCs, and microglia, the cell types involved in the inflammatory response. Thus, our results provide direct clinical evidence linking the induction of an immune response and the transcriptional activation of ERVs. Our mechanistic *in vitro* modeling in human glia progenitor cultures demonstrates that the induction of an innate immune response causes a transcriptional activation of ERVs. We do not understand why an IFN response results in transcriptional activation of ERVs. It will be interesting to investigate the nature of the transcription factors recruited to ERVs upon IFN treatment and if this results in chromatin remodeling that releases their transcriptional silencing. In addition, exactly what initially triggers the IFN response remains unclear, but it may be linked to the release of mitochondrial DNA into the cytosol or to genomic DNA damage.

From our results, it is also not possible to pinpoint the role ERVs have in driving and boosting the immune response following their transcriptional activation. Previous studies have demonstrated that aberrant expression of ERVs results in the formation of double-stranded RNAs and reverse-transcribed DNA molecules, which can trigger viral mimicry.[18–20] However, there is also growing awareness that the transcriptional activation of ERVs results in the production of peptides that can activate immune pathways. For example, the same HERV-K provirus on chromosome 1 (also known as *ERVK-7*[52]) that we found upregulated in TBI tissue and IFNγ-treated hGPCs (Figures 4C and 5G) was recently reported to trigger the production of ERV-reactive antibodies in human patients with lung cancer through the expression if its envelope glycoprotein.[52] In addition, ERV-derived peptides may contribute to protein aggregation, which is a process directly linked to chronic effects following TBI and which also may further enhance an inflammatory response.[53] The development of robust reagents and protocols to identify ERV-derived peptides with specificity to unique loci will be key to clarify if ERV-derived proteins are involved in neuroinflammatory responses.

In summary, we here describe a role for OPCs and oligodendrocytes in the initiation of an IFN response following acute severe human TBI. We describe the activation of an IFN response in oligodendroglia that is linked to the activation of regulatory genes of molecules present in immune cells. This transcriptional switch in OPCs and oligodendrocytes is mechanistically linked to the activation of ERVs. Our results could lead to treatment opportunities for acute TBI, as well as for chronic neurodegenerative disorders.

Limitations of the study

The TBI tissue samples used in this study come from rare cases of surgically evacuated tissue from emergent, life-saving procedures. This greatly limits the possibility to match individuals regarding brain region, sex, tissue composition, and time post-impact. The samples also display severely disrupted tissue integrity because of the injury, making it extremely challenging to obtain high-quality sequencing data. We initially attempted sequencing experiments from a total of 18 TBI tissue samples and achieved high-quality snRNA-seq data from 12 samples (listed in Table S1). Such a small cohort size could limit the strength of some of our conclusions. Still, the expression of housekeeping genes was homogeneous among cell types and conditions, and our observations regarding the immune response in oligodendroglia were replicated using different analytical approaches. In addition, two of our samples contained a high degree of oligodendrocytes—likely a consequence of those samples being rich in white matter. This sample heterogeneity could potentially skew the data. However, despite these limitations, we are confident that our observations are robust and accurate. When our dataset was reanalyzed with the two samples rich in oligodendrocytes omitted (TBI 2 and TBI 3), we still observed a robust activation of immune-related genes in oligodendrocytes after TBI (Figure S6). In addition, the key observation regarding activation of an IFN response and activation of ERVs in oligodendrocytes after TBI can also be seen in OPCs, and the distribution of OPCs is not skewed in our different samples. This suggest that the immune response in oligodendroglia is robust among our samples.

Another challenge is the choice of control tissue. There is no perfect control material to obtain for the study of human TBI tissue. We chose to use post-mortem tissue obtained from acute, non-neurological deaths from our clinic. These samples have been handled in a similar way to the TBI samples, with the exception that they are coming from deceased individuals. Aware of these limitations, we decided to look only at transcriptional alterations with robust differences in magnitude. For example, many of the genes related to an IFN response are completely silent in oligodendrocytes and OPCs (such as STAT1 and STAT2) and are then robustly transcriptionally activated in TBI tissue. By using strict thresholds, we limit the possibilities to draw wrong conclusions due to differences in tissue origin or tissue quality.

Nevertheless, future validation of our observations in different cohorts, including the use of orthogonal technical approaches, will be essential to clarify the role of oligodendroglia in the initiation of neuroinflammation after TBI. This is important since a detailed understanding of the underlying mechanisms linked to their cell-fate change could lead to treatment opportunities.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE

- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Ethical statement
- METHOD DETAILS
  - Sampling and preparation of brain tissue
  - Immunohistochemistry
  - Single-nuclei RNA sequencing
  - Bulk RNA sequencing
  - Generation of glial progenitors from hES cells
  - Immunocytochemistry
  - IFN-γ stimulation of glial progenitors
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Immunohistochemistry statistical analysis
  - Single-nuclei RNA sequencing analysis
  - TE quantification per cell cluster
  - Bulk RNA sequencing analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2023.113395.

ACKNOWLEDGMENTS

We would like to thank Roger A. Barker for critical input on the study and Alafuzoff, J. Johansson, U. Jarl, M. Vejgården, B. Mattsson, and A. Hammarberg for technical assistance. We are grateful to all members of the Jakobsson and Marklund labs. The work was supported by grants from the Swedish Research Council (2018-02694 to J.J., 2018-02500 to N.M., and 2021-01740 to P.J.), the Swedish Brain Foundation (FO2019-0098 to J.J. and FO2020-0147 to N.M.), the Novo Nordisk Foundation (NNF21CC0073729 to A.K.), and the Swedish Government Initiative for Strategic Research Areas (MultiPark and StemTherapy).

AUTHOR CONTRIBUTIONS

All of the authors took part in designing the study and interpreting the data. R.G., J.J., and N.M. conceived and designed the study. D.A.M.A., A.T., S.A.H., V.H., A.A., and M.E.J. performed the clinical and experimental research. R.G. and Y.S. performed the bioinformatic analyses. M.I., M.G.H., P.J., E.E., and A.K. contributed reagents and expertise. R.G., J.J., and N.M. wrote the manuscript, and all authors reviewed the final version.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 20, 2023
Revised: September 5, 2023
Accepted: October 20, 2023
Published: November 14, 2023

REFERENCES

1. Cole, J.H., Jolly, A., de Simoni, S., Bourke, N., Patel, M.C., Scott, G., and Sharp, D.J. (2018). Spatial patterns of progressive brain volume loss after moderate-severe traumatic brain injury. Brain *141*, 822–836. https://doi.org/10.1093/brain/awx354.
2. Johnson, V.E., Stewart, J.E., Begbie, F.D., Trojanowski, J.Q., Smith, D.H., and Stewart, W. (2013). Inflammation and white matter degeneration persist for years after a single traumatic brain injury. Brain *136*, 28–42. https://doi.org/10.1093/brain/aws322.
3. Graham, N.S.N., Jolly, A., Zimmerman, K., Bourke, N.J., Scott, G., Cole, J.H., Schott, J.M., and Sharp, D.J. (2020). Diffuse axonal injury predicts neurodegeneration after moderate-severe traumatic brain injury. Brain *143*, 3685–3698. https://doi.org/10.1093/brain/awaa316.
4. Raj, R., Kaprio, J., Jousilahti, P., Korja, M., and Siironen, J. (2022). Risk of Dementia After Hospitalization Due to Traumatic Brain Injury: A Longitudinal, Population-Based Study. Neurology *98*, e2377–e2386. https://doi.org/10.1212/WNL.0000000000200290.
5. Loane, D.J., Kumar, A., Stoica, B.A., Cabatbat, R., and Faden, A.I. (2014). Progressive neurodegeneration after experimental brain trauma: association with chronic microglial activation. J. Neuropathol. Exp. Neurol. *73*, 14–29. https://doi.org/10.1097/NEN.0000000000000021.
6. Ramlackhansingh, A.F., Brooks, D.J., Greenwood, R.J., Bose, S.K., Turkheimer, F.E., Kinnunen, K.M., Gentleman, S., Heckemann, R.A., Gunanayagam, K., Gelosa, G., and Sharp, D.J. (2011). Inflammation after trauma: microglial activation and traumatic brain injury. Ann. Neurol. *70*, 374–383. https://doi.org/10.1002/ana.22455.
7. Ritzel, R.M., Doran, S.J., Barrett, J.P., Henry, R.J., Ma, E.L., Faden, A.I., and Loane, D.J. (2018). Chronic Alterations in Systemic Immune Function after Traumatic Brain Injury. J. Neurotrauma *35*, 1419–1436. https://doi.org/10.1089/neu.2017.5399.
8. Barrett, J.P., Henry, R.J., Shirey, K.A., Doran, S.J., Makarevich, O.D., Ritzel, R.M., Meadows, V.A., Vogel, S.N., Faden, A.I., Stoica, B.A., and Loane, D.J. (2020). Interferon-beta Plays a Detrimental Role in Experimental Traumatic Brain Injury by Enhancing Neuroinflammation That Drives Chronic Neurodegeneration. J. Neurosci. *40*, 2357–2370. https://doi.org/10.1523/JNEUROSCI.2516-19.2020.
9. Flygt, J., Ruscher, K., Norberg, A., Mir, A., Gram, H., Clausen, F., and Marklund, N. (2018). Neutralization of Interleukin-1beta following Diffuse Traumatic Brain Injury in the Mouse Attenuates the Loss of Mature Oligodendrocytes. J. Neurotrauma *35*, 2837–2849. https://doi.org/10.1089/neu.2018.5660.
10. Xing, J., Ren, L., Xu, H., Zhao, L., Wang, Z.H., Hu, G.D., and Wei, Z.L. (2022). Single-Cell RNA Sequencing Reveals Cellular and Transcriptional Changes Associated With Traumatic Brain Injury. Front. Genet. *13*, 861428. https://doi.org/10.3389/fgene.2022.861428.
11. Raabe, F.J., Stephan, M., Waldeck, J.B., Huber, V., Demetriou, D., Kannaiyan, N., Galinski, S., Glaser, L.V., Wehr, M.C., Ziller, M.J., et al. (2022). Expression of Lineage Transcription Factors Identifies Differences in Transition States of Induced Human Oligodendrocyte Differentiation. Cells *11*, 241. https://doi.org/10.3390/cells11020241.
12. Floriddia, E.M., Lourenço, T., Zhang, S., van Bruggen, D., Hilscher, M.M., Kukanja, P., Gonçalves Dos Santos, J.P., Altınkök, M., Yokota, C., Llorens-Bobadilla, E., et al. (2020). Distinct oligodendrocyte populations have spatial preference and different responses to spinal cord injury. Nat. Commun. *11*, 5860. https://doi.org/10.1038/s41467-020-19453-x.
13. Seeker, L.A., and Williams, A. (2022). Oligodendroglia heterogeneity in the human central nervous system. Acta Neuropathol. *143*, 143–157. https://doi.org/10.1007/s00401-021-02390-4.
14. Falcão, A.M., van Bruggen, D., Marques, S., Meijer, M., Jäkel, S., Agirre, E., Samudyata, Floriddia, E.M., Floriddia, E.M., Vanichkina, D.P., Ffrench-Constant, C., et al. (2018). Disease-specific oligodendrocyte lineage cells arise in multiple sclerosis. Nat. Med. *24*, 1837–1844. https://doi.org/10.1038/s41591-018-0236-y.
15. Meijer, M., Agirre, E., Kabbe, M., van Tuijn, C.A., Heskol, A., Zheng, C., Mendanha Falcão, A., Bartosovic, M., Kirby, L., Calini, D., et al. (2022). Epigenomic priming of immune genes implicates oligodendroglia in multiple sclerosis susceptibility. Neuron *110*, 1193–1210.e13. https://doi.org/10.1016/j.neuron.2021.12.034.

16. Jönsson, M.E., Garza, R., Johansson, P.A., and Jakobsson, J. (2020). Transposable Elements: A Common Feature of Neurodevelopmental and Neurodegenerative Disorders. Trends Genet. *36*, 610–623. https://doi.org/10.1016/j.tig.2020.05.004.

17. Jönsson, M.E., Garza, R., Sharma, Y., Petri, R., Södersten, E., Johansson, J.G., Johansson, P.A., Atacho, D.A., Pircs, K., Madsen, S., et al. (2021). Activation of endogenous retroviruses during brain development causes an inflammatory response. EMBO J. *40*, e106423. https://doi.org/10.15252/embj.2020106423.

18. Hurst, T.P., and Magiorkinis, G. (2015). Activation of the innate immune response by endogenous retroviruses. J. Gen. Virol. *96*, 1207–1218. https://doi.org/10.1099/jgv.0.000017.

19. Ishak, C.A., Classon, M., and De Carvalho, D.D. (2018). Deregulation of Retroelements as an Emerging Therapeutic Opportunity in Cancer. Trends Cancer *4*, 583–597. https://doi.org/10.1016/j.trecan.2018.05.008.

20. Saleh, A., Macia, A., and Muotri, A.R. (2019). Transposable Elements, Inflammation, and Neurological Disease. Front. Neurol. *10*, 894. https://doi.org/10.3389/fneur.2019.00894.

21. Abu Hamdeh, S., Ciuculete, D.M., Sarkisyan, D., Bakalkin, G., Ingelsson, M., Schiöth, H.B., and Marklund, N. (2021). Differential DNA Methylation of the Genes for Amyloid Precursor Protein, Tau, and Neurofilaments in Human Traumatic Brain Injury. J. Neurotrauma *38*, 1679–1688. https://doi.org/10.1089/neu.2020.7283.

22. Slyper, M., Porter, C.B.M., Ashenberg, O., Waldman, J., Drokhlyansky, E., Wakiro, I., Smillie, C., Smith-Rosario, G., Wu, J., Dionne, D., et al. (2020). A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. Nat. Med. *26*, 792–802. https://doi.org/10.1038/s41591-020-0844-1.

23. Atasheva, S., Frolova, E.I., and Frolov, I. (2014). Interferon-stimulated poly(ADP-Ribose) polymerases are potent inhibitors of cellular translation and virus replication. J. Virol. *88*, 2116–2130. https://doi.org/10.1128/JVI.03443-13.

24. Malgras, M., Garcia, M., Jousselin, C., Bodet, C., and Leveque, N. (2021). The Antiviral Activities of Poly-ADP-Ribose Polymerases. Viruses *13*, 582. https://doi.org/10.3390/v13040582.

25. Kobayashi, K.S., and Van Den Elsen, P.J. (2012). NLRC5: a key regulator of MHC class I-dependent immune responses. Nat. Rev. Immunol. *12*, 813–820. https://doi.org/10.1038/nri3339.

26. Paul, P., van den Hoorn, T., Jongsma, M.L.M., Bakker, M.J., Hengeveld, R., Janssen, L., Cresswell, P., Egan, D.A., van Ham, M., Ten Brinke, A., et al. (2011). A Genome-wide Multidimensional RNAi Screen Reveals Pathways Controlling MHC Class II Antigen Presentation. Cell *145*, 268–283. https://doi.org/10.1016/j.cell.2011.03.023.

27. Jin, Y., Tam, O.H., Paniagua, E., and Hammell, M. (2015). TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics *31*, 3593–3599. https://doi.org/10.1093/bioinformatics/btv422.

28. Sperber, G.O., Airola, T., Jern, P., and Blomberg, J. (2007). Automated recognition of retroviral sequences in genomic data—RetroTector. Nucleic Acids Res. *35*, 4964–4976. https://doi.org/10.1093/nar/gkm515.

29. Chuong, E.B., Elde, N.C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science *351*, 1083–1087. https://doi.org/10.1126/science.aad5497.

30. Holmin, S., Söderlund, J., Biberfeld, P., and Mathiesen, T. (1998). Intracerebral inflammation after human brain contusion. Neurosurgery *42*, 291–298, discussion 298-299. https://doi.org/10.1097/00006123-199802000-00047.

31. Marklund, N., Vedung, F., Lubberink, M., Tegner, Y., Johansson, J., Blennow, K., Zetterberg, H., Fahlström, M., Haller, S., Stenson, S., et al. (2021). Tau aggregation and increased neuroinflammation in athletes after sports-related concussions and in traumatic brain injury patients - A PET/MR study. Neuroimage. Clin. *30*, 102665. https://doi.org/10.1016/j.nicl.2021.102665.

32. Todd, B.P., Chimenti, M.S., Luo, Z., Ferguson, P.J., Bassuk, A.G., and Newell, E.A. (2021). Traumatic brain injury results in unique microglial and astrocyte transcriptomes enriched for type I interferon response. J. Neuroinflammation *18*, 151. https://doi.org/10.1186/s12974-021-02197-w.

33. Witcher, K.G., Bray, C.E., Chunchai, T., Zhao, F., O'Neil, S.M., Gordillo, A.J., Campbell, W.A., McKim, D.B., Liu, X., Dziabis, J.E., et al. (2021). Traumatic Brain Injury Causes Chronic Cortical Inflammation and Neuronal Dysfunction Mediated by Microglia. J. Neurosci. *41*, 1597–1616. https://doi.org/10.1523/JNEUROSCI.2469-20.2020.

34. Giacci, M.K., Bartlett, C.A., Smith, N.M., Iyer, K.S., Toomey, L.M., Jiang, H., Guagliardo, P., Kilburn, M.R., and Fitzgerald, M. (2018). Oligodendroglia Are Particularly Vulnerable to Oxidative Damage after Neurotrauma In Vivo. J. Neurosci. *38*, 6491–6504. https://doi.org/10.1523/JNEUROSCI.1898-17.2018.

35. Xu, G.Y., Liu, S., Hughes, M.G., and McAdoo, D.J. (2008). Glutamate-induced losses of oligodendrocytes and neurons and activation of caspase-3 in the rat spinal cord. Neuroscience *153*, 1034–1047. https://doi.org/10.1016/j.neuroscience.2008.02.065.

36. Dent, K.A., Christie, K.J., Bye, N., Basrai, H.S., Turbic, A., Habgood, M., Cate, H.S., and Turnley, A.M. (2015). Oligodendrocyte birth and death following traumatic brain injury in adult mice. PLoS One *10*, e0121541. https://doi.org/10.1371/journal.pone.0121541.

37. Flygt, J., Djupsjö, A., Lenne, F., and Marklund, N. (2013). Myelin loss and oligodendrocyte pathology in white matter tracts following traumatic brain injury in the rat. Eur. J. Neurosci. *38*, 2153–2165. https://doi.org/10.1111/ejn.12179.

38. Mierzwa, A.J., Marion, C.M., Sullivan, G.M., McDaniel, D.P., and Armstrong, R.C. (2015). Components of myelin damage and repair in the progression of white matter pathology after mild traumatic brain injury. J. Neuropathol. Exp. Neurol. *74*, 218–232. https://doi.org/10.1097/NEN.0000000000000165.

39. Flygt, J., Gumucio, A., Ingelsson, M., Skoglund, K., Holm, J., Alafuzoff, I., and Marklund, N. (2016). Human Traumatic Brain Injury Results in Oligodendrocyte Death and Increases the Number of Oligodendrocyte Progenitor Cells. J. Neuropathol. Exp. Neurol. *75*, 503–515. https://doi.org/10.1093/jnen/nlw025.

40. Abdullah, A., Zhang, M., Frugier, T., Bedoui, S., Taylor, J.M., and Crack, P.J. (2018). STING-mediated type-I interferons contribute to the neuroinflammatory process and detrimental effects following traumatic brain injury. J. Neuroinflammation *15*, 323. https://doi.org/10.1186/s12974-018-1354-7.

41. Jern, P., and Coffin, J.M. (2008). Effects of retroviruses on host genome function. Annu. Rev. Genet. *42*, 709–732. https://doi.org/10.1146/annurev.genet.42.110807.091501.

42. Andrews, W.D., Tuke, P.W., Al-Chalabi, A., Gaudin, P., Ijaz, S., Parton, M.J., and Garson, J.A. (2000). Detection of reverse transcriptase activity in the serum of patients with motor neurone disease. J. Med. Virol. *61*, 527–532. https://doi.org/10.1002/1096-9071(200008)61:4<527::aid-jmv17>3.0.co;2-a.

43. Douville, R., Liu, J., Rothstein, J., and Nath, A. (2011). Identification of active loci of a human endogenous retrovirus in neurons of patients with amyotrophic lateral sclerosis. Ann. Neurol. *69*, 141–151. https://doi.org/10.1002/ana.22149.

44. Garson, J.A., Usher, L., Al-Chalabi, A., Huggett, J., Day, E.F., and McCormick, A.L. (2019). Quantitative analysis of human endogenous retrovirus-K transcripts in postmortem premotor cortex fails to confirm elevated expression of HERV-K RNA in amyotrophic lateral sclerosis. Acta Neuropathol. Commun. *7*, 45. https://doi.org/10.1186/s40478-019-0698-2.

45. Guo, C., Jeong, H.H., Hsieh, Y.C., Klein, H.U., Bennett, D.A., De Jager, P.L., Liu, Z., and Shulman, J.M. (2018). Tau Activates Transposable Elements in Alzheimer's Disease. Cell Rep. *23*, 2874–2880. https://doi.org/10.1016/j.celrep.2018.05.004.

46. Li, W., Jin, Y., Prazak, L., Hammell, M., and Dubnau, J. (2012). Transposable elements in TDP-43-mediated neurodegenerative disorders. PLoS One *7*, e44099. https://doi.org/10.1371/journal.pone.0044099.

47. Li, W., Lee, M.H., Henderson, L., Tyagi, R., Bachani, M., Steiner, J., Campanac, E., Hoffman, D.A., von Geldern, G., Johnson, K., et al. (2015). Human endogenous retrovirus-K contributes to motor neuron disease. Sci. Transl. Med. *7*, 307ra153. https://doi.org/10.1126/scitranslmed.aac8201.

48. MacGowan, D.J.L., Scelsa, S.N., Imperato, T.E., Liu, K.N., Baron, P., and Polsky, B. (2007). A controlled study of reverse transcriptase in serum and CSF of HIV-negative patients with ALS. Neurology *68*, 1944–1946. https://doi.org/10.1212/01.wnl.0000263188.77797.99.

49. Mayer, J., Harz, C., Sanchez, L., Pereira, G.C., Maldener, E., Heras, S.R., Ostrow, L.W., Ravits, J., Batra, R., Meese, E., et al. (2018). Transcriptional profiling of HERV-K(HML-2) in amyotrophic lateral sclerosis and potential implications for expression of HML-2 proteins. Mol. Neurodegener. *13*, 39. https://doi.org/10.1186/s13024-018-0275-3.

50. Steele, A.J., Al-Chalabi, A., Ferrante, K., Cudkowicz, M.E., Brown, R.H., Jr., and Garson, J.A. (2005). Detection of serum reverse transcriptase activity in patients with ALS and unaffected blood relatives. Neurology *64*, 454–458. https://doi.org/10.1212/01.WNL.0000150899.76130.71.

51. Krug, L., Chatterjee, N., Borges-Monroy, R., Hearn, S., Liao, W.W., Morrill, K., Prazak, L., Rozhkov, N., Theodorou, D., Hammell, M., and Dubnau, J. (2017). Retrotransposon activation contributes to neurodegeneration in a Drosophila TDP-43 model of ALS. PLoS Genet. *13*, e1006635. https://doi.org/10.1371/journal.pgen.1006635.

52. Ng, K.W., Boumelha, J., Enfield, K.S.S., Almagro, J., Cha, H., Pich, O., Karasaki, T., Moore, D.A., Salgado, R., Sivakumar, M., et al. (2023). Antibodies against endogenous retroviruses promote lung cancer immunotherapy. Nature *616*, 563–573. https://doi.org/10.1038/s41586-023-05771-9.

53. Chang, Y.H., and Dubnau, J. (2022). Endogenous Retroviruses and TDP-43 Proteinopathy Form a Sustaining Feedback to Drive the Intercellular Spread of Neurodegeneration. Preprint at bioRxiv. https://doi.org/10.1101/2022.07.20.500816.

54. Svedung Wettervik, T., Howells, T., Ronne-Engström, E., Hillered, L., Lewén, A., Enblad, P., and Rostami, E. (2019). High Arterial Glucose is Associated with Poor Pressure Autoregulation, High Cerebral Lactate/Pyruvate Ratio and Poor Outcome Following Traumatic Brain Injury. Neurocrit. Care *31*, 526–533. https://doi.org/10.1007/s12028-019-00743-2.

55. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.R., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296. https://doi.org/10.1038/s41592-019-0619-0.

56. Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287. https://doi.org/10.1089/omi.2011.0118.

57. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

58. Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. *42*, W187–W191. https://doi.org/10.1093/nar/gku365.

59. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26. https://doi.org/10.1038/nbt.1754.

60. Hayward, A., Cornwallis, C.K., and Jern, P. (2015). Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. Proc. Natl. Acad. Sci. USA *112*, 464–469. https://doi.org/10.1073/pnas.1414980112.

61. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. https://doi.org/10.1093/bioinformatics/btt656.

62. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit anti-STAT1, monoclonal	Abcam	Cat# ab109320; RRID:AB_10863383
Mouse anti-STAT1, monoclonal	Abcam	Cat# ab281999
Rabbit anti-OLIG2, polyclonal	Sigma-Aldrich	Cat# AB9610; RRID:AB_570666
Goat anti-OLIG2, polyclonal	R and D Systems	Cat# AF2418; RRID:AB_2157554
Chicken anti-Vimentin, polyclonal	Millipore	Cat# AB5733; RRID:AB_11212377
Mouse anti-S-100beta, monoclonal	Sigma-Aldrich	Cat# S2532; RRID:AB_477499
Rabbit anti-AQP4, polyclonal	Sigma-Aldrich	Cat# HPA014784; RRID:AB_1844967
Rabbit anti-GFAP, polyclonal	DAKO; Sigma-Aldrich	Cat# Z0334; RRID:AB_10013382
Chemicals, peptides, and recombinant proteins		
Laminin-521	Biolamina	Cat# LN521
Laminin-111	Biolamina	Cat# LN111-02
Penicillin-Streptomycin	GIBCO (ThermoFisher)	Cat# 15140122
StemPro Accutase	GIBCO (ThermoFisher)	Cat# A11105-01
StemMACS iPS-Brew XF, human	Miltenyi biotec	Cat# 130-104-368
DMEM/F12, w/Glutamax	Invitrogen	Cat# 31331-093
StemMACS SB431542 in solution	Miltenyi	Cat# 130-106-543
Human Noggin, research grade	Miltenyi	Cat# 130-103-456
Human SHH (C24II), premium grade	Miltenyi	Cat# 130-095-730
Human FGF-8b, premium grade	Miltenyi	Cat# 130-095-740
MACS NeuroMedium	Miltenyi	Cat# 130-093-570
StemMacs CHIR99021 in solution	Miltenyi	Cat# 130-106-539
N2 supplement	Invitrogen	Cat# 17502048
NB21 supplement	Miltenyi	Cat# 130-097-263
L-Glutamine	GIBCO (ThermoFisher)	Cat# 25030032
Rock-inhibitor, Y27632	Miltenyi Biotec	Cat# 130-106-538
Non-Essential Amino Acids (NEAA)	GIBCO (ThermoFisher)	Cat# 11140
CNTF, Research grade	Mintenyi	Cat# 130-108-972
BDNF, Research grade	Mintenyi	Cat# 130-096-286
GDNF, premium grade	Mintenyi	Cat# 130-129-542
L-ASCORBIC ACID	Sigma-Aldrich	Cat# A4403-100MG
Dibutyryl-cAMP	Sigma-Aldrich	Cat# D0627-1G
NB-21 supplement without vitamin A	Miltenyi biotec	Cat# 130-093-566
PBS	ThermoFisher	Cat# AM9625
EGF	Invitrogen	Cat# PHG0311
FGF2, premium grade	Miltenyi	Cat# 130-093-564
PFA	GIBCO (ThermoFisher)	Cat# 28908
IFN-γ, premium grade	Miltenyi biotec	Cat#130-096-482
Critical commercial assays		
Chromium Single Cell 3′ Library	10X Genomics	Cat# PN-1000268
Rneasy mini kit (miniRNeasy)	QIAGEN	Cat# 74104
TruSeq Stranded mRNA	Illumina	Cat# 20020594
Deposited data		
Raw and processed data of single-nuclei and bulk RNA-seq of the control postmortem samples and cell cultures	This paper	GSE209552

(*Continued on next page*)

Continued		
REAGENT or RESOURCE	SOURCE	IDENTIFIER
Visualization, statistical analyses, and preprocessing pipelines, including QC filtering for the single nuclei RNAseq and bulk RNAseq	This paper	https://github.com/raquelgarza/TBI_Garza_2023; https://doi.org/10.5281/zenodo.8392734
Experimental models: Cell lines		
Human Embryonic stem cells (RC17)	Roslin Cells	RRID:CVCL_L206
Software and algorithms		
Seurat	10.1016/j.cell.2019.05.031	RRID:SCR_007322
STAR	10.1093/bioinformatics/bts635	RRID:SCR_004463
TEcount	10.1093/bioinformatics/btv422	RRID:SCR_023208
featureCounts	10.1093/bioinformatics/btt656	RRID:SCR_012919
deeptools	10.1093/nar/gku365	RRID:SCR_016366
CellRanger	10.1038/ncomms14049	RRID:SCR_017344
Integrative Genomics Viewer	10.1038/nbt.1754	RRID:SCR_011793
DESeq2	10.1186/s13059-014-0550-8	RRID:SCR_015687
bamtofastq	NA	RRID: SCR_023215
subset-bam	NA	RRID:SCR_023216
trusTEr	10.5281/zenodo.7589548	
Other		
Paraffin hardware Tissue tek VIP; Sakura Finetek	Tissue-Tek	Cat# 62580-01
Buffered formalin	Histo-Lab Products AB	Cat# 02176
Blocking buffer/permeabilization buffer (Triton X-100)	Cell Signaling Technology	Cat# 39487
DAPI	Sigma-Aldrich	Cat# D9542

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact Johan Jakobsson (johan.jakobsson@med.lu.se).

Materials availability

No new materials were generated for the purpose of this paper.

Data and code availability

- Raw data of single-nuclei and bulk RNA-seq of the control postmortem samples and cell cultures, as well as processed data of all samples can be found at the GEO: GSE209552.
- Visualization, statistical analyses, and preprocessing pipelines, including QC filtering for the single nuclei RNAseq can be found at https://github.com/raquelgarza/TBI_Garza_2023 (https://doi.org/10.5281/zenodo.8392734). Code for the quantification of TEs per cluster from single nuclei RNAseq can be found at https://github.com/raquelgarza/truster (https://doi.org/10.5281/zenodo.7589548), documentation at https://raquelgarza.github.io/truster/.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Ethical statement

All clinical and experimental research described herein was approved by the regional ethical review board (decision numbers 2005/103, 2008/303, 2009/89, and 2010/379). Written informed consent was obtained from the TBI patients' closest relatives and from the patients themselves if they had sufficiently recovered from their injury at >6 months post injury.

Detailed demographic and clinical characteristics are shown in Table S1. Twelve patients with severe TBI, defined as post-resuscitation Glasgow Coma Scale score $\leq$8, were included. The patients were >18 years old, and no patient had any other known neurological disorder or Down syndrome. All patients had been mechanically ventilated and sedated and continuous measurements of

intracranial pressure (ICP) and cerebral perfusion pressure (CPP) were performed at the neurocritical care unit at Uppsala University Hospital.[54]

The age (mean ± SD) of the TBI patients (10 males, 2 females) was 49.5 ± 18.2 years. The tissue samples were obtained from patients suffering severe, life-threatening focal TBI. Due to space-occupying brain swelling causing midline shift and compression of the basal cisterns, or due to markedly increased intracranial pressure (ICP) refractory to medical neurointensive care measures, contused brain regions (typically the injured part of a temporal or frontal lobe) were surgically removed between 4 h and 8 days after injury.[21] There were no complications associated with the surgical method. The patients needed prolonged neurocritical, neurosurgical, and then neurorehabilitative care. At the time of follow-up, one patient was deceased. All available patients' outcomes are shown in Table S1.

METHOD DETAILS

Sampling and preparation of brain tissue

Surgically removed brain tissues were immediately placed in a sterile pre-labeled container and subsequently stored at −80°C until analyzed. Half of the tissue was put in a routinely-used fixative, 4% buffered formalin (Histo-Lab Products AB, Gothenburg, Sweden, catalog no. 02176). The samples were fixed for 24–72 h and then paraffin-embedded and processed by hardware Tissue tek VIP (Sakura, CA, USA). Small samples were taken from the fresh-frozen contused brain tissue for snRNA-seq.

Immunohistochemistry

Immunohistochemistry was performed on microtome sections according to previously published protocols.[17]

For OLIG2 and STAT1 two sets of antibodies were used with similar results. OLIG2 (R&D Systems, AF2418, 1:500), OLIG2 (Sigma-Aldrich, AB9610, 1:500), STAT1 (Abcam, ab109320, 1:500), STAT1 (Abcam, ab281999, 1:2000).

Confocal images were captures using a Leica TCS SP8 confocal laser-scanning microscope. The Operetta CLS (PerkinElmer) instrument and Harmony analysis software (version 4.9.) were used for high-content screening and analysis. We identified the number of DAPI+ (Sigma-Aldrich, 1:1,000), STAT1+ and OLIG2+ cells and defined the number of cells with STAT1 staining. Tissue samples from five individuals with TBI and four individuals from non-neurological deaths were stained for DAPI, OLIG2 and STAT1. Images were acquired from 37 to 382 fields using the 20× objective. Valid nuclei were defined by DAPI staining based on intensity and area. We excluded DAPI cells which were clumped together or where the separation of nuclei by the software was not efficient enough by setting a maximum area and shape. STAT1+ and OLIG2+ cells were identified using a threshold considering the background of the staining.

Single-nuclei RNA sequencing

The FACS-based isolation of nuclei from frozen brain tissue was performed as previously described and loaded onto 10x Genomics Single Cell 3′ Chip.[17] Sequencing library samples were multiplexed and sequenced on a Novaseq 6000 machine using a 150-cycle kit with the recommended read length from 10x Genomics.

Bulk RNA sequencing

Total RNA was isolated from nuclei using the Rneasy Mini Kit (Qiagen). Libraries were generated using Illumina TruSeq Stranded mRNA library prep kit (poly-A selection) and were sequenced on an Illumina NextSeq500 machine (paired-end 2 × 150 bp).

Generation of glial progenitors from hES cells

Human Embryonic stem cells (RC17) were maintained in IPS-brew on laminin 521 (0.5 μg/cm^2) coated plates. Cells were passaged every 5 days with 0.5 mM EDTA, followed by seeding at a density of 2,500 cells per cm2 with ROCK inhibitor (10 μM Y-27632- only first 24 h after plating). RC17 ES cells were differentiated toward ventral midbrain (vMB) fate based on a previously published protocol (Nolbrant et al., 2017). Post quality control for ventral midbrain specification, day 16 progenitors were maintained in a Neurobasal medium containing NB-21 supplement without vitamin A (1:500), penicillin/streptomycin (1:1000), l-glutamine (1:1000), NEAA (1:1000), FGF2 (20 ng mL−1) and EGF (100 ng mL−1) in suspension on a non-adherent flask. vMB progenitors self-organise and form homogeneous embryoid bodies (Ebs) in suspension. Ebs were dissociated every 18–20 days and maintained on a non-adherent flask for another 100 days for cells to undergo a gliogenic switch. Post 100 days of differentiation, cells were dissociated into single cells with accutase and plated on laminin 521 (2 μg/cm^2) coated plates for terminal differentiation medium (Neurobasal medium containing NB-21 supplement without vitamin A (1:500), penicillin/streptomycin (1:1000), l-glutamine (1:1000), NEAA (1:1000) and CNTF (50 ng mL−1)) for glial progenitors.

Immunocytochemistry

Cells were fixed in 4% PFA for 20 min, and post-fixation cells were washed with 1XPBS and incubated in blocking buffer/permeabilization buffer (Triton X-100) for 1 h. Cells were incubated with primary antibodies overnight in PBS. Antibodies: GFAP (DAKO, Z0334, 1:500), AQP4 (Sigma, HPA014784, 1:500), s100-Beta (Sigma, S2532, 1:500) and Vimentin (Millipore, AB5733, 1:500). After incubation with secondary antibodies and DAPI, images were acquired in Lieca SP8 confocal microscope.

IFN-γ stimulation of glial progenitors

RC17-derived glial progenitor cells were plated on laminin 521 (2 μg/cm^2) coated plates in a terminal differentiation medium. Cells were matured for 14 days and stimulated with IFN-γ (5 ng/mL) for 48hrs in the culture medium. Post stimulation, cells were washed twice with 1XPBS and detached with accutase and pellets were frozen down on dry ice for bulk-RNA seq.

QUANTIFICATION AND STATISTICAL ANALYSIS

Immunohistochemistry statistical analysis

The percentage of STAT1 positive cells among OLIG2 positive cells was calculated only considering cells with valid DAPI staining as:

$$100\left(\frac{STAT1\ positive\ cells}{OLIG2\ positive\ cells + STAT1\ positive\ cells}\right)$$

The data was normally distributed (p = 0.1397, Shapiro test for normality). We used Students t-test to identify the differences between the two groups. Statistical details specified on figure legend.

Single-nuclei RNA sequencing analysis

Raw base calls were demultiplexed to obtained sample-specific FastQ files and reads were aligned to GRCh38 genome assembly using the Cell Ranger pipeline (10x Genomics, Cellranger count v5; RRID:SCR_017344) with default parameters (–include-introns was used for nuclei mapping). The resulting matrix files were used for further downstream analysis. Seurat (version 3.1.1; RRID:SCR_007322) and R (version 3.4) were used for bioinformatics analysis.

Cells which have lower than three mean absolute deviation from the median number of reads present in the sample were removed (R function *scater::isOutlier,* parameters *type* = ''both'', *nmads = 3, log = TRUE).* Given the difference between quality of TBI and control tissue, a further cut off at least 1,000 detected genes per cell was implemented to ensure that only high quality cells were retained for downstream analysis. Data was log-normalized to reduce sequencing depth variability (LogNormalize, Seurat). All samples were merged to generate integrated UMAP using Harmony integration (RunHarmony, 'sample', Seurat).[55] For visualization and clustering, manifolds were calculated using UMAP methods (*RunUMAP,* Seurat) and 20 precomputed principal components and the shared nearest neighbor algorithm modularity optimization-based clustering algorithm (FindClusters, Seurat) with a resolution of 0.1. The clusters obtained were annotated using canonical marker gene expression. Differential gene expression between TBI and control samples across respective cell types was carried out using the Seurat function FindMarkers (Wilcoxon Rank-Sum test, padj <0.01). Cell-cycle scores were computed using the Seurat function CellCycleScoring. Gene ontology overrepresentation analysis was performed using the gseGO function in the clusterProfiler R package using differentially-expressed genes in each cell type (padj <0.01).[56] Enrichment scores for genes related to selected immune-related GO terms were calculated using the AddModuleScore function from Seurat using five features as controls (ctrl = 5).

TE quantification per cell cluster

To ease the usage of the workflow we made use of, we created a Python package named trusTEr (version 0.1.1; https://doi.org/10.5281/zenodo.7589548) to process single-cell RNA-seq data to quantify TE expression per cluster of cells. The package consists of three classes (Experiment, Sample, and Cluster) for the abstraction of the projects, and a module (jobHandler.py) to handle SLURM jobs. For more detailed information about how the classes are related, please refer to https://raquelgarza.github.io/truster/. The workflow has the following steps.

1. Extraction of reads from samples' BAM files (tsv_to_bam() in class Cluster). After the clustering formation, trusTEr expects one text file per cluster containing all barcodes of a sample in a cluster (done by default by get_clusters() and merge_clusters()). It will extract the given barcodes from the sample's BAM file (outs/possorted_genome_bam.bam output from Cell Ranger count) using the subset-bam software from 10x Genomics (version 1.0; RRID:SCR_023216). Outputs a BAM file per cluster containing all alignments from the cells in the cluster.
2. Filter duplicates (filter_UMIs() in class Cluster). Given the number of PCR duplicates we could be carrying in these BAM files, we wrote a Python script (filterUMIs) to filter most of these out. The filter_UMIs function keeps reads with unique combinations of cell barcodes, UMI, and sequence, so that only unique molecules are kept.
3. Convert to FastQ (bam_to_fastq() in class Cluster). Using bamtofastq from 10x Genomics (version 1.2.0; RRID: SCR_023215), the previous BAM files are output as fastQ files.
4. Concatenate lanes (concatenate_lanes() in class Cluster). Concatenates the different lanes from the same library as output from bamtofastq. This step outputs one fastQ file per cluster.
5. The quantification was performed in groups of samples (TBI and control groups) to increase statistical power. This step concatenates the fastQ files of the samples' clusters within a single group (see the groups parameters in the process_clusters function in the Experiment class, or in each of the steps' corresponding functions).
6. Map cluster (map_cluster() in class Cluster). Mapping the reads to a reference genome using STAR aligner (version 2.7.9a; RRID:SCR_004463).[57] For subfamily quantification (default in the map_cluster function) –outFilterMultimapNmax

100, –winAnchorMultimapNmax 200 were used. To quantify individual elements (argument unique set to True), unique mapping was performed using –outFilterMultimapNmax 1 and –outFilterMismatchNoverLmax 0.03. Visualization of tracks per cell type (Figures S5B and S5C) was performed using the uniquely mapped reads per cluster, grouped by condition. These were filtered by strand using deeptools bamCoverage (version 2.4.3; RRID:SCR_016366)[58] with –filterRNAstrand set to ''forward'' to get reverse transcription, and ''reverse'' to get forward transcription – as bamCoverage assumes a dUTP-based library preparation. Signal was normalized using a scale factor (–scaleFactor) of 1e+7 divided by the number of cells in the cluster. Tracks of cell types with more than one cluster were overlayed in the Integrative Genomics Viewer (IGV) (version 2.11.1; RRID:SCR_011793).[59] Matrices for upregulated HERVs and young L1s (Figure S5C) were performed using deeptools computeMatrixOperations (version 2.4.3; RRID:SCR_016366)[58] and visualized as profile plots using deeptools plotHeatmap (version 3.5.2; RRID:SCR_016366).[58]

7. TE count (TE_count() in class Cluster). TE quantification of the BAM files produced for each cluster. For the purposes of this paper, we used TEcount from the TEToolkit (version 2.0.3; RRID:SCR_023208) with the curated TE GTF for hg38 provided by the authors (–TE), gencode version 36 as the gene GTF (–GTF). We ran it in multi mode (–mode multi) as forward stranded (–stranded yes).[27]
8. Normalization of TE counts. TE quantification is normalized by cluster size (number of cells in a cluster), and it's stored within the Seurat object. The matrices output and the Seurat assay only includes the normalized TE subfamily information output by TEcount.

Bulk RNA sequencing analysis

Reads were mapped using STAR (2.6.0c; RRID:SCR_004463)[57] with GRCh38.p13 as the genome index, and gencode version 38 to guide the mapping (–sjdbGTFfile). To avoid ambiguous reads due to the quantification of ERVs, the number of loci allowed for a read to map (–outFilterMultimapNmax) was set to 1, and the number of mismatches allowed (–outFilterMismatchNoverLmax) was set to 3%.

The RetroTector software,[28] (https://github.com/PatricJernLab/RetroTector) was used to mine the human genome (version GRCh38/hg38 downloaded from https://hgdownload.soe.ucsc.edu/) for ERVs as previously described.[60] Predicted positions and structures for ERVs scoring 300 and above are summarized in Tables S2 and S3. Using the GTF version of the output file, read were quantified using featureCounts (Subread version 1.6.3; RRID:SCR_012919),[61] forcing matching strandness of the reads to the features being quantified (-s 2).

Read count matrices were then input to DESeq2 (version 1.28.1; RRID:SCR_015687) and fold changes shrunk using DESeq2lfcShrink.[62] Statistical details specified in figure or table legend. For further details, please refer to the corresponding Rmarkdown (TBI_bulk.Rmd) on github.

Gene quantification was performed using featureCounts (Subread version 1.6.3; RRID:SCR_012919),[61] forcing matching strands (-s 2). Gene differential expression analysis was then performed using DESeq2 (version 1.28.1; RRID:SCR_015687) and fold changes shrunk using DESeq2lfcShrink.[62] Statistical details specified in figure or table legend. Gene set enrichment analysis was performed using the gseGO function in the clusterProfiler R package.[56]

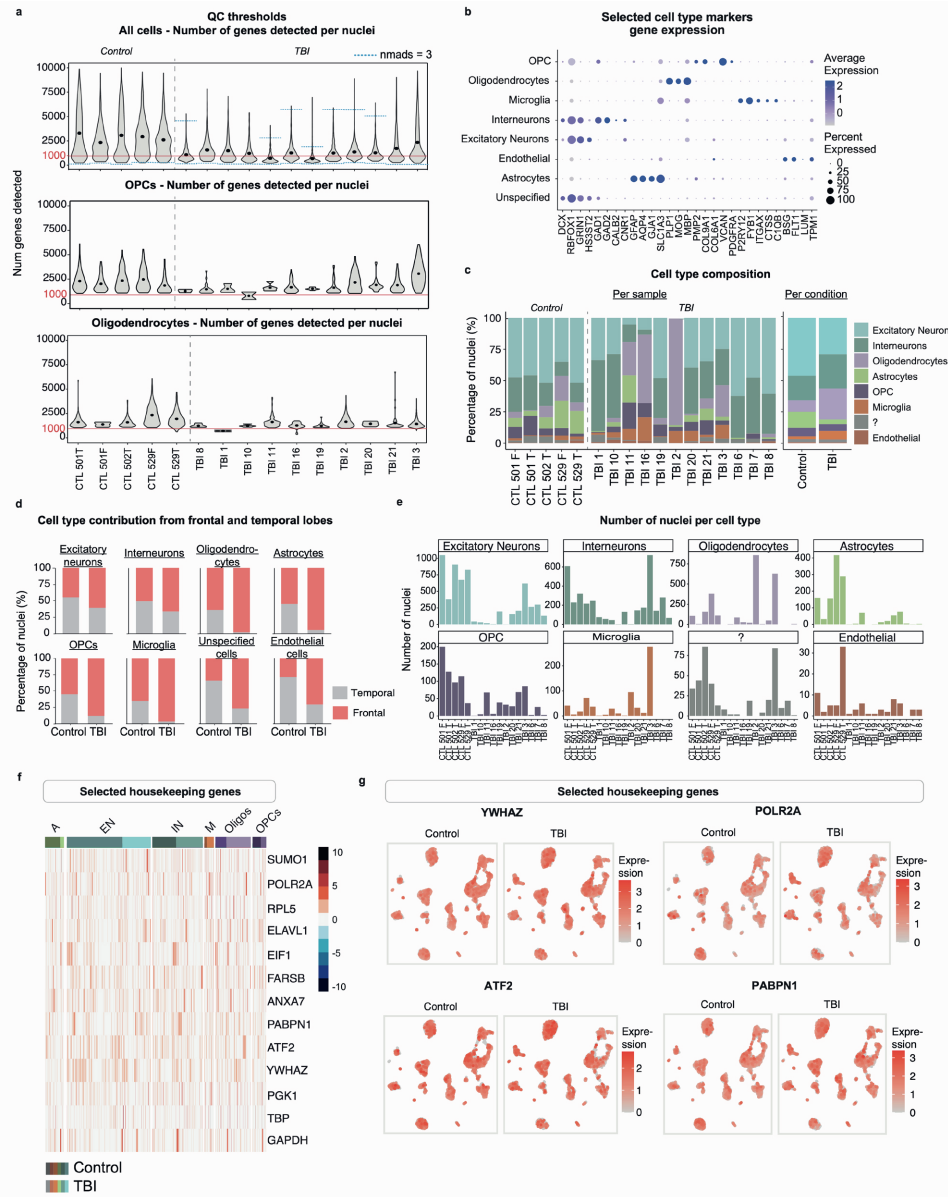
Supplemental information

Single-cell transcriptomics of human traumatic brain injury reveals activation of endogenous retroviruses in oligodendroglia

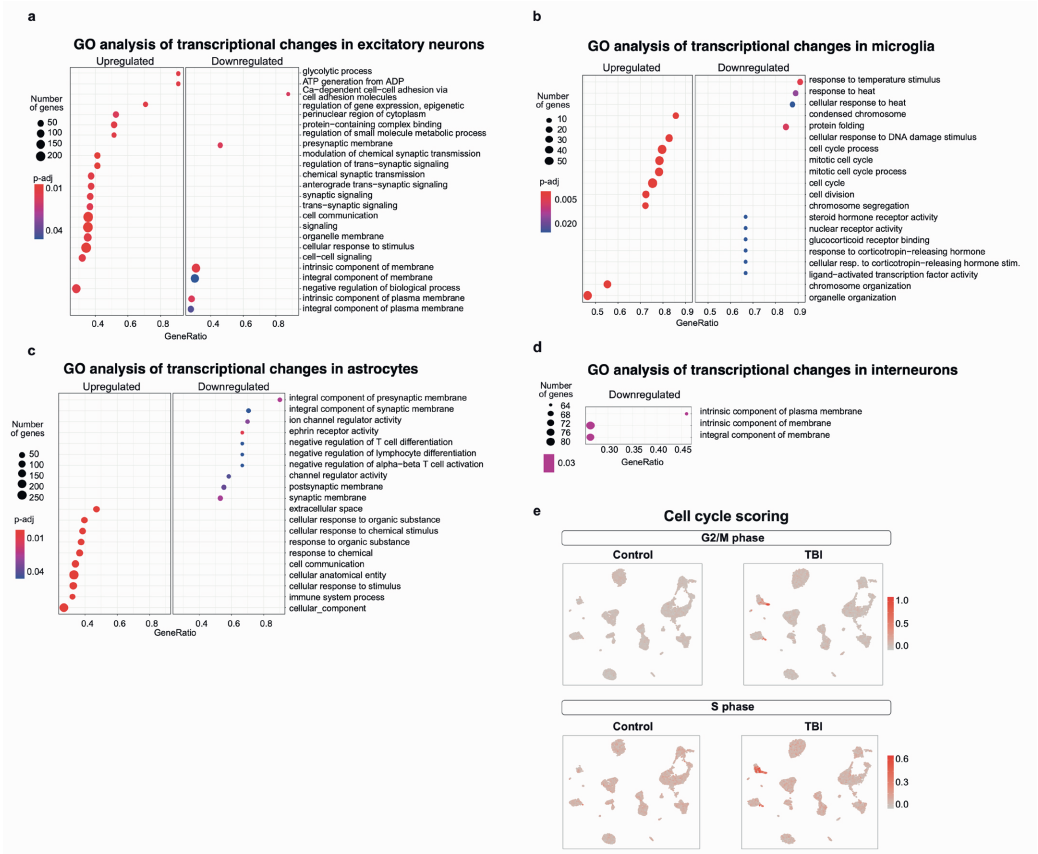
Raquel Garza, Yogita Sharma, Diahann A.M. Atacho, Arun Thiruvalluvan, Sami Abu Hamdeh, Marie E. Jönsson, Vivien Horvath, Anita Adami, Martin Ingelsson, Patric Jern, Molly Gale Hammell, Elisabet Englund, Agnete Kirkeby, Johan Jakobsson, and Niklas Marklund

Sample	Region	Age	Sex	Injury	Time post injury (h)	GOSE
TBI 1	Temporal lobe	22	M	MVA	9	3
TBI 2	Frontal lobe	72	M	Fall	4	7
TBI 3	Frontal lobe	42	M	MVA	4	8
TBI 6	Temporal lobe	74	M	Fall	4	4
TBI 7	Temporal lobe	58	M	Fall	9	7
TBI 8	Temporal lobe	49	M	Fall	84	1
TBI 10	Temporal lobe	65	M	Fall	180	3
TBI 11	Parietal lobe	25	M	SPR	24	3
TBI 16	Temporal lobe	52	M	Fall	42	4
TBI 19	Frontal lobe	49	F	HOB	57	-
TBI 20	Frontal lobe	62	F	MVA	192	8
TBI 21	Frontal lobe	24	M	Fall	106	6
CTL 501	Frontal lobe	87	M	-	-	-
CTL 501	Temporal lobe	87	M	-	-	-
CTL 502	Temporal lobe	75	M	-	-	-
CTL 529	Frontal lobe	69	M	-	-	-
CTL 529	Temporal lobe	69	M	-	-	-

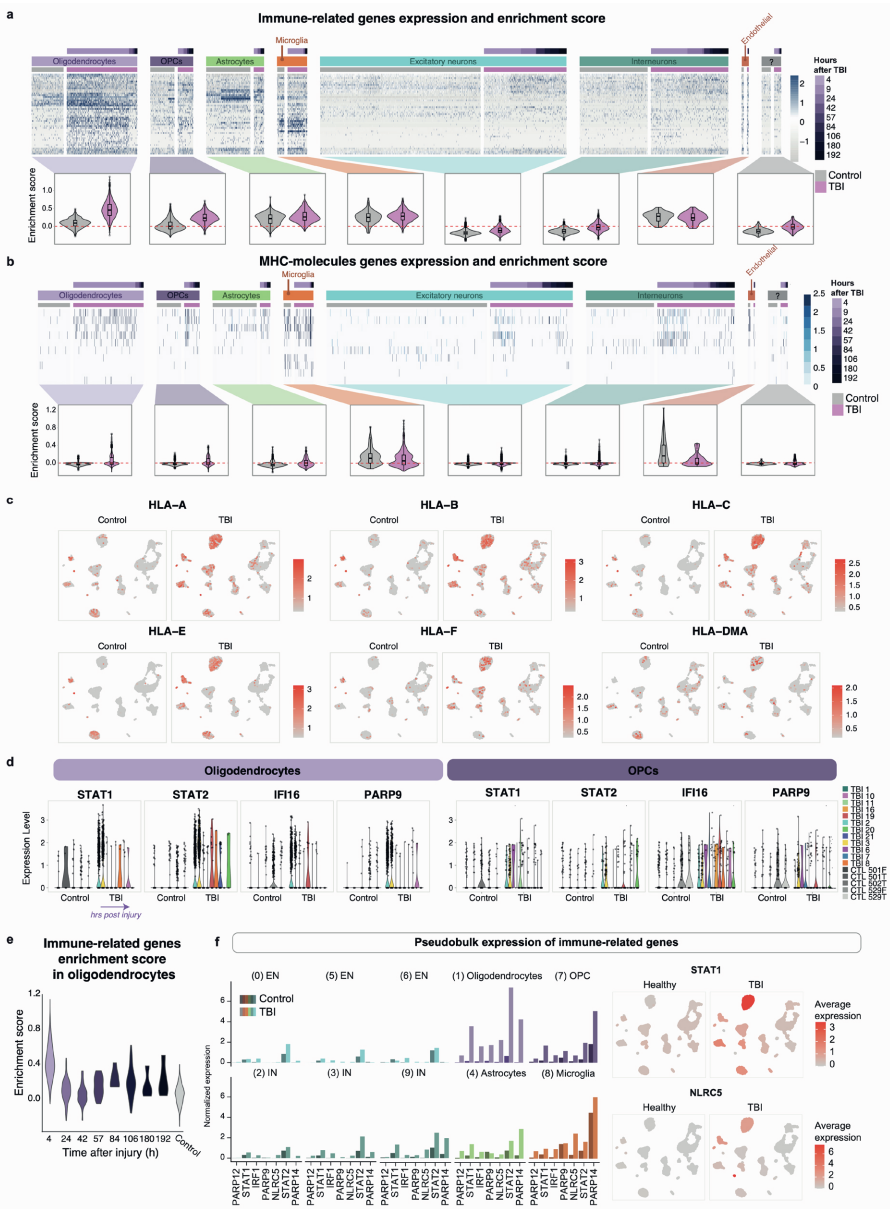
Suppl. Table 1 Details of TBI and control samples. *Demographics of TBI cases and controls, including the brain region of sample collection. MVA = motor vehicle accident, SPR = sports related, HOB = hit by object. GOSE = Glasgow Outcome Scale extended*



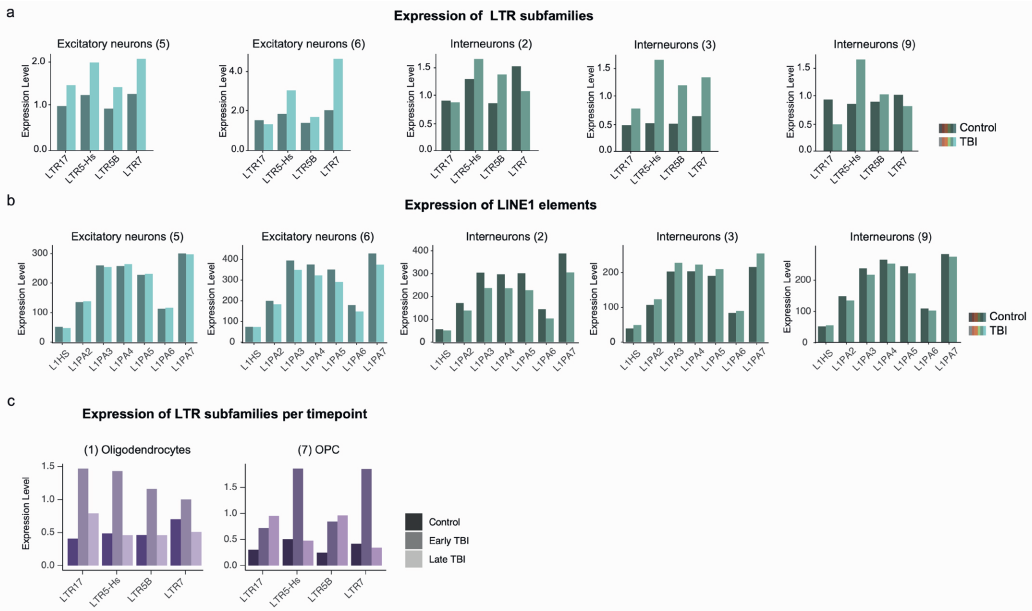
Suppl. Figure 1 Quality control of snRNA-seq data. a) Quality control thresholds for number of genes detected. Top: All cells, middle: OPCs, bottom: oligodendrocytes. Blue dotted lines indicating thresholds of nuclei excluded as outliers (isOutlier, scater, nmads = 3; see methods). Nuclei below red line (1,000 genes detected) were excluded from downstream analysis. b) Dot plot grouped by cell type showing the average gene expression and percentage of cell expressing the selected gene markers c) Cell-type ratio found in each sample and condition. d) Percentage of cells belonging to samples collected from the frontal or temporal lobe in each cell type. e) Number of nuclei per cell type in each sample. f) Heatmap showing level of expression of housekeeping genes through the different cell types (signal scaled by gene). Column annotation indicating condition (dark = Controls, light = TBI) and cell types (A = Astrocytes, EN = Excitatory neurons, IN = interneurons, M = microglia, Oligos = Oligodendrocytes, OPCs = Oligodendrocyte precursor cells). g) Selected housekeeping genes expression projected onto UMAP, split by condition.



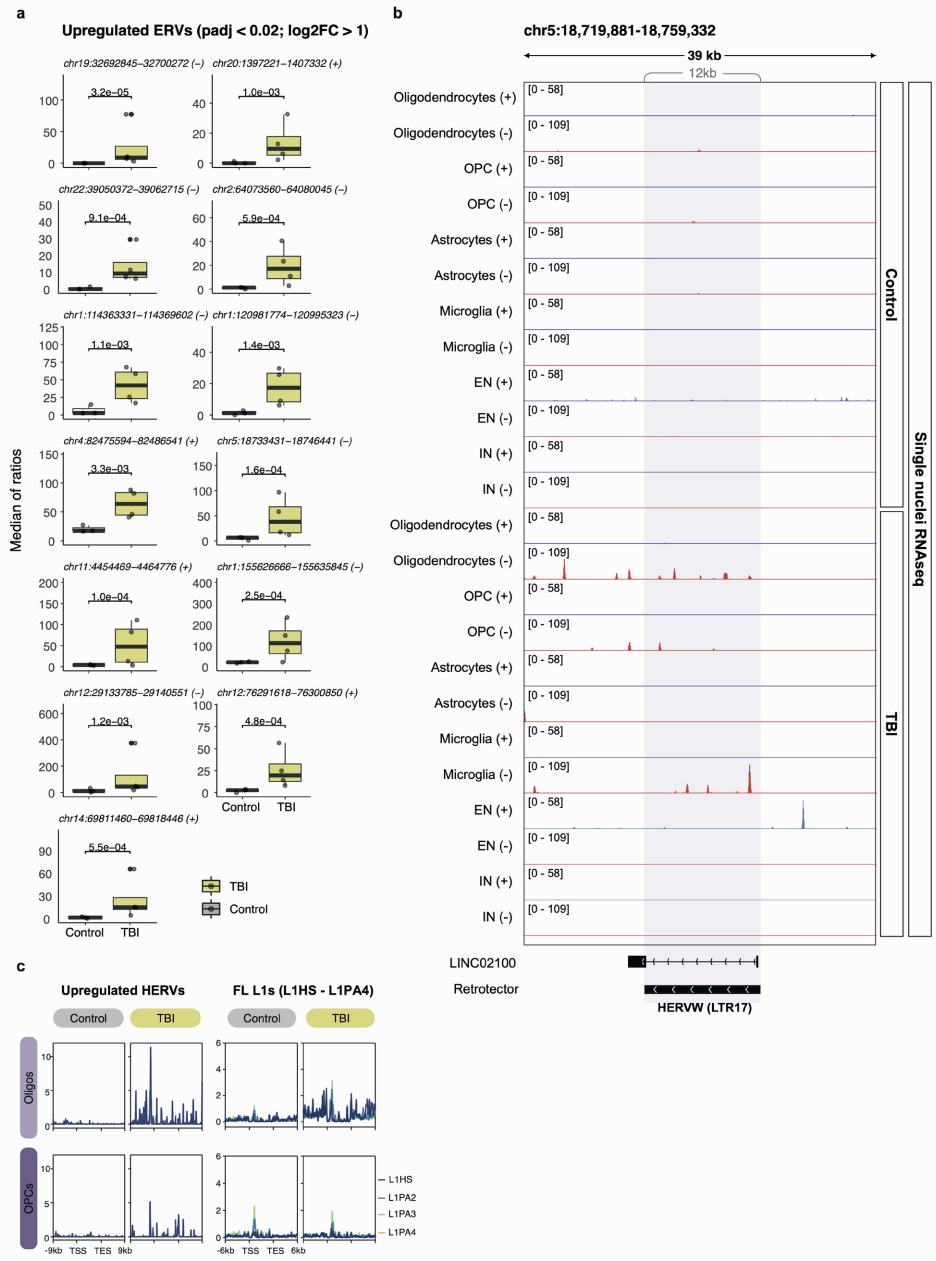
Suppl. Figure 2 Molecular cascade per cell type upon TBI. a) Gene set enrichment analysis (GSEA) of differentially-expressed genes (padj < 0.01) found in excitatory neurons, b) microglia, c) astrocytes, and d) interneurons. e) Enrichment score of cell-cycle-related genes (top G2/M phase, bottom S phase).



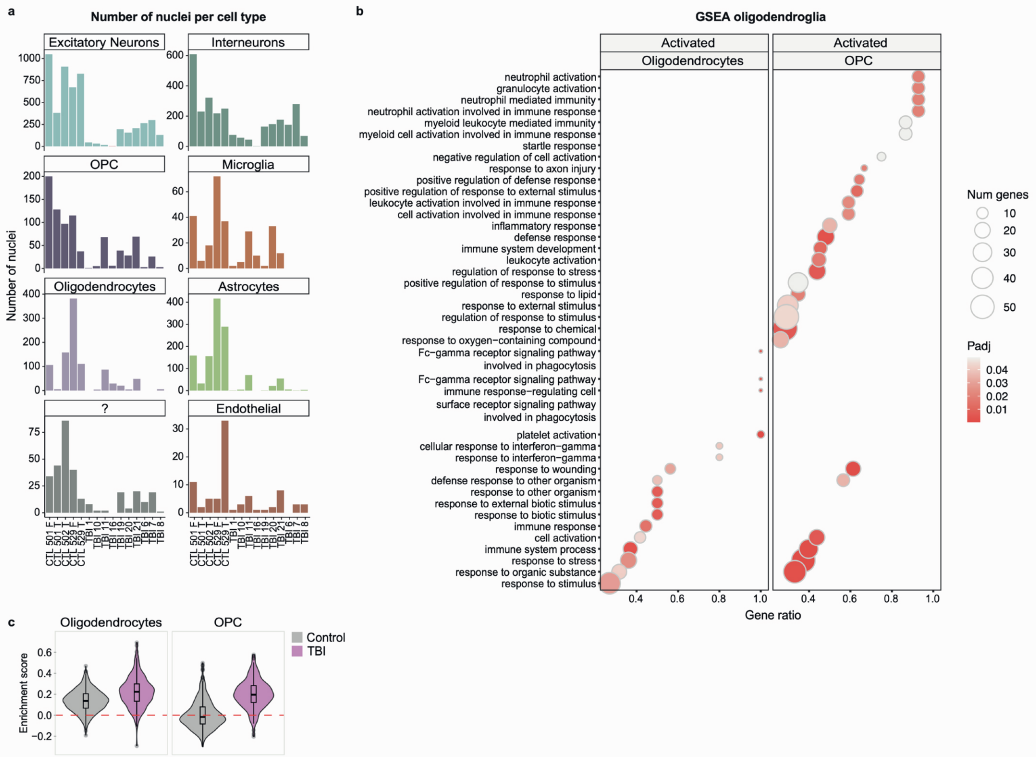
Suppl. Figure 3 Prevalent innate immune response in TBI oligodendroglia. a) Heatmap showing expression of selected genes (upregulated genes in oligodendrocytes related to Interferon-gamma-mediated signaling pathway, Response to interferon-gamma, Cellular response to interferon-gamma, Innate immune response, Cellular-response to cytokine stimulus, Response to cytokine, and Defense response), as shown in Figure 2a and violin plots showing enrichment scores per cell type and condition (AddModuleScore, Seurat). b) Heatmap showing expression of detected genes encoding for MHC molecules, and violin plots showing enrichment scores per cell type and condition (AddModuleScore, Seurat). c) Expression of selected MHC molecules. d)Expression of selected IFN-related genes among oligodendroglia in each sample. TBI samples sorted from left to right by hours after injury. e) Violin plots showing enrichment scores of immune-related genes per time point of injury (GO terms: Interferon-gamma-mediated signaling pathway, Response to interferon-gamma, Cellular response to interferon-gamma, Innate immune response, Cellular-response to cytokine stimulus, Response to cytokine, and Defense response). f) Pseudobulk quantification of gene expression showing immune-related genes (EN = Excitatory neurons, IN = interneurons).



Suppl. Figure 4 Quantification of TEs per cluster and time point. a) Expression of LTR subfamilies in clusters not shown in Figure 3d. b) Expression of LINE1 subfamilies in clusters not shown in Figure 3e. c) Expression of LTR subfamilies in oligodendroglia grouping nuclei by time point (early = 4-9hrs post injury, late = from 9hrs onwards).



Suppl. Figure 5 Validation of ERV proviruses activation in oligodendroglia. a) Boxplots showing normalized expression of upregulated ERVs per sample, grouped by condition (DESeq2 pvalue, unadjusted). b) Genome browser tracks showing an example of an upregulated HERV-W (Figure 4c). Cells were grouped by cluster and condition (TBI/Control) and uniquely mapped as pseudobulk. Tracks of clusters of same cell type are overlayed. Tracks marked as + (blue) or – (red) referring to forward and reverse transcription, respectively. c) Average signal over upregulated ERVs (left panel) and evolutionary young full-length L1s (right panel). Column annotation the condition of the sample (control/TBI; grey/yellow). Row annotation indicating cell type (oligos = oligodendrocytes in light purple, OPCs in purple). Different lines in the L1 panel indicate the specific L1-subfamily colour.



Suppl. Figure 6 Validation of results excluding outlier samples. a) Number of nuclei per cell type in each sample removing outlier samples (TBI 2 and 3). b) Selected activated terms from GSEA of differentially expressed genes (TBI vs control; padj < 0.01, Wilcoxon Rank Sum test (FindMarkers, Seurat)) in oligodendroglia (biological-process ontology). c) Violin plots showing enrichment scores in oligodendroglia (AddModuleScore, Seurat) of immune-related genes (as shown in Figure 2a: upregulated genes in TBI oligodendrocytes related to Interferon-gamma-mediated signaling pathway, Response to interferon-gamma, Cellular response to interferon-gamma, Innate immune response, Cellular-response to cytokine stimulus, Response to cytokine, and Defense response).

PAPER IV

Mini-heterochromatin domains constrain the cis-regulatory impact of SVA transposons in human brain development and disease

Vivien Horváth[1], Raquel Garza[1], Marie E. Jönsson[1], Pia A. Johansson[1], Anita Adami[1], Georgia Christoforidou[1,2], Ofelia Karlsson[1], Laura Castilla Vallmanya[1], Patricia Gerdes[1], Ninoslav Pandiloski1,[2], Christopher H. Douse[2], Johan Jakobsson[1]

Abstract

SVA retrotransposons remain active in humans and contribute to individual genetic variation. Polymorphic SVA alleles harbor gene-regulatory potential and can cause genetic disease. However, how SVA insertions are controlled and functionally impact human disease is unknown. Here, we dissect the epigenetic regulation and influence of SVAs in cellular models of X-linked dystonia-parkinsonism (XDP), a neurodegenerative disorder caused by an SVA insertion at the TAF1 locus. We demonstrate that the KRAB zinc finger protein ZNF91 establishes H3K9me3 and DNA methylation over SVAs, including polymorphic alleles, in human neural progenitor cells. The resulting mini-heterochromatin domains attenuate the cis-regulatory impact of SVAs. This is critical for XDP pathology; removal of local heterochromatin severely aggravates the XDP molecular phenotype, resulting in increased TAF1 intron retention and reduced expression. Our results provide unique mechanistic insights into how human polymorphic transposon insertions are recognized, and their regulatory impact constrained by an innate epigenetic defense system.

*Correspondence and lead contact:
Johan Jakobsson
Lund Stem Cell Center
Lund University
221 84 Lund, SWEDEN
Email: johan.jakobsson@med.lu.se
Phone: +46 46 2224225, +46709286443

Introduction

More than 50% of the human genome is made up of transposable elements (TEs) [1-3]. Three families of TEs are still active in humans: the autonomous long interspersed nuclear element 1 (LINE-1); the non-autonomous Alu short interspersed element (Alu); and the composite element SINE-R-VNTR-Alu (SVA) [4-9]. The mobilization of these elements represents a significant source of genomic variation in the human population and is the underlying cause of some genetic diseases [6,10-15].

SVAs are a class of hominoid-specific TEs. They are non-autonomous, depending on the LINE1 machinery for retrotransposition, and consist of a fusion of two TE fragments separated by a variable number of tandem repeats (VNTR) 16-19. Based on their evolutionary age, SVAs are divided into different subfamilies (A-F), of which SVA-E and SVA-F are human-specific [5]and make up about half of the approximately 3800 fixed SVAs annotated in the human genome [20]. In addition to the annotated SVAs, there are thousands of polymorphic SVA alleles in the human population. Current estimates suggest about one new germline SVA insertion in every 60 births [5,10,21-24]. The individual genetic variation caused by polymorphic SVA insertions is thought to contribute to phenotypic variation in the human population and contribute to, or cause, disease [11,12,25]. However, SVAs have been notoriously challenging to study due to their highly repetitive na-

[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]Epigenetics and Chromatin Dynamics, Department of Experimental Medical Science, Wallenberg Neuroscience Center and Lund Stem Cell Center, BMC A11, Lund University, 221 84 Lund, Sweden.

ture, and little is known about how polymorphic SVA insertions are regulated by the human genome or how they influence phenotypic traits and disease.

SVAs harbor strong gene regulatory sequences that can function both as transcriptional activators and repressors, influencing the expression of genes in the vicinity of their integration site [25-32]. Notably, SVAs appear to be particularly potent as cis-regulatory elements in the human brain, where they have been linked to enhancer-like activities [29,32]. In line with this, polymorphic SVAs have been linked to several genetic neurological disorders [23,33-37]. The most well-characterized of these is X-linked dystonia-parkinsonism (XDP), a recessive adult-onset autosomal genetic neurodegenerative disorder [38-40]. XDP is caused by a germline SVA retrotransposition event in intron 32 of TAF1, a gene that encodes TATA-box binding protein associated factor 1, an essential part of the transcriptional machinery [39,41]. The SVA insertion interferes with the transcription and/or splicing of TAF1 mRNA, resulting in reduced expression [39,41]. The example of XDP illustrates the important role of polymorphic SVA insertions in human brain disorders. However, although the SVA insertion is the underlying genetic cause of XDP, the molecular mechanism behind how the SVA interferes with TAF1 expression is unknown and there is still no mechanistic insight into why certain SVA insertions cause brain disorders. For example, there are hundreds of intronic SVA insertions in the human genome that do not cause disease. How is the human brain protected against the strong regulatory impact of SVAs in these cases and what makes the disease-causing SVA insertion in TAF1 unique?

In this study we demonstrate that the DNA-binding KRAB zinc finger protein (KZFP) ZNF91 plays a key role in protecting the human genome against the cis-regulatory impact of SVAs by establishing a dual layer of repressive epigenetic modifications over SVAs in neural cells, including new polymorphic alleles such as the disease-causing XDP-SVA. The resulting mini-heterochromatin domains are characterized by the presence of both DNA methylation and H3K9me3. Notably, the presence of ZNF91-mediated heterochromatin on the polymorphic XDP-SVA is highly relevant for XDP pathology, as the removal of this heterochromatin domain aggravates the molecular XDP phenotype, resulting in increased intron-retention and reduced TAF1 expression. In summary, our results provide unique mechanistic insights into how human polymorphic TE insertions are recognized, and how their potential regulatory impact in neural cells is minimized by an innate epigenetic defense system based on a KZFP.

Results

Establishing XDP-NPCs to study the epigenetic regulation of SVAs

To investigate the molecular mechanisms controlling SVAs in human neural cells, including the polymorphic XDP-SVA, we established a neural progenitor cell (NPC) model system using induced pluripotent stem cell (iPSC) lines derived from three XDP patients and three control individuals (Figure 1A, Table 1). The XDP-SVA carriers presented initially with dystonia at a mean age at onset of 42.6 years (±13.6), similar to what has previously been reported (42.3±8.3 years) (Table 1) [39,42]. The controls used were unaffected sons of two of the XDP-SVA carriers (Table 1). The six iPSC lines were converted into stable NPC lines [43] (XDP- and Ctrl-NPCs) that could be extensively expanded or differentiated into different neural cell types. The XDP- and Ctrl-NPCs exhibited NPC morphology and expressed NPC markers such as SOX2 and NESTIN, monitored with immunocytochemistry (Figure 1B, Figure S1A). The expression of NPC markers, as well as the lack of expression of pluripotency markers, was also confirmed by RNA-seq (Figure 1C).

The presence of the XDP-SVA insertion, which is ~2.6 kbp long and located in intron 32 of the TAF1 gene, was confirmed using PCR (Figure 1D, E). RNA-seq analysis confirmed that XDP-NPCs displayed a characteristic retention of intron 32 of TAF1 (p-val-

Table 1.

Status	Subject	Sample ID	Age at onset (years)	Age at collection (years)	Relationship
XDP	33363.C	XNPC 1	38	44	Father of CNPC 1
	33109.2B	XNPC 2	58	72	Father of CNPC 2,3
	32517.B	XNPC 3	32	35	Not related
Control	33362.C	CNPC 1	-	18	Son of XNPC 1
	33114.C	CNPC 2	-	34	Son of XNPC 2
	33113.2I	CNPC 3	-	42	Son of XNPC 2

Subjects as described in Ito et al. 2016 [94]

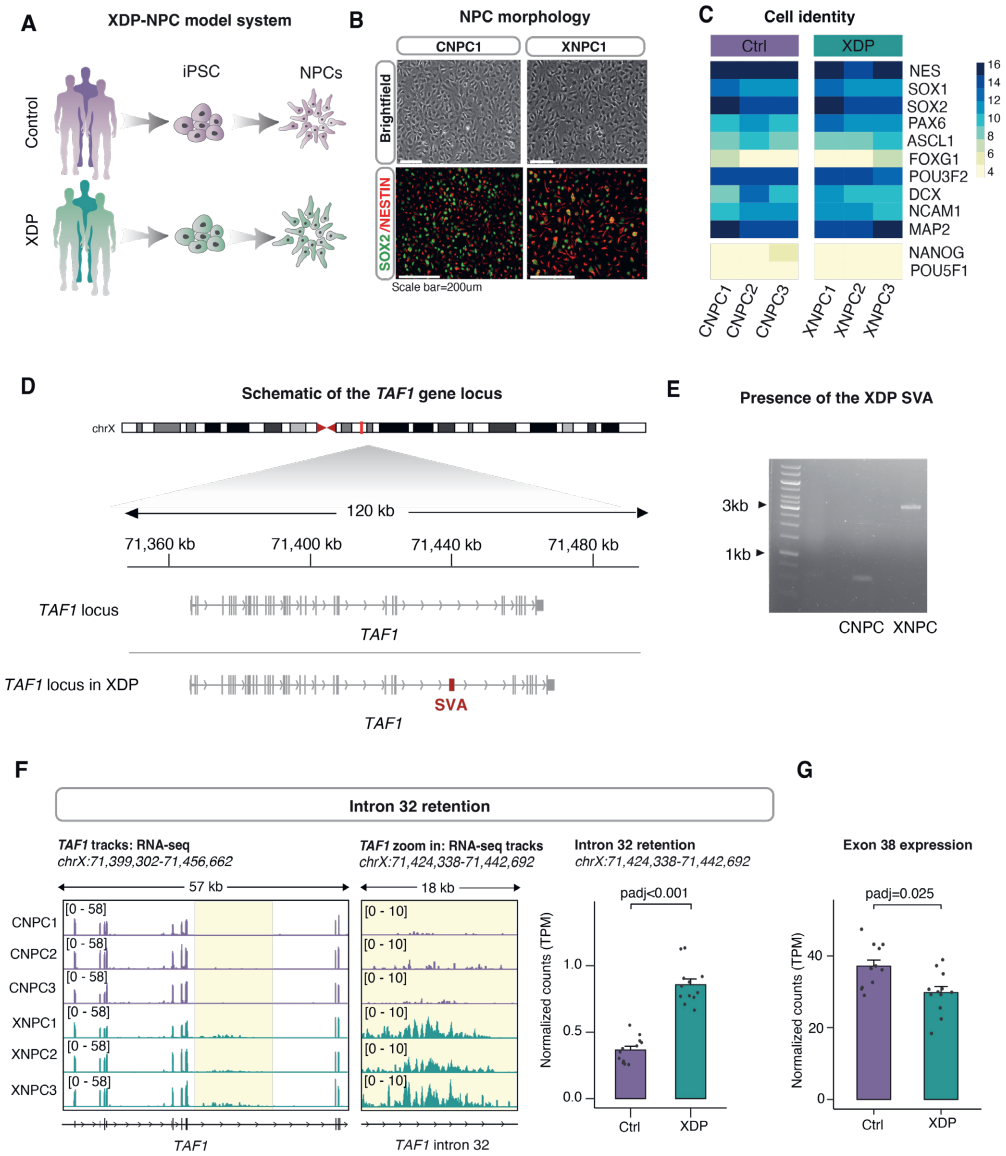


Figure 1. Characterization of the XDP-NPC model system.

(A) Schematic of the generation of XDP-NPCs. (B) Brightfield images of Ctrl- (CNPC1) and XDP-NPCs (XNPC1) (top). Immunocytochemistry (bottom) of Sox2 (green) and nestin (red) in Ctrl- and XDP-NPCs. (C) Heatmap of NPC marker-gene expression in Ctrl-NPCs (n=3) and XDP-NPCs (n=3) measured using RNA-seq. (D) Schematic of the TAF1 gene locus. The polymorphic XDP-SVA is depicted in red. (E) PCR analysis of genomic DNA identifying the XDP-SVA. (F) Genome browser tracks showing gene expression of the TAF1 gene (left) and a magnification of intron 32 of TAF1, highlighting the characteristic intron retention in XDP-NPCs. Quantification of TAF1 intron 32 retention (right) in Ctrl- (n=12) and XDP-NPCs (n=12) (padj, DESeq2). (G) Quantification of TAF1 exon 38 expression in Ctrl- (n=12) and XDP-NPCs (n=12) (padj, DESeq2).

ue<0.001, DESeq2) and lower expression of downstream TAF1 exons, such as exon 38, when compared to Ctrl-NPCs (p=0.025, DESeq2) (Figure 1F, G). These observations are similar to those previously described for XDP-iPSC and NPC lines [39].

SVAs are covered by H3K9me3 in NPCs

TEs, including SVAs, are associated with heterochromatin in somatic tissues that correlate with their transcriptional silencing, and which may impact their regulatory potential [44]. We chose to characterize the repressive histone mark H3K9me3, which is linked to heterochromatin, in fetal human forebrain tissue, two XDP-NPCs, and two Ctrl-NPCs using CUT&RUN analysis (Figure 2A). The computational analysis of histone marks on SVAs using CUT&RUN data is challenging due to their repetitive nature. This results in a large proportion of ambiguous reads. To avoid false conclusions due to multi-mapping artefacts, we used a strict unique mapping approach to investigate individual SVA elements (Figure 2A). With this bioinformatic approach, it is only possible to investigate the epigenetic status of the flanking regions of the SVAs where the unique genomic context allows us to discriminate reads without ambiguity, and the epigenetic modification can be traced to unique loci in the human genome. The boundaries of nearly all SVAs (>1 kbp in length) of the different subfamilies (A-F), both in the developing human forebrain and in NPCs, were enriched with H3K9me3 (Figure 2B, C, Figure S2A). However, the genomic context did not enable us to analyze the XDP-SVA with this approach. To resolve this issue, we developed a qPCR-based technique in combination with CUT&RUN (Figure 2A, see Materials and methods). This showed a clear enrichment of H3K9me3 at the boundary of the XDP-SVA in XDP-NPCs (Figure 2D, Figure S2C).

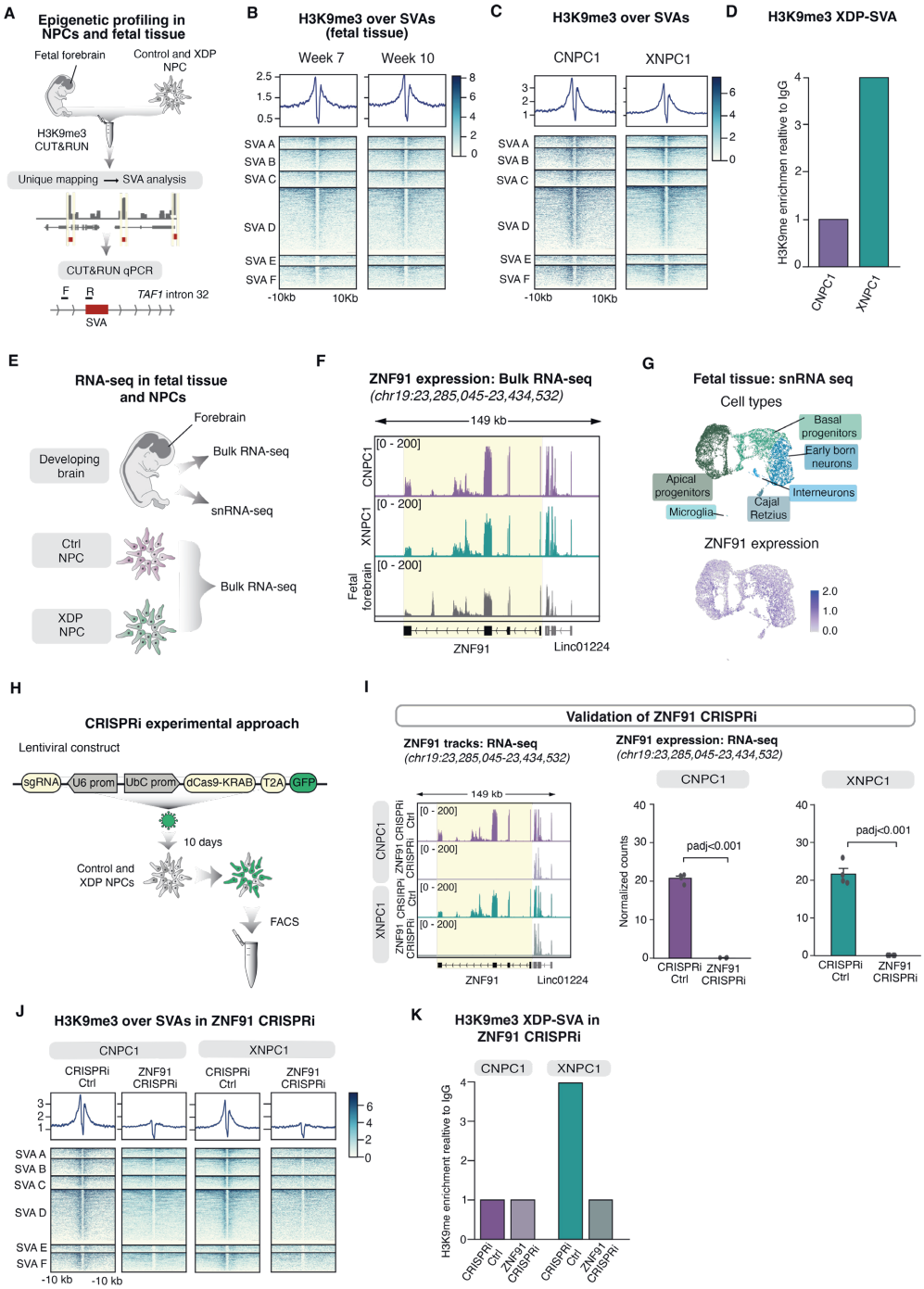
H3K9me3 deposition at SVAs is dependent on the KZFP ZNF91

To protect genomic integrity against TE insertions, organisms have evolved cellular defense mechanisms [45,46]. KZFP genes have amplified and diversified in mammalian species in response to transposon colonization 47-52; recent profiling efforts have identified several KZFPs that bind to SVAs, including ZNF91 and ZNF611 [29,50,53]. We noted that ZNF91 is highly expressed in human fetal forebrain tissue and XDP- and Ctrl-NPC cultures, as monitored by bulk and snRNA-seq (Figure 2E-G) [54]. Thus, we hypothesized that ZNF91 could be a KZFP that binds SVAs and recruits the epigenetic machinery that deposits H3K9me3 at these sites in NPCs.

To investigate a role for ZNF91 in SVA repression in NPCs, we designed a lentiviral CRISPR inhibition (CRISPRi) strategy to silence ZNF91 expression. We targeted two guide RNAs (gRNAs) to a genomic region located next to the ZNF91 transcription start site (TSS) and co-expressed gRNAs with a KRAB transcriptional repressor domain fused to catalytically-dead Cas9 (dCas9) (Figure 2H). As a control, we used a gRNA targeting lacZ, representing a sequence not found in the human genome. The transduction of XDP- and Ctrl-NPCs resulted in efficient silencing of ZNF91-expression, monitored with RNA-seq (Figure 2I, Figure S2D). CUT&RUN analysis of ZNF91-CRISPRi NPCs (XDP- and Ctrl-NPCs) revealed almost complete loss of H3K9me3 around SVAs (Figure 2J). This finding was reproduced in one ad-

Figure 2. ZNF91 is required for H3K9me3 maintenance at SVAs in NPCs.

(A) Schematic of CUT&RUN approaches to profiling H3K9me3 at SVAs in NPCs and human fetal forebrain tissue. (B) Heatmap showing enrichment of H3K9me3 over SVAs in human fetal forebrain tissue. (C) Heatmap showing H3K9me3 enrichment in NPCs. Displayed are the genomic regions spanning ±10 kbp up and downstream from the element. (D) Barplots showing the enrichment of H3K9me3 over the XDP-SVA and the lack of enrichment in control samples. (E) Schematic of RNA-seq and snRNA-seq experiments in NPCs and human fetal forebrain tissue. (F) RNA-seq tracks of ZNF91 expression in NPCs and fetal forebrain. (G) UMAP showing characterized cell types (top). UMAP representing ZNF91 expression (bottom) in different cell types in the fetal brain. (H) Schematic of the CRISPRi approach including the lentiviral construct and experimental design. (I) RNA-seq tracks (left) and quantification (right) of ZNF91 expression in CRISPRi-Ctrl and ZNF91-CRISPRi in Ctrl-NPCs (n=4) and XDP-NPCs (n=4) (padj, DESeq2). (J) Heatmap showing H3K9me3 over SVAs in CRISPRi-Ctrl and ZNF91-CRISPRi in Ctrl-NPCs and XDP-NPCs. (K) Barplots showing the effect of ZNF91-CRISPRi on H3K9me3 over the XDP-SVA in XDP-NPC.


A
Epigenetic profiling in NPCs and fetal tissue
Fetal forebrain
Control and XDP NPC
H3K9me3 CUT&RUN
Unique mapping
SVA analysis
CUT&RUN qPCR
F
R
TAF1 intron 32
SVA
B
H3K9me3 over SVAs (fetal tissue)
Week 7
Week 10
SVA A
SVA B
SVA C
SVA D
SVA E
SVA F
-10kb
10Kb
C
H3K9me3 over SVAs
CNPC1
XNPC1
D
H3K9me3 XDP-SVA
H3K9me enrichmen realitve to IgG
E
RNA-seq in fetal tissue and NPCs
Forebrain
Developing brain
Bulk RNA-seq
snRNA-seq
Ctrl NPC
XDP NPC
F
ZNF91 expression: Bulk RNA-seq
(chr19:23,285,045-23,434,532)
149 kb
[0 - 200]
Fetal forebrain
ZNF91
Linc01224
G
Fetal tissue: snRNA seq
Cell types
Basal progenitors
Early born neurons
Apical progenitors
Interneurons
Microglia
Cajal Retzius
ZNF91 expression
H
CRISPRi experimental approach
Lentiviral construct
sgRNA
U6 prom
UbC prom
dCas9-KRAB
T2A
GFP
10 days
Control and XDP NPCs
FACS
I
Validation of ZNF91 CRISPRi
ZNF91 tracks: RNA-seq
(chr19:23,285,045-23,434,532)
ZNF91 expression: RNA-seq
(chr19:23,285,045-23,434,532)
Normalized counts
padj<0.001
CRISPRi Ctrl
ZNF91 CRISPRi
J
H3K9me3 over SVAs in ZNF91 CRISPRi
K
H3K9me3 XDP-SVA in ZNF91 CRISPRi

ditional XDP-NPC line and one additional Ctrl-NPC line (Figure S2E). CUT&RUN qPCR confirmed that the XDP-SVA also lost H3K9me3 in a ZNF91-dependent manner in XDP-NPCs (Figure 2K). Using a similar CRISPRi strategy we also confirmed that the H3K9me3 at SVAs, including the XDP-SVA, also depend on TRIM28, an epigenetic corepressor protein that is essential for the repressive action of KZFPs, in Ctrl- and XDP-NPCs (Figure S2F-H) [47,55]. Together, these results demonstrate that a ZNF91/TRIM28-dependent mechanism establishes local H3K9me3 heterochromatin over SVAs in human NPCs, including the polymorphic disease-causing XDP-SVA.

SVAs are covered by DNA methylation in human NPCs

In addition to H3K9me3, TE silencing in somatic tissues has also been extensively linked to DNA CpG-methylation [46,56-59]. To investigate the presence of DNA methylation on SVAs in NPCs, we performed genome-wide methylation profiling using Oxford Nanopore Technologies (ONT) long-read sequencing (Figure 3A) [60,61] on one XDP-NPC line (XNPC1) and one Ctrl-NPC line (CNPC1). The long-read DNA methylation analysis revealed that the SVA elements of different subfamilies (A-F), which are CG-rich sequences, were all heavily methylated in human NPCs (Figure 3B). In addition, the polymorphic XDP-SVA was fully covered by DNA methylation (Figure 3C). Furthermore, we performed Cas9-targeted ONT sequencing over the XDP-SVA on the ZNF91-CRISPRi NPCs and CRISPRi-Ctrl XDP-NPCs (Figure 3D). These results demonstrated that the XDP-SVA was fully methylated in both the ZNF91-CRISPRi and CRISPRi-Ctrl XDP-NPCs (Figure 3E). Thus, SVAs in human NPCs, including the XDP-SVA, are covered by both DNA methylation and H3K9me3. Our results also indicate that the presence of DNA methylation at SVAs is not dependent on ZNF91-binding or H3K9me3 in this cell type.

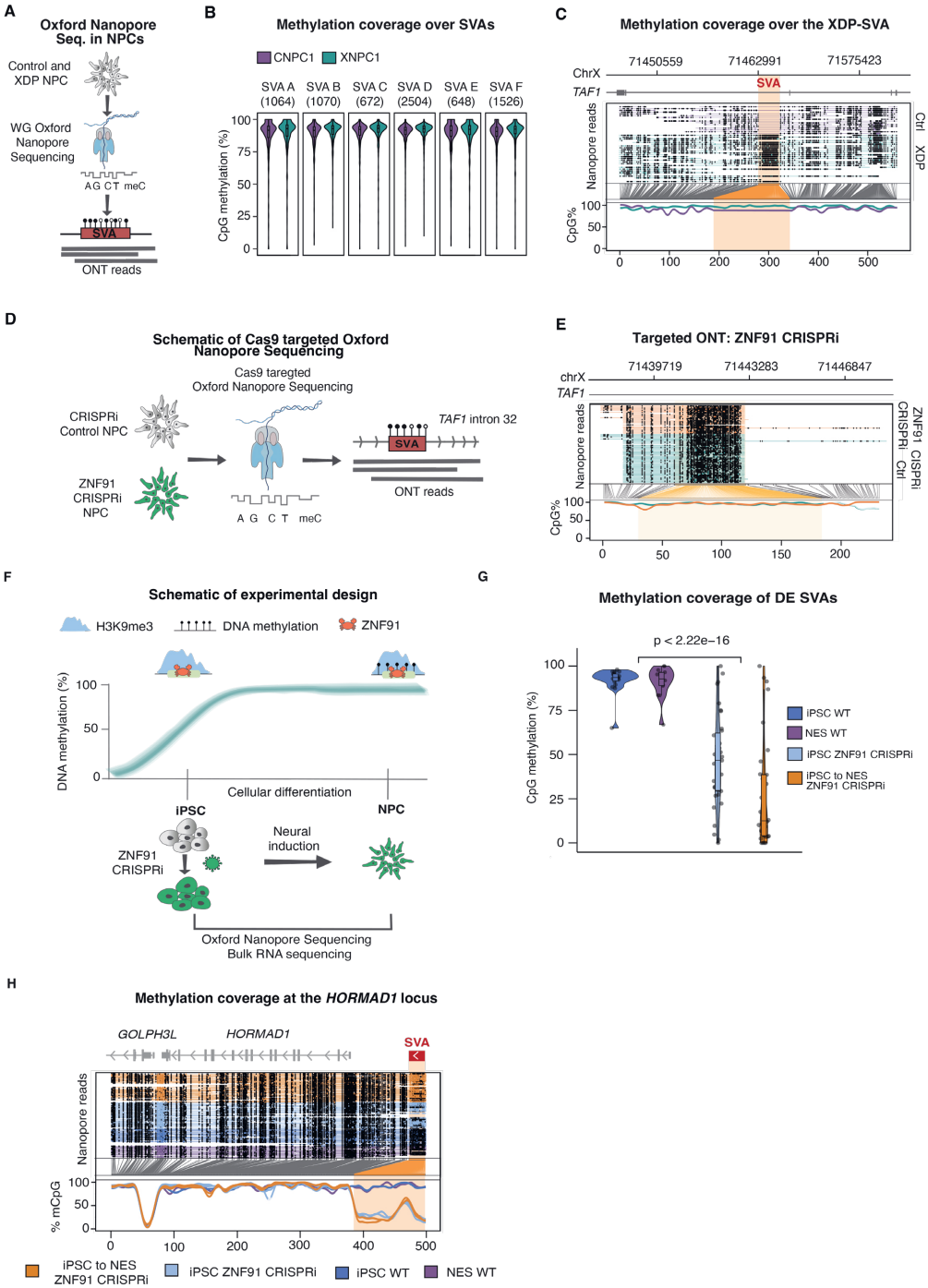
ZNF91 establishes DNA methylation on SVAs during early human development

Since the presence of DNA methylation on SVAs in NPCs did not depend on ZNF91, we wondered how and when DNA methylation on SVAs is established. DNA methylation is reprogrammed during the first few days of early development where global DNA methylation patterns, including that of many TEs, are erased and reinstated [62-66]. During this process, TEs are initially silenced by dynamic epigenetic mechanisms which are then gradually replaced by other [67], more stable epigenetic mechanisms in somatic cell types such as NPCs 55,68-71. We and others have implicated TRIM28-KZFP complexes in this process [55,69,71-75]. Thus, we hypothesized that ZNF91/TRIM28 may be involved in dynamically establishing DNA methylation of SVAs during earlier phases of human embryonic development (Figure 3F).

To test this hypothesis, we used iPSCs that resemble the epiblast stage of early human development[76], where DNA methylation patterns are more dynamically regulated (Figure 3F). In contrast, NPCs are somatic cells with a stably methylated genome [59,66,77]. We found clear evidence of dynamic ZNF91-mediated DNA methylation patterning of SVAs when we generated ZNF91-CRISPRi iPSCs. By performing genome-wide ONT analysis we found numerous SVAs (n=39) where DNA methylation was lost upon inhibition of ZNF91 (Figure 3G). In contrast, the same SVAs were covered by DNA methylation in control iPSCs and NPCs (Figure 3G). Notably, when we differentiated the ZNF91-CRISPRi iPSCs into NPCs we found that these SVAs remained hypomethylated (Figure 3G). Thus, without ZNF91 expression DNA methylation could not be established on these SVAs upon differ-

Figure 3. SVAs are covered by DNA methylation in NPCs.

(A) Schematic of ONT sequencing experiment to monitor DNA methylation over SVAs. (B) Methylation coverage over SVAs in Ctrl- and XDP-NPCs. The different SVA families (A-F) are shown. (C) Methylation coverage over the XDP-SVA in Ctrl- and XDP-NPCs. (D) Schematic of Cas9 targeted ONT sequencing. (E) Targeted ONT sequencing in CRISPRi-Ctrl and ZNF91-CRISPRi NPCs. The TAF1 XDP-SVA locus is shown. (F) Schematic of DNA methylation patterns during development in iPSCs and NPCs. ZNF91-CRISPRi in iPSCs and conversion to NPCs is also shown. (G) Violin plot showing DNA methylation over the first quarter (from their transcription start site) of the differentially-expressed SVAs (p-value, Student's t-test). (H) DNA methylation pattern over an SVA element near the HORMAD1 gene.


A
Oxford Nanopore Seq. in NPCs
Control and XDP NPC
WG Oxford Nanopore Sequencing
A G C T meC
SVA
ONT reads
B
Methylation coverage over SVAs
CNPC1
XNPC1
SVA A (1064)
SVA B (1070)
SVA C (672)
SVA D (2504)
SVA E (648)
SVA F (1526)
CpG methylation (%)
C
Methylation coverage over the XDP-SVA
71450559
71462991
71575423
ChrX
TAF1
SVA
Nanopore reads
Ctrl
XDP
CpG%
D
Schematic of Cas9 targeted Oxford Nanopore Sequencing
Cas9 taregted Oxford Nanopore Sequencing
CRISPRi Control NPC
ZNF91 CRISPRi NPC
TAF1 intron 32
SVA
ONT reads
A G C T meC
E
Targeted ONT: ZNF91 CRISPRi
71439719
71443283
71446847
chrX
TAF1
Nanopore reads
ZNF91 CRISPRi
CRISPRi
Ctrl
CpG%
F
Schematic of experimental design
H3K9me3
DNA methylation
ZNF91
DNA methylation (%)
Cellular differentiation
iPSC
NPC
ZNF91 CRISPRi
Neural induction
Oxford Nanopore Sequencing
Bulk RNA sequencing
G
Methylation coverage of DE SVAs
p < 2.22e−16
CpG methylation (%)
iPSC WT
NES WT
iPSC ZNF91 CRISPRi
iPSC to NES ZNF91 CRISPRi
H
Methylation coverage at the HORMAD1 locus
GOLPH3L
HORMAD1
SVA
Nanopore reads
% mCpG
iPSC to NES ZNF91 CRISPRi
iPSC ZNF91 CRISPRi
iPSC WT
NES WT

entiation. For example, an SVA-F element located upstream of the HORMAD1 gene was fully methylated in control NPCs and iPSCs. The DNA methylation over this SVA was completely lost in ZNF91-CRISPRi iPSCs, and remained absent when the ZNF91-CRISPRi iPSCs were differentiated into NPCs (Figure 3H). These results demonstrate that the DNA methylation patterns over some SVAs are dynamic in iPSCs and depend on ZNF91. In addition, ZNF91 is essential for establishing the stable layer of DNA methylation found over these SVAs in NPCs. Thus, cellular context is important for the downstream consequence of ZNF91 binding to SVAs. In early development, ZNF91 mediates the establishment of both H3K9me3 and DNA methylation, while in somatic cells only H3K9me3 depends on ZNF91; DNA methylation is propagated through other mechanisms.

DNA methylation and H3K9me3 co-operate to silence SVA expression in NPCs

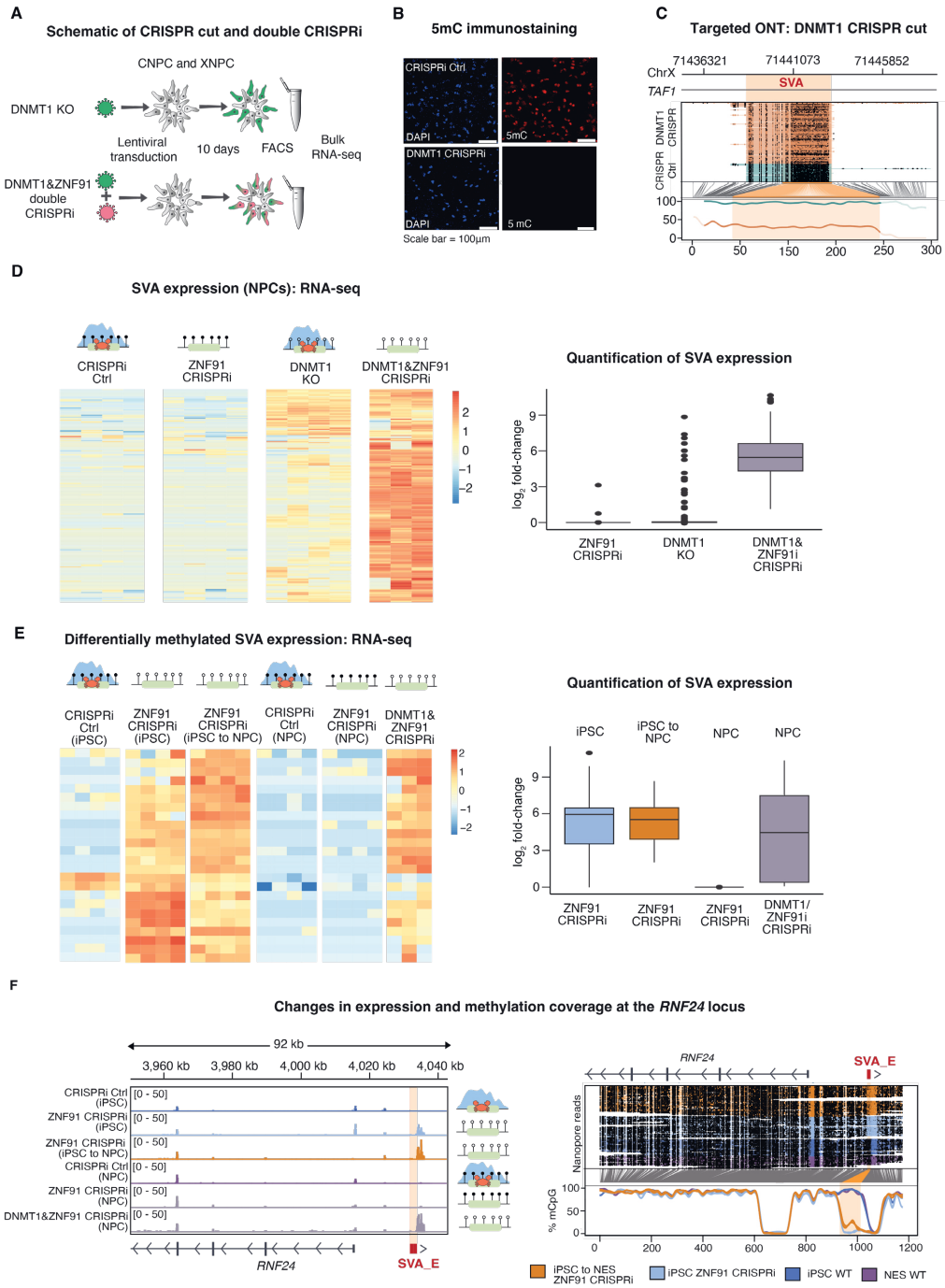
To investigate the role of H3K9me3 and DNA methylation in the transcriptional silencing of SVAs in NPCs we combined loss-of-function experiments with RNA-seq analysis. To remove H3K9me3 we used the ZNF91-CRISPRi NPCs. To remove DNA methylation we deleted DNA methyltransferase 1 (DNMT1), which is the enzyme that maintains DNA methylation during cell division [78]. We used a previously-described CRISPR-cut approach, resulting in a global loss of DNA methylation including over SVAs [59,79] as well as a CRISPRi combination strategy targeting the expression of both DNMT1 and ZNF91 in XDP- and Ctrl-NPCs (Figure 4A). Both approaches resulted in a global loss of DNA methylation, as monitored with 5mC immunocytochemistry (Figure 4B, Figure S3A), and a loss of DNA methylation over the XDP-SVA as demonstrated by targeted ONT long-read sequencing methylation analysis (Figure 4C). CUT&RUN analysis on DNMT1-KO NPCs revealed that loss of DNA methylation did not affect the presence of H3K9me3 at SVAs (Figure S3B).

We used an in-house 2×150bp, polyA-enriched stranded library preparation for bulk RNA-seq using a reduced fragmentation step to optimize the read length for SVA analysis. Such reads can be uniquely assigned to many SVA loci. We obtained ~40 million reads per sample. To quantify SVA expression we discarded all ambiguously mapping reads and only quantified those that map uniquely to a single location (unique mapping) [80]. We found that ZNF91-CRISPRi in NPCs, which removes H3K9me3 at SVAs, did not result in activation of SVA expression (Figure 4D). When DNMT1 was deleted in NPCs, which removes DNA methylation, we also found only a small number of SVAs transcriptionally upregulated (Figure 4D). However, in the ZNF91&DNMT1-CRISPRi NPCs, where both H3K9me3 and DNA methylation over SVAs is lost, we found a massive transcriptional activation of hundreds of SVAs (Figure 4D).

We also analyzed the expression of SVAs in the iPSC-NPC conversion experiments (Figure 3F). RNA-seq revealed that the SVAs that lost DNA methylation after ZNF91 deletion in iPSCs (ZNF91-CRISPRi iPSCs, Figure 3G) were also transcriptionally upregulated (Figure 4E). This contrasted with ZNF91-CRISPRi NPCs, where the same SVA elements were not upregulated upon inhibition of ZNF91 (Figure 4E). When we analyzed the ZNF91-CRISPRi iPSCs that were differentiated to NPCs we found that the SVAs were expressed in these NPCs (Figure 4E, F). These SVAs were also found to be upregulated upon ZNF91&DNMT1-CRISPRi in NPCs (Figure 4E). One example was an SVA-E element located upstream of the RNF24 gene (Figure 4F). This SVA was transcriptionally silent in control iPSCs and control NPCs. In ZNF91-CRISPRi iPSCs we detected a robust activation of the expression of this SVA-E element that

Figure 4. DNA methylation and H3K9me3 co-operate to silence SVAs in NPCs.

(A) Schematic of CRISPR cut and double CRISPRi experiment in NPCs. (B) 5mC immunostaining showing the global loss of DNA methylation upon DNMT1-CRISPRi. (C) DNA methylation coverage over the XDP-SVA in CRISPR-Ctrl and DNMT1 CRISPR-cut conditions. (D) Heatmap (left) showing upregulated SVAs in ZNF91&DNMT1-CRISPRi. The same SVAs are also shown in ZNF91-CRISPRi and DNMT1-KO experiments. Boxplot (right) showing SVA expression in ZNF91-CRISPRi, DNMT1-KO, and ZNF91&DNMT1-CRISPRi. I Heatmap (left) showing the expression level of differentially-methylated SVAs. Boxplot (right) showing the expression of differentially-expressed SVAs. (F) Genome browser tracks and DNA methylation pattern over an SVA element near to the RNF24 gene.


A
Schematic of CRISPR cut and double CRISPRi
CNPC and XNPC
DNMT1 KO
Lentiviral transduction
10 days
FACS
Bulk RNA-seq
DNMT1&ZNF91 double CRISPRi
B
5mC immunostaining
CRISPRi Ctrl
DAPI
5mC
DNMT1 CRISPRi
DAPI
5 mC
Scale bar = 100µm
C
Targeted ONT: DNMT1 CRISPR cut
71436321
71441073
71445852
ChrX
SVA
TAF1
DNMT1 CRISPR
CRISPR Ctrl
100
50
0
0
50
150
200
250
300
D
SVA expression (NPCs): RNA-seq
CRISPRi Ctrl
ZNF91 CRISPRi
DNMT1 KO
DNMT1&ZNF91 CRISPRi
2
1
0
−1
−2
Quantification of SVA expression
log2 fold-change
9
6
3
0
ZNF91 CRISPRi
DNMT1 KO
DNMT1& ZNF91i CRISPRi
E
Differentially methylated SVA expression: RNA-seq
CRISPRi Ctrl (iPSC)
ZNF91 CRISPRi (iPSC)
ZNF91 CRISPRi (iPSC to NPC)
CRISPRi Ctrl (NPC)
ZNF91 CRISPRi (NPC)
DNMT1& ZNF91 CRISPRi
2
1
0
−1
−2
Quantification of SVA expression
iPSC
iPSC to NPC
NPC
NPC
log2 fold-change
9
6
3
0
ZNF91 CRISPRi
ZNF91 CRISPRi
ZNF91 CRISPRi
DNMT1/ ZNF91i CRISPRi
F
Changes in expression and methylation coverage at the RNF24 locus
92 kb
3,960 kb
3,980 kb
4,000 kb
4,020 kb
4,040 kb
CRISPRi Ctrl (iPSC) [0 - 50]
ZNF91 CRISPRi (iPSC) [0 - 50]
ZNF91 CRISPRi (iPSC to NPC) [0 - 50]
CRISPRi Ctrl (NPC) [0 - 50]
ZNF91 CRISPRi (NPC) [0 - 50]
DNMT1&ZNF91 CRISPRi (NPC) [0 - 50]
RNF24
SVA_E
RNF24
SVA_E
Nanopore reads
% mCpG
100
50
0
0
200
400
600
800
1000
1200
iPSC to NES ZNF91 CRISPRi
iPSC ZNF91 CRISPRi
iPSC WT
NES WT

correlated with the loss of DNA methylation. When the ZNF91-CRISPRi iPSCs were differentiated to NPCs, the SVA remained expressed; this also correlated with a lack of DNA methylation. Thus, the loss of DNA methylation patterns over SVAs in iPSCs upon ZNF91-CRISPRi correlates with the transcriptional activation of SVAs, including when these cells are differentiated to NPCs. These experiments demonstrate that ZNF91 dynamically represses the expression of at least some SVAs in iPSCs, and is essential for establishing stable transcriptional repression of these SVAs.

DNA methylation and H3K9me3 co-operate to protect the human genome from the cis-regulatory influence of SVAs

SVAs carry regulatory sequences which can mediate cis-acting transcriptional effects on the surrounding genome [26-31]. We therefore investigated whether the ZNF91-mediated heterochromatin domains found over SVAs in NPCs influenced this activity. When investigating transcriptional changes of genes monitored via RNA-seq upon removal of H3K9me3 (ZNF91-CRISPRi), removal of DNA methylation (DNMT1-KO), or removal of both repressive marks (ZNF91&DNMT1-CRISPRi) we only found profound effects on nearby gene expression in the ZNF91&DNMT1-CRISPRi NPCs. The expression of genes located in the vicinity of an SVA element were significantly increased upon DNMT1&ZNF91-CRISPRi but not when deleting only one of the factors (Figure 5A). This effect could be detected when the SVA was located up to 50 kbp from the TSS, but was stronger when the SVA was closer to the TSS (Figure 5A).

Notably, the dynamics of the SVA-mediated influence on gene expression was distinct between different loci. Most genes in the vicinity of an SVA were completely unaffected by ZNF91 deletion or DNMT1 deletion alone, but transcriptionally upregulated when both factors were removed (Figure 5A). Thus, in most instances the presence of one of the heterochromatin marks was sufficient to protect flanking genomic regions from the regulatory impact of SVAs. However, we also found examples where both marks were needed to block the regulatory impact of SVAs. For example, the expression of HORMAD1 was upregulated due to the activation of an upstream SVA-F element acting as an alternative promoter in both ZNF91-CRISPRi and DNMT1-KO NPCs (Figure 5B). When both ZNF91 and DNMT1 were inhibited, HORMAD1 expression was even more strongly activated, suggesting a co-operative mode of action (Figure 5B). This demonstrates that at some loci both epigenetic marks are necessary to block the regulatory impact from SVAs at some loci.

Loss of H3K9me3 and DNA methylation over the XDP-SVA results in an aggravated molecular phenotype at the TAF1 locus

We next used the XDP-NPCs to investigate if the presence of H3K9me3 and DNA methylation over the XDP-SVA has any impact on TAF1 expression. Removing H3K9me3 alone (ZNF91-CRISPRi) did not affect intron retention in the TAF1 loci in the XDP-NPCs nor exon 38 expression of the TAF1 gene (Figure 5C, D). When we investigated the TAF1 loci in XDP-NPCs that lacked DNA methylation (DNMT1-KO), retention of intron 32 was significantly increased and exon 38 expression was reduced (Figure 5C, D). Removing both DNA methylation and H3K9me3 (ZNF91&DNMT1-CRISPRi) had an even stronger effect on TAF1 expression, including a considerable increase in intron 32 retention of TAF1 and lower expression of exon 38 in XDP-NPCs (Figure 5C, D). Thus, the loss of both DNA methylation and H3K9me3 aggravates the molecular pathology in XDP-NPCs. These data demonstrate that the regulatory impact of the polymorphic XDP-SVA is negatively influenced by the presence of a local mini-heterochromatin domain. When this heterochromatin domain is lost, the cis-regulatory effect of the XDP-SVA is strongly and significantly enhanced.

Polymorphic SVA insertions are silenced by ZNF91/H3K9me3 and DNA methylation

To investigate whether the local heterochromatin observed over the polymorphic XDP-SVA represented a unique event or if it was a general effect, we extended our analysis to other polymorphic SVAs in the genomes of two of the individuals in this study. We took advantage of the whole genome ONT long-read sequencing data from the XNPC1 and CNPC1

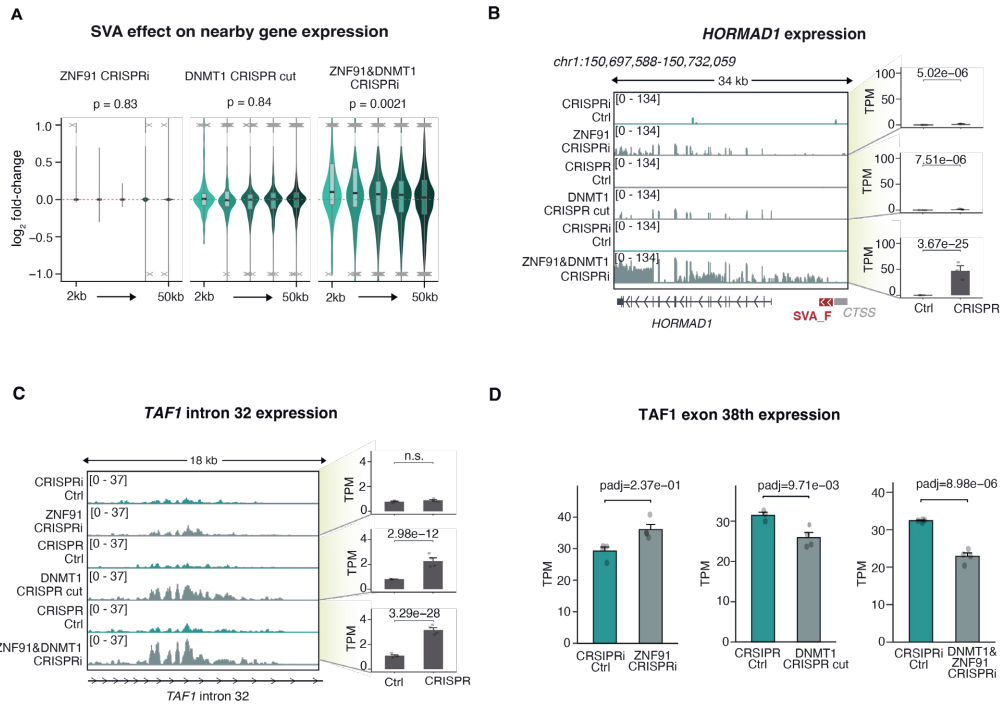


Figure 5. SVAs have a regulatory influence on nearby genes when heterochromatin marks are lost.

(A) Violin plot showing the effect of SVAs on nearby gene expression (2-50 kbp) in ZNF91-CRISPRi, DNMT1-KO, and ZNF91&DNMT1-CRISPRi (One-way ANOVA). (B) Genome browser tracks (left) showing HORMAD1 expression in ZNF91-CRISPRi, DNMT1-KO, and ZNF91&DNMT1-CRISPRi. Barplots (right) showing HORMAD1 expression in ZNF91-CRISPRi, DNMT1-KO, and ZNF91&DNMT1-CRISPRi (padj, DESeq2). (C) Genome browser tracks (left) showing TAF1 intron 32 expression in ZNF91-CRISPRi, DNMT1-KO and ZNF91&DNMT1-CRISPRi. Barplots (right) showing TAF1 intron 32 expression in ZNF91-CRISPRi, DNMT1-KO and ZNF91&DNMT1-CRISPRi (padj, DESeq2). (D) Barplots showing TAF1 exon 38 expression in ZNF91-CRISPRi, DNMT1 CRISPR-cut, and DNMT1&ZNF91-CRISPRi (padj, DESeq2).

lines and used the Transposons from Long DNA Reads (TLDR) pipeline to identify non-reference SVA insertions (Figure 6A) [61]. We identified 22 high-confidence polymorphic insertions of the SVA-E and -F subfamilies (average 2.5 kbp in length, range 1.23.8 kbp), of which 14 were shared between the two genomes (Figure 6B). Notably, several of these polymorphic SVAs, which represent recent TE insertions into the germline of these two individuals, displayed clear hallmarks of local heterochromatinization, including the presence of DNA methylation and H3K9me3.

For example, we found a polymorphic SVA insertion present only in XNPC1 in an intron of SLC12A6, which is a gene encoding a potassium/chloride transporter linked to neurological disorders [81,82]. This SVA insertion site displayed H3K9me3 at its boundaries and was fully covered by DNA methylation (Figure 6C). ZNF91-CRISPRi led to loss of H3K9me3 over this SVA, while ZNF91&DNMT1-CRISPRi in XNPC1 led to its transcriptional activation, resulting in the expression of an antisense readthrough transcript extending into the SLC12A6 gene (Figure 6C). Another example was a polymorphic SVA insertion

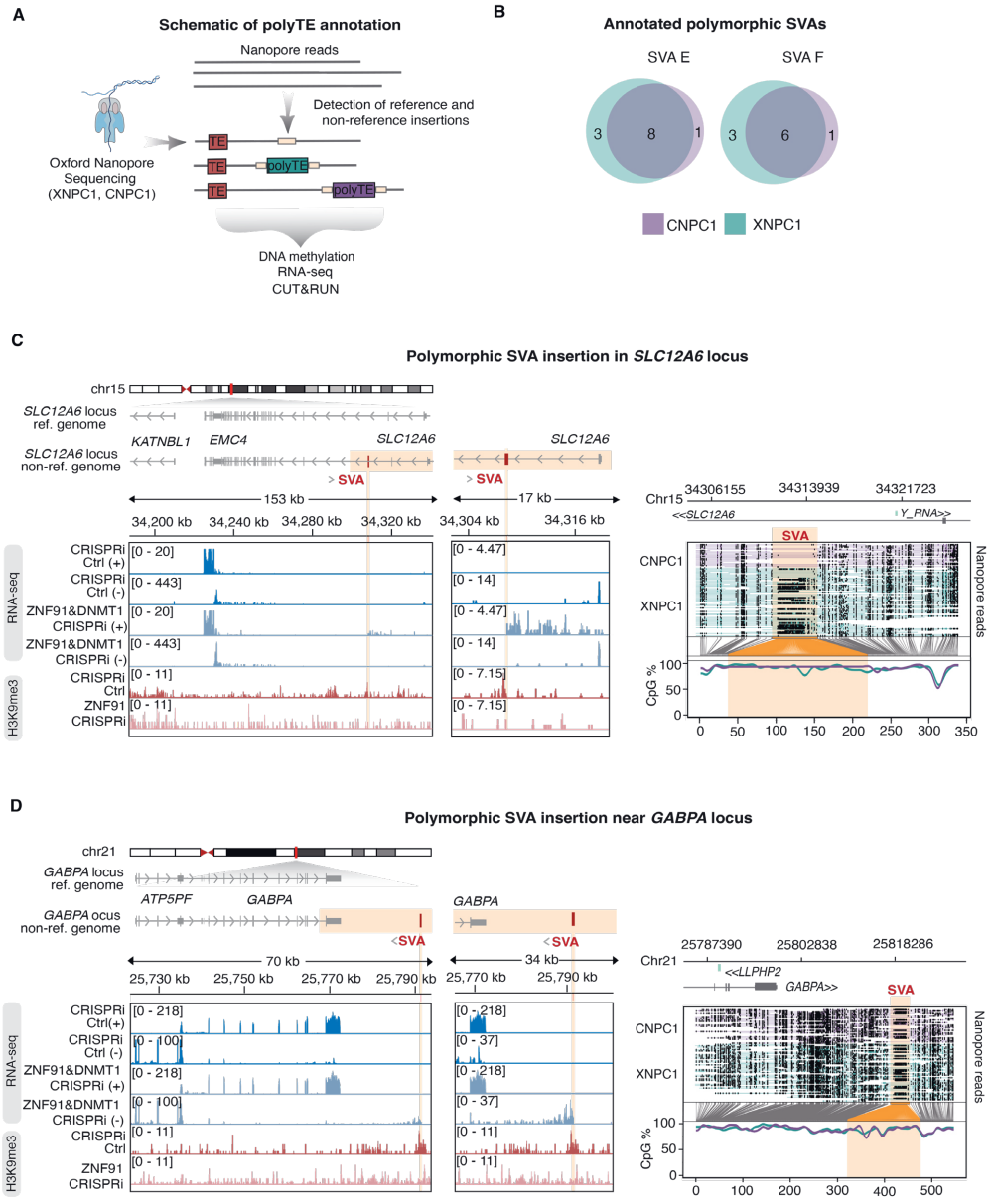


Figure 6. Polymorphic SVA insertions are repressed by DNA methylation and H3K9me3.

(A) Schematic of ONT sequencing and annotation of polymorphic SVAs. (B) Venn diagram showing annotated polymorphic SVA insertions. (C) Genome browser tracks (left) showing gene expression and H3K9me3 over a polymorphic SVA insertion and the nearby gene SLC12A6. ONT sequencing data (right) showing DNA methylation over the annotated polymorphic insertion. (D) Genome browser tracks (left) showing gene expression and H3K9me3 over a polymorphic SVA insertion and the nearby gene GABPA. ONT sequencing data (right) showing DNA methylation over the annotated polymorphic insertion.

shared between CNPC1 and XNPC1 downstream of GABPA, which is a gene encoding a DNA binding protein involved in mitochondrial function [83,84]. Similarly, the SVA insertion results in the accumulation of H3K9me3 and DNA methylation; ZNF91-CRISPRi and ZNF91&DNMT1-CRISPRi resulted in the loss of H3K9me3 and transcriptional activation of the SVA respectively, generating a readthrough antisense transcript (Figure 6D).

These data confirm that recent, polymorphic germline SVA insertions are recognized by ZNF91 in human NPCs and are covered by a dual layer of repressive epigenetic marks. Loss of this local heterochromatin domain results in transcriptional activation of these elements and the formation of novel transcripts, which are likely to have a regulatory impact on nearby genes. In these cases this is illustrated by the production of polymorphic antisense transcripts to SLC12A6 and GABPA.

Discussion

Our data support a model in which the KZNF ZNF91 binds to SVAs in early human development and throughout brain development, here modeled using iPSCs and NPCs. ZNF91 binding results in the establishment of local heterochromatin over SVAs, characterized by DNA methylation and H3K9me3, in a TRIM28-dependent manner. In early development (here represented by iPSCs) both modifications are dependent on the binding of ZNF91 to SVAs. In later phases of development, after epigenetic reprogramming of the genome (here represented by NPCs) only H3K9me3 depends on ZNF91: DNA methylation over SVAs is propagated by DNMT1. These two repressive chromatin modifications work together to limit the influence of SVAs on the host genome. They prevent the expression of the SVA elements and restrict the cis-acting influence of SVAs on the surrounding regions in neural cells. It is worth noting that this mechanism does not only involve evolutionarily older SVA insertions that are fixed in the human population, but also include recent polymorphic germline SVA insertions including the disease-causing XDP-SVA.

SVAs carry regulatory sequences with the potential to provide strong cis-acting influences on gene regulatory networks [23,26-31]. The ZNF91-mediated mini-heterochromatin domains prevent this cis influence in the NPC model system. Our data demonstrate that for most SVA loci only one of the heterochromatin marks is necessary to silence SVA expression and to prevent its regulatory influence on nearby gene expression. However, there are examples where the cooperation of the two mechanisms appears necessary. The most striking example is the HORMAD1 locus, where an upstream SVA can act as an alternative promoter [53]. The regulatory effect of this SVA is activated when H3K9me3 and DNA methylation are removed individually, demonstrating the need for both marks to prevent the cis influence of this SVA. Removal of both marks results in a massive activation of HORMAD1 expression, indicating that dual removal has synergistic effects. It is not yet understood why some SVA loci, such as the one upstream of HORMAD1, require both mechanisms for their control. However, it is likely that the transcriptional and epigenetic state of the integration sites is important, as well as structural variants within SVAs. For example, it is known that the VNTR region of SVAs is highly variable and has expanded recently in human evolution [61,85]. It is also evident that ZNF91-heterochromatin domains are not able to prevent the regulatory influence of all new SVA germline insertions. SVA insertions on the sense strand within genes are less abundant than expected by chance, suggesting that such SVA insertions are selected against [16,86]. In addition, there is a growing number of polymorphic SVA insertions linked to genetic disorders, with XDP being the best characterized example.

The SVA insertion linked to XDP is in intron 32 of the essential gene TAF1; the molecular phenotype includes intron retention and reduced TAF1 expression [39,40]. Our data demonstrate that ZNF91 binds to the XDP-SVA and establish a polymorphic mini-heterochromatin domain. The epigenetic status of the XDP-SVA is highly relevant for XDP pathology, as this layer of heterochromatin protects against the gene regulatory impact of the SVA. When DNA methylation and H3K9me3 are lost, intron retention is greatly increased, and TAF1 expression levels are further reduced. Thus, while ZNF91 can limit the impact of the XDP-SVA insertion, it cannot entirely remove the regulatory impact over the TAF1 gene. This explains why the XDP-SVA causes disease while most other SVA insertions are inert. However, we still do not understand

why ZNF91 is unable to fully block the cis-regulatory impact of the XDP-SVA.

Our data are limited to cell-culture models; the epigenetic status of the XDP-SVA in human brain tissue has not been investigated. It is worth noting that DNA methylation patterns in the human brain change with age [87-89]. It will be interesting to investigate if DNA methylation over the XDP-SVA is stable in the human brain, or if it is lost with ageing. Such phenomena could explain the late-onset phenotype of XDP, where a gradual increase in the loss of TAF1 function ultimately results in cellular dysfunction. Such a scenario would also open new therapeutic possibilities where restoration of DNA methylation on the XDP-SVA could block or reverse the pathology.

KZFPs have been implicated in an evolutionary arms race with TEs, where KZFP gene expansions and modifications limit the activity of newly-emerged transposon classes [50,90]. This event is followed by mutations in the TEs to avoid repression in an ongoing cycle. One such example is ZNF91, which appeared in the last common ancestor of humans and Old-World monkeys and underwent a series of structural changes about 8-12 million years ago that enabled it to bind to SVA elements [50,53]. However, the presence of the SVA-binding ZNF91 in hominoids has not prevented the expansion of SVAs in their genomes. On the contrary, SVAs are highly active in the human germline, providing a substantial source of genome variation in the population [5,10,21-25]. Although ZNF91 is not able to completely prevent new SVA germline insertions, sustained expression during brain development greatly limits the cis-regulatory impact of these insertions. Thus, it appears that in this case ZNF91 may facilitate the expansion of SVA insertions by limiting their gene-regulatory impact on the human genome. Thus, our data are consistent with a model where KZFPs are not only TE repressors, but also facilitators of inert germline transposition events, thereby fueling genome complexity and evolution.

Our results demonstrate how a KZFP prevents the regulatory impact of TEs in human neural cells, but it is still not known if the ZNF91-SVA partnership represents a unique event, or if these results can be extrapolated to other TE families and lineages. In humans, there are three active TE classes: Alus, LINE-1s, and SVAs. The relationship between LINE-1s and SVAs is of special interest, since SVA retrotransposition depends on co-expression of the LINE-1 machinery. We and others have previously found that LINE-1s are controlled by DNA methylation in human NPCs, and that their transcriptional activation is recognized and silenced by the HUSH complex by a mechanism that is independent of KZFPs and TRIM28 [59,91-93]. Thus, in human brain development LINE-1 activity appears to be controlled by fundamentally different mechanisms to SVAs. This suggests that the control of TE activity in the human brain is not only multilayered, but also highly specialized. Our data are limited to cell models of early development and neural cells, where ZNF91 is particularly highly expressed. We do not know how SVAs are controlled in other human tissues. In addition, our data indicate that there are additional KZFPs controlling SVAs in early human development (see one example in Figure S4). ZNF91 deletion in iPSCs activates only a fraction of SVAs, all others stayed silenced. It is likely that these SVAs contain binding sites for additional KZFPs, such as ZNF611, that co-operate to control SVAs in early development [29,53]. ZNF91 may then play a unique role in neural tissues. It is the only KZFP which binds SVAs and protects against cis-acting mechanisms from regulatory sequences in SVAs that are highly active in neural cells.

In summary, our results provide a unique mechanistic insight into an epigenetic defense system, based on a KZFP, active against the regulatory impact of SVA transposons in the human brain. On the one hand, this system protects the genome from any negative impact of SVAs, and SVA insertions result in genetic disease only in very rare instances, here exemplified by XDP. On the other hand, this system has likely contributed to the expansion of SVAs in our genomes, maximizing the potential for TEs to contribute to increased genome complexity and suggesting that SVAs are likely to have played an important role in primate brain evolution.

Data and code availability

Processed sequencing data has been deposited at GSE245093. Additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

This paper includes analyses of existing, publicly

available data. The accession numbers for these datasets are:

GSE224747: 3' single nuclei RNAseq, bulk RNAseq, and H3K9me3 CUT&RUN of human fetal forebrain tissue. GSE242143: H3K9me3 CUT&RUN from the DNMT1-KO NPCs

All original code has been deposited at GitHub and is publicly available at:
https://github.com/raquelgarza/XDP_Horvath_2023

Acknowledgements

We would like to thank Frank Jacobs, Didier Trono, Cristopher Bragg, and Amy Alessi for comments on the manuscript and their support throughout this project. We also thank Jenny Johansson, Marie Persson Vejgården, Anna Hammarberg, and Ulla Jarl for their technical assistance. We are grateful to all members of the Jakobsson lab. The work was supported by grants from the Collaborative Center for X-Linked Dystonia-Parkinsonism (J.J and C.H.D.), the Swedish Research Council (2018-02694 to J.J. and 2021-03494 to C.H.D.), the Swedish Brain Foundation (FO2019-0098 to J.J.), Cancerfonden (190326 to J.J.), Barncancerfonden (PR2017-0053 to J.J.), the Swedish Society for Medical Research (S19-0100 to C.H.D.), and the Swedish Government Initiative for Strategic Research Areas (MultiPark & StemTherapy).

Author Contributions

All of the authors took part in designing the study and interpreting the data. V.H. and J.J. conceived the study. V.H., M.J., P.J., A.A., G.C., O.K., L.C.V. and C.D. performed the experimental research. R.G. and N.P., performed the bioinformatic analyses. C.D., P.G. contributed expertise. V.H., C.D. and J.J. wrote the manuscript, and all authors reviewed the final version.

Competing interests

The authors declare no competing interests.

METHODS

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Prior to experimental use, all cell lines were confirmed to be mycoplasma free.

Induced pluripotent stem cells (iPSCs)

We used iPSC lines derived from three XDP patients and three healthy individuals from WiCell (See table 1). iPSCs were maintained on Biolaminin 521-coated (0.7 mg/cm^2; Biolamina) Nunc multidishes in iPS media (StemMACS iPS-Brew XF and 0.5% penicillin/streptomycin (GIBCO)). Cells were passaged 1:3 every 2-3 days. Briefly, cells were rinsed once with DPBS (GIBCO) and dissociated using Accutase (GIBCO) at 37°C for 5 minutes. Following incubation, Accutase was carefully aspirated from the well, and the cells were washed off from the dish using washing medium (9.5 ml DMEM/F12 (GIBCO) and 0.5 ml knockout serum replacement (GIBCO)). The cells were then centrifuged at 400 g for 5 minutes and resuspended in iPS brew medium supplemented with 10 mM Y27632 Rock inhibitor (Miltenyi Biotech) for expansion. The media was changed daily [95].

Neural progenitor cells (NPCs)

The neural progenitor cells (NPCs) were generated from iPSCs from the three XDP patients and three unaffected individuals (Table 1). The neural induction was done as described in [96]. The NPCs were cultured in DMEM/F12 (Thermo Fisher Scientific) supplemented with glutamine (2 mM, Sigma), penicillin/streptomycin (1×, Gibco), N2 supplement (1×, Thermo Fisher Scientific), B27 (0.05×, Thermo Fisher Scientific), EGF and FGF2 (both 10 ng/ml, Thermo Fisher Scientific). 10 mM Y27632 Rock inhibitor (Miltenyi) was also used. Cells were grown on Nunc multidishes or in T25 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 TryplLE™ express enzyme (GIBCO) and trypsin inhibitor (GIBCO).

Immunocytochemistry

24-well Nunc plates were pre-coated with Poly L-Ornithine (15 μg/ml, Sigma) and Laminin (2 μg/ml, Sigma). Approximately 50,000 cells were plated in the wells and were allowed to expand, until they reached 70-80% confluency. At this point, cells were washed three times with DPBS (GIBCO) and fixed with 4% paraformaldehyde (Merck Millipore) solution for 15 minutes at room temperature, and washed again three times with DPBS. Fixed cells were stored in DPBS at 4°C for a maximum of one month until staining and imaging.

For blocking, cells were incubated for one hour with 5% normal donkey serum (NDS) in TKBPS (KBPS with 0.25% Triton X-100 (Fisher Scientific)). Subsequently, they were incubated overnight at 4°C with the primary antibody (5mC, Active Motif, cat.no. 39649, lot 02617020, used 1:250; SOX2, R&D Systems, AF2018, 1:100; and Nestin, Abcam, AB176571, 1:100). For a negative control, cells were incubated overnight with TKPBS + 5% NDS. After overnight incubation, cells were washed two times for five minutes in TKPS, followed by five minutes in TKBPS with NDS. Next, they were incubated at room temperature for two hours with the secondary antibody (donkey anti-rabbit Alexa fluor 647, 1:200, Jackson Lab, and donkey anti-goat cy3, 1:200, Jackson Lab) and for five minutes with DAPI (1:1000, Sigma Aldrich) as a nuclear counterstain. This was followed by two five-minute washes with KPBS, then cells were stored in PBS until imaging.

5mC staining

As described in Jönsson et al. 2019 59, cells stained for 5mC were pre-treated with 0.9% Triton in PBS for 15 minutes, followed by 2 N HCl for 15 minutes, then 10 mM Tris-HCl, pH 8, for 10 minutes prior to incubation with the primary antibody. Cells were imaged using a fluorescence microscope (Leica).

RNA sequencing

Total RNA was isolated using the RNeasy Mini Kit (Quiagen) with on-column DNAse treatment following the manufacturer's instructions. The isolated RNA was used for qPCRs (see below) and RNA sequencing. RNA sequencing was performed using four biological replicates. Libraries for RNA sequencing were generated using Illumina TruSeq Stranded mRNA library prep kit (poly-A selection), optimized for long fragments, and sequenced on a Novaseq6000 (paired end, ~250 bp) yielding an average of 46M reads. The reads were mapped to the human reference genome (hg38) using STAR aligner v2.7.8a [97] and gene quantification was performed using FeatureCounts (Subread package v1.6.3; hg38 Gencode v38), setting -p for paired-end, and -s 2 for reversely stranded reads (TruSeq) [98].

To quantify TE expression, reads were re-mapped using STAR aligner and discarded if mapped to more than one location (–outFilterMultimapNmax 1). A maximum of 0.03 mismatches per base were allowed (–outFilterMismatchNoverLmax 0.03). FeatureCounts (Subread package v1.6.3) using hg38 RepeatMasker annotation "parsed to filter out low complexity and simple repeats, rRNA, scRNA, snRNA, srpRNA and tRNA" was used to quantify reads [99].

Bigwig files for genome browser tracks were generated using bamCoverage (deeptools v2.5.4), set to –normalizeUsingRPKM and –filterRNAstrand to split signal between strands. Visualization was performed in the Integrative Genome Browser (IGV) [100]. Matrices for deeptools heatmaps were generated including only SVAs longer than 1 kbp (grouped by subfamily; individual BED files), using computeMatrix scale-regions setting –regionBodyLength to 1kbp, and flanking regions (-a and -b) to 10kbp. Heatmaps were generated using plotHeatmap (v3.5.1). Profile plots were generated the same way, ungrouping the SVAs (input a single BED file) prior to the matrix computation.

Normalization of counts to visualize the expression of different features on barplots (genes, TAF1 intron 32 or exon 38) was performed as TPM: the length of the feature was used to calculate an approximate TPM value. Statistical tests, however, were performed using DESeq2, which normalizes using median of ratios [101]. Intron 32 and exon 38 of the TAF1 gene were added as part of the gene count matrix.

CUT&RUN

CUT&RUN analysis was done on CNPC1, CNPC2, XNPC1, and XNPC3 in both ZNF91-

CRISPRi and TRIM28-CRISPRi, including CRISPRi-Ctrl (lacZ). We followed the protocol described in Skene and Henikoff 2018 [102]. Briefly, 300,000 cells were washed twice (20 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM spermidine, 1× Roche cOmplete protease inhibitors) and attached to 10 ConA-coated magnetic beads (Bangs Laboratories) that had been pre-activated in binding buffer (20 mM HEPES pH 7.9, 10 mM KCl, 1 mM $CaCl_2$, 1 mM MnCl2). Bead-bound cells were resuspended in 50 ml buffer (20 mM HEPES pH 7.5, 0.15 M NaCl, 0.5 mM spermidine, 1× Roche complete protease inhibitors, 0.02% w/v digitonin, 2 mM EDTA) containing primary antibody (rabbit anti H3K9me3, Abcam ab8898, RRID:AB_306848, or goat anti-rabbit IgG, Abcam ab97047, RRID:AB_10681025) at 1:50 dilution and incubated at 4°C overnight with gentle shaking. Beads were washed thoroughly with digitonin buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM spermidine, 1× Roche complete protease inhibitors, 0.02% digitonin). After the final wash, pA-MNase (a generous gift from Steve Henikoff) was added in digitonin buffer and incubated with the cells at 4°C for 1 hour. Bead-bound cells were washed twice, resuspended in 100 ml digitonin buffer, and chilled to 0-2°C. Genome cleavage was stimulated by adding 2 mM CaCl2 at 0°C for 30 minutes. The reaction was quenched by adding 100 ml 2× stop buffer (0.35 M NaCl, 20 mM EDTA, 4 mM EGTA, 0.02% digitonin, 50 ng/ml glycogen, 50 ng/ml Rnase A, 10 fg/ml yeast spike-in DNA (a generous gift from Steve Henikoff) and vortexing. After 10 minutes incubation at 37°C to release genomic fragments, cells and beads were pelleted by centrifugation (16,000 g, 5 minutes, 4°C) and fragments from the supernatant were purified. Illumina sequencing libraries were prepared using the Hyperprep kit (KAPA) with unique dual-indexed adapters (KAPA), pooled, and sequenced on a Nextseq500 instrument (Illumina). Paired-end reads (2×75) were aligned to the human and yeast genomes (hg38 and R64-1-1 respectively) using bowtie2 (–local –very-sensitive- local –no-mixed –no-discordant –phred33 -I 10 -X 700) and converted to bam files using samtools [103]. Normalized bigwig coverage tracks were made using bamCoverage (deeptools) [104], with a scaling factor accounting for the number of reads arising from the spike-in yeast DNA (104 per aligned yeast read number). Tracks were displayed in IGV.

CRISPR approaches

CRISPRi

To silence the transcription of ZNF91 and TRIM28, we used the catalytically inactive Cas9 (deadCas9) fused to the transcriptional repressor KRAB 105. Single-guide sequences were designed to recognize DNA regions just down-stream of the transcription start site (TSS), according to the GPP Portal (Broad Institute). ZNF91 sgRNA: GAGTTTCCAGGTCTCGACTT (No PAM). The guides were inserted into a deadCas9-KRAB-T2A-GFP lentiviral backbone containing both the guide RNA under the U6 promoter and deadCas9-KRAB and GFP under the Ubiquitin C promoter (pLV hU6-sgRNA hUbC-dCas9-KRAB-T2a-GFP, a gift from Charles Gersbach, Addgene plasmid #71237 RRID:Addgene_71237). The guides were inserted into the backbone using annealed oligos and the BsmBI cloning site. Lentiviruses were produced as described below, yielding titers of 10^8-10^9 TU/ml, which was determined using qRT–PCR. Control virus with a gRNA sequence absent from the human genome (LacZ) was also produced and used in all experiments. All lentiviral vectors were used with an MOI of 2.5 unless stated differently. GFP cells were FACS isolated (FACSAria, BD sciences) on day 10 at 10°C (reanalysis showed >97% purity) and pelleted at 400 g for 5 minutes, snap frozen on dry ice and stored at −80°C until RNA isolation. All groups were performed in 4 biological replicates unless indicated differently. Knockdown efficiency was validated using RNA sequencing.

DNMT1 CRISPR cut

LV.gRNA.CAS9-GFP vectors were used to target DNMT1 79 or LacZ (control) as described in 59. Lentiviral vectors were produced as described previously and had a titer of 10^8-10^9 TU/ml which was determined using qRT-PCR. hNPCs were transduced with an MOI of 10-15, allowed to expand for 10 days, and were FACS-sorted as described previously.

DNMT1 and ZNF91 double CRISPRi

We used a double-transduction method to do a double CRISPRi of DNMT1 and ZNF91. We transduced ZNF91 with the previously-mentioned deadCas9-KRAB-T2A-GFP lentivirus containing the dead Cas9 protein and a GFP. At the same time, the cells were transduced with lentivirus containing the pLV.U6B-

smBI.EFS-NS.H2b-RFPW lentiviral backbone wihth the gRNA for DNMT1 and mCherry as a marker but without deadCas9 to knock down DNMT1. DNMT1 sgRNA: TGCTGAAGCCTCCGAGATGC (no PAM). Double-positive (mCherry and GFP) cells were FACS sorted as previously described and stored at −80°C until RNA extraction.

Lentiviral vector production

Lentiviral vectors were produced according to Zufferey et al. [106]. Briefly, HEK293T cells were grown to a confluency of 70-90% at the day of transfection for lentiviral production. We used third-generation packaging and envelop vectors (pMDL, psRev, and pMD2G), together with polyethyleneimine (PEI Polysciences PN 23966, in DPBS (Gibco)). The lentivirus was harvested 2 days after transfection. The supernatant was then collected, filtered, and centrifuged at 25,000 g for 1.5 hours at 4°C. The supernatant was removed from the tubes and the virus was resuspended in PBS and left at 4°C. The resulting lentivirus was aliquoted and stored at −80°C.

CUT&RUN qRT-PCR

To identify whether the XDP-SVA was surrounded by an H3K9me3 mark, we designed a qPCR approach. Briefly, two primers were designed on the 5' flanking region of the XDP-SVA: one in the flanking region and one in the SVA. As the positive control, primers were designed for a genomic region (hg38, chr5:141253464-141255143) known to be covered by H3K9me3. The CUT&RUN library was used as a template for amplification. The qRT-PCR was done with SYBR Green I Master (Roche) on a LightCycler 480 (Roche). The primer pairs used were: XDP-SVA forward (5'-3'): GAATGGTATATGTTTAGTTTTACA; XDP-SVA reverse (5'-3'): CATGACCCTGCCAAATCCCCCT; positive-control forward (5'-3'): AAATGGGAATTAAAATCAGTGAGGC; positive-control reverse (5'-3'): TTGACATATCATTAAGGGGGCA.

Oxford Nanopore sequencing

Whole-genome Nanopore sequencing

DNA was extracted from frozen pellets using the Nanobind HMW DNA Extraction kit (PacBio) following the manufacturer's instructions. Final product was eluted in 100 µl of elution buffer provided in the kit. DNA concentration and quality were measured using Nanodrop and Qubit from the top, middle, and bottom of each tube. Only DNA with a quality of 260/280 1.8-2.0 and 260/230 2.0-2.2 was further processed. Whole-genome sequencing on samples XNPC1 and CNPC1 was done at the SciLife lab in Uppsala using SQK-LSK109 Ligation Sequencing kit (Oxford Nanopore Technologies) and FLO-PRO002 PromethION Flow Cell R9 Version, and was sequenced on a PromethION (Oxford Nanopore Technologies).

Cas9-targeted Nanopore sequencing

To target the XDP locus we used the Cas9 sequencing (SQK-CS9109) kit following the manufacturer's instructions (Oxford Nanopore Technologies). To enrich for the fragment of interest, four previously-described guide RNAs were used [107]. Briefly, two guides were designed upstream and two were designed downstream of the XDP-SVA insertion. The excision using these guides resulted in a 5.5 kbp product, including the XDP-SVA (2.6 kbp). 5 µg of DNA was used. Samples were sequenced on a MinION Mk1Mc using flow cell R9.4.1 (Oxford Nanopore Technologies). One flow cell per sample was used. For cas9-targeted enrichment we obtained 20770 reads from 3 samples (XNPC3 CRISPRi-Ctrl 7805 reads; XNPC3 DNMT1-KO 7581 reads; XNPC3 ZNF91-CRISPRi 4384 reads; samtools view -c BAM). The proportion of reads that mapped to the target was on average 1.25% (CRISPRi-Ctrl 52 reads; DNMT1-KO 171 reads; ZNF91-CRISPRi 36 reads; primary alignments over TAF1 intron 32 only: samtools view -c -F 260 BAM chrX:71424238-71457170). XNPC3 CRISPRi Ctrl, ZNF91-CRISPRi, and DNMT1-KO samples were sequenced using the targeted approach.

Fastq files were indexed using the nanopolish index (v0.13.3) on default parameters [108]. To build an index of the XDP genome (with the XDP-SVA insertion), a consensus of the SVA sequences (two enrichment methods via PCR and CRISPR) as reported by Reyes et al 107 (https://github.com/nanopol/xdp_sva/blob/main/) was created using EMBOSS cons (v6.6.0.0)

(http://emboss.open-bio.org/) resulting in a 2,638 bp long sequence.

A TAF1 fasta file was generated using grep -w TAF1 from hg38 gencode v38, and bedtools getfasta (v2.30.0) [109]. The SVA consensus sequence and the TAF1 sequence were then aligned using clustalw2 (v2.1). We observed that the breaking point between the sequence extracted by Reyes et al. 2022 and the reference genome's sequence of TAF1 occurred at nucleotide chrX:71,440,502 [107].

Fasta file of chrX was then chopped using bedtools getfasta using breaking points:

chrX	1	71440502
chrX	71440503	156040895

The three sequences (chrX:1-71440502, the SVA sequence, and chrX:71440503-156040895) were then stitched (concatenated) together. A new genome fasta file was created concatenating all hg38 chromosomes fasta files (except for chrX) and the XDP-chrX fasta file. A minimap2 (v2.24) index was then created using the Nanopore present (-x map-ont) [110]. Mapping of the reads was performed using minimap2 (v2.24) using the Nanopore preset (-a -x map-ont) with the XDP genome index. BAM files were sorted and indexed using samtools (v1.16.1).

Polymorphic insertions were identified using TLDR (v1.2.2), using GRCh38.p13 as reference genome (-r) and a TE library (-e) including the consensus sequences for TE subfamilies: L1Ta, L1preTa, L1PA2, SVA A-F, and HERVK (sequences provided by TLDR developers) [61]. Insertions were considered for further analysis if they were found in the two individuals analyzed (CNPC1 and XNPC1). Insertions were required to have an UnmapCover of at least 80% (percentage of insertion with TEsequence), have a sequence similarity (TEMatch) of at least 80% to the TE consensus sequence, and a minimum of three reads supporting it (SpanReads).

The local consensus sequences of the polymorphic insertions (as output from TLDR) were introduced to the reference genome. A custom script (add_polymorphic_insertions_fa.py) was used to read the TLDR output table, sort the polymorphic insertions from the end to the start of each chromosome, and perform the following operations for each of the chromosomes:

1. Read its fasta file (chr.fa)
2. Extract its sequence before and after the insertion using bedtools getfasta (-fi chr.fa -bed coordinates.bed), where coordinates.bed included two coordinates: One spanning from the beginning of the chromosome to the start of the insertion (as reported by TLDR), and the second spanning from the end of the insertion (as reported by TLDR) to the end of the chromosome.
3. Concatenate the sequences:
    a. The chromosome before the insertion
    b. The sequence of the polymorphic insertion
    c. The chromosome after the insertion

And re-write the chromosome's fasta with it.

This process was repeated for each polymorphic insertion, introducing them in order from end to start of each of the chromosomes. Similarly, an updated gene annotation GTF file (to fit the coordinates including all polymorphic insertions) was created using a custom script following the same logic (add_polymorphic_insertions.py).

The reads were re-mapped to the custom genome using an indexed version of the output fasta (output from add_polymorphic_insertions_fa.py) using minimap2 (index using -x map-ont; mapping using -a -x map-ont).

Methylation for each of the regions of interest was called using nanopolish call-methylation (v0.13.3) with the raw reads (-r), the alignment files to the custom genome (-b), and the custom genome's fasta file as a reference (-g). Databases for methylartist were produced using methylartist db-nanopolish (v1.2.2), using the methylation calls files as input (default parameters). Specific loci were visualized using methylartist locus (v1.2.2) [60].

SUPPLEMENTARY FIGURES

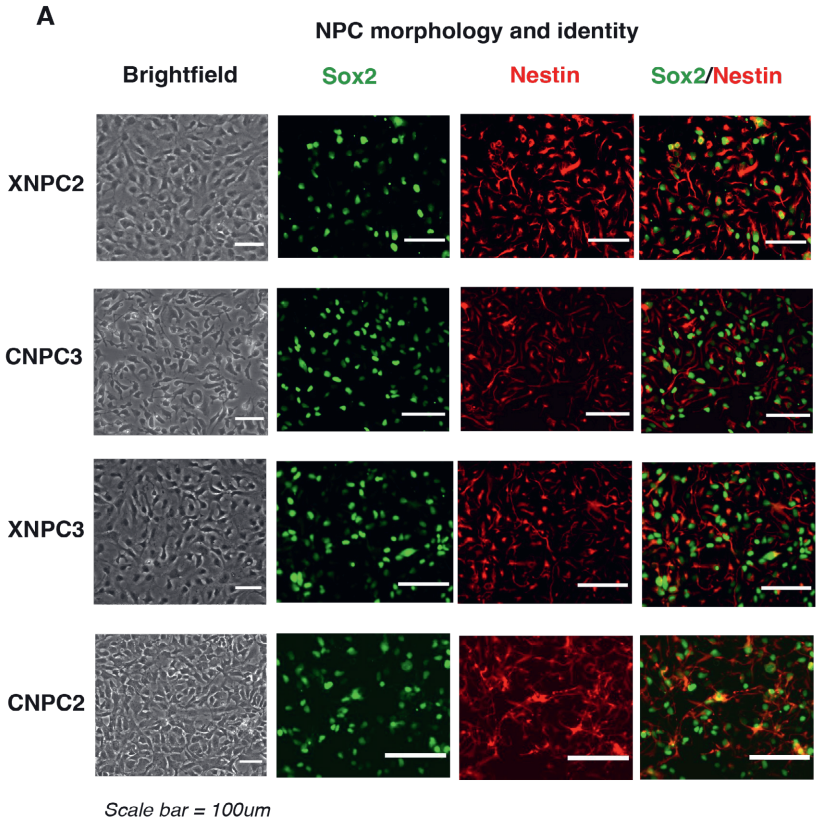
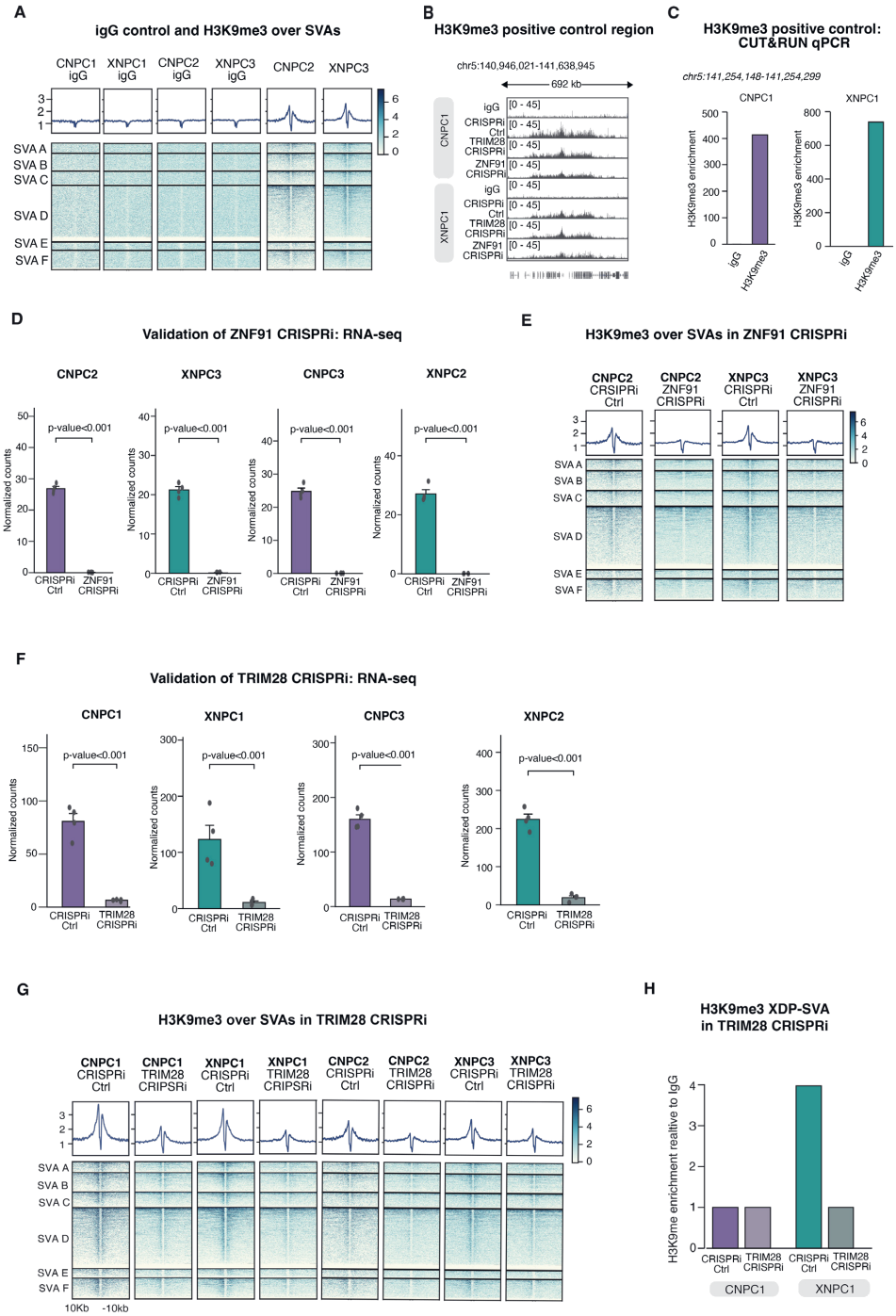


Figure S1. The XDP-NPCs display NPC morphology and expression of NPC markers.

(A) Brightfield images of XDP-NPCs and Ctrl-NPCs (left). Immunostainings (right) of Sox2 (green) and Nestin (red) in XDP-NPCs and Ctrl-NPCs.

Figure S2. ZNF91 and TRIM28 orchestrate H3K9me3 deposition over SVAs in NPCs.

(A) Heatmap showing igG and H3K9me3 enrichment in Ctrl-NPCs and XDP-NPCs. The genomic regions spanning ±10 kbp from the peak center are displayed. (B) Genome browser tracks showing H3K9me3 signal over a region known to be covered by H3K9me3 as a positive control. Tracks are shown for Ctrl-NPC and XDP-NPC for H3K9me3 and igG in control, TRIM28, and ZNF91-CRISPRi. (C) Barplots showing igG and H3K9me3 coverage over a positive-control region using CUT&RUN qPCR in Ctrl-NPC and XDP-NPC. (D) Barplots showing ZNF91 expression (RNA-seq) in CRISPRi-Ctrl and ZNF91-CRISPRi (padj, DESeq2). (E) Heatmaps showing H3K9me3 signal around SVAs in CRISPRi-Ctrl and ZNF91-CRISPRi (n=2). (F) Barplots showing TRIM28 expression (RNA-seq) in CRISPRi-Ctrl and TRIM28-CRISPRi in Ctrl-NPC and XDP-NPC (n=4) (padj, DESeq2). (G) Heatmaps showing H3K9me3 signal around SVAs in CRISPRi-Ctrl and TRIM28-CRISPRi (n=4). (H) Barplots showing the H3K9me3 status of the XDP-SVA in CRISPRi-Ctrl and TRIM28-CRISPRi (n=2) (padj, DESeq2).


A
igG control and H3K9me3 over SVAs
CNPC1 igG
XNPC1 igG
CNPC2 igG
XNPC3 igG
CNPC2
XNPC3
SVA A
SVA B
SVA C
SVA D
SVA E
SVA F
B
H3K9me3 positive control region
chr5:140,946,021-141,638,945
692 kb
CNPC1
XNPC1
igG
CRISPRi Ctrl
TRIM28 CRISPRi
ZNF91 CRISPRi
C
H3K9me3 positive control: CUT&RUN qPCR
chr5:141,254,148-141,254,299
H3K9me3 enrichment
D
Validation of ZNF91 CRISPRi: RNA-seq
CNPC2
XNPC3
CNPC3
XNPC2
p-value<0.001
Normalized counts
E
H3K9me3 over SVAs in ZNF91 CRISPRi
CNPC2 CRSIPRi Ctrl
CNPC2 ZNF91 CRISPRi
XNPC3 CRISPRi Ctrl
XNPC3 ZNF91 CRISPRi
F
Validation of TRIM28 CRISPRi: RNA-seq
TRIM28 CRISPRi
G
H3K9me3 over SVAs in TRIM28 CRISPRi
CNPC1 CRISPRi Ctrl
CNPC1 TRIM28 CRIPSRi
XNPC1 CRISPRi Ctrl
XNPC1 TRIM28 CRIPSRi
CNPC2 CRISPRi Ctrl
CNPC2 TRIM28 CRISPRi
XNPC3 CRISPRi Ctrl
XNPC3 TRIM28 CRISPRi
10Kb
-10kb
H
H3K9me3 XDP-SVA in TRIM28 CRISPRi
H3K9me enrichment realtive to IgG

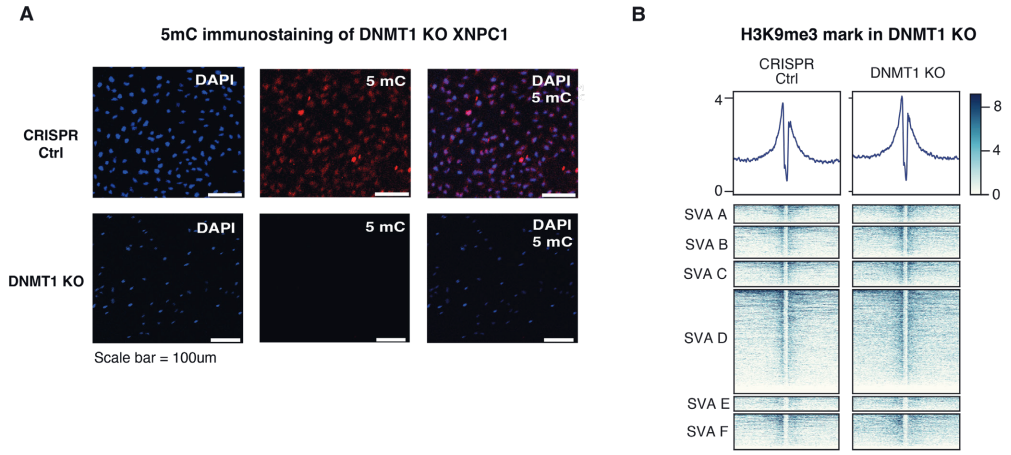


Figure S3. DNMT1 does not regulate H3K9me3

(A) Fluorescent 5mC immunostaining shows successful DNMT1-KO in XNPC1 10 days post transduction. Blue=Dapi, red=5mC. (B) Heatmap showing H3K9me3 around SVAs in a genome-wide scale in Ctrl and DNMT1-KO NPCs.

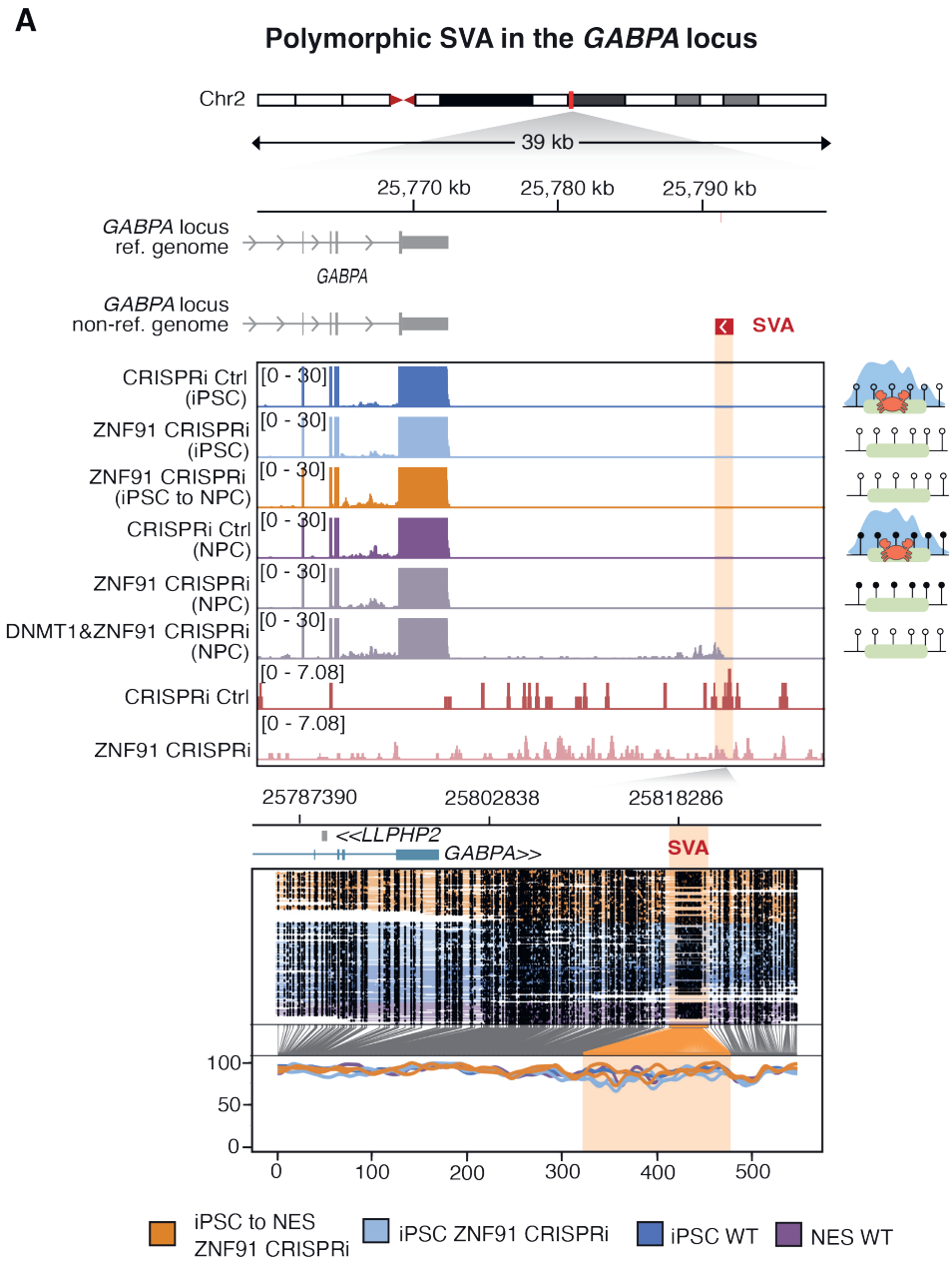


Figure S4. ZNF91 independent SVA regulation.

(A) Polymorphic SVA loci near GABPA. Genome browser tracks (top) showing gene expression and H3K9me3. ONT reads (bottom) showing SVA DNA methylation.

References

1. Lander, E.S. et al. Initial sequencing and analysis of the human genome. Nature 409, 860-921 (2001).
2. de Koning, A.P., Gu, W., Castoe, T.A., Batzer, M.A. & Pollock, D.D. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet 7, e1002384 (2011).
3. Nurk, S. et al. The complete sequence of a human genome. Science 376, 44-53 (2022).
4. Ostertag, E.M., Goodier, J.L., Zhang, Y. & Kazazian, H.H. SVA elements are nonautonomous retrotransposons that cause disease in humans. Am J Hum Genet 73, 1444-51 (2003).
5. Wang, H. et al. SVA elements: a hominid-specific retroposon family. J Mol Biol 354, 994-1007 (2005).
6. Kazazian, H.H. et al. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature 332, 164-6 (1988).
7. Batzer, M.A. et al. Amplification dynamics of human-specific (HS) Alu family members. Nucleic Acids Res 19, 3619-23 (1991).
8. Brouha, B. et al. Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci U S A 100, 5280-5 (2003).
9. Beck, C.R. et al. LINE-1 retrotransposition activity in human genomes. Cell 141, 1159-70 (2010).
10. Wang, L., Norris, E.T. & Jordan, I.K. Human Retrotransposon Insertion Polymorphisms Are Associated with Health and Disease via Gene Regulatory Phenotypes. Front Microbiol 8, 1418 (2017).
11. Payer, L.M. & Burns, K.H. Transposable elements in human genetic disease. Nat Rev Genet 20, 760-772 (2019).
12. Jönsson, M.E., Garza, R., Johansson, P.A. & Jakobsson, J. Transposable Elements: A Common Feature of Neurodevelopmental and Neurodegenerative Disorders. Trends Genet 36, 610-623 (2020).
13. Chen, J.M., Stenson, P.D., Cooper, D.N. & Férec, C. A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. Hum Genet 117, 411-27 (2005).
14. O'Donnell, K.A. & Burns, K.H. Mobilizing diversity: transposable element insertions in genetic variation and disease. Mobile DNA 1(2010).
15. Goodier, J.L. & Kazazian, H.H., Jr. Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell 135, 23-35 (2008).
16. Hancks, D.C. & Kazazian, H.H. SVA retrotransposons: Evolution and genetic instability. Semin Cancer Biol 20, 234-45 (2010).
17. Ono, M., Kawakami, M. & Takezawa, T. A novel human nonviral retroposon derived from an endogenous retrovirus. Nucleic Acids Res 15, 8725-37 (1987).
18. Zhu, Z.B., Hsieh, S.L., Bentley, D.R., Campbell, R.D. & Volanakis, J.E. A variable number of tandem repeats locus within the human complement C2 gene is associated with a retroposon derived from a human endogenous retrovirus. J Exp Med 175, 1783-7 (1992).
19. Shen, L. et al. Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. J Biol Chem 269, 8466-76 (1994).
20. Hoyt, S.J. et al. From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. Science 376, eabk3112 (2022).
21. Kim, H.S., Wadekar, R.V., Takenaka, O., Hyun, B.H. & Crow, T.J. Phylogenetic analysis of a retroposon family in african great apes. J Mol Evol 49, 699-702 (1999).
22. Feusier, J. et al. Pedigree-based estimation of human mobile element retrotransposition rates. Genome Res 29, 1567-1577 (2019).
23. Pfaff, A.L., Bubb, V.J., Quinn, J.P. & Koks, S. Reference SVA insertion polymorphisms are associated with Parkinson's Disease progression and differential gene expression. Npj Parkinsons disease 7(2021).
24. Wang, L., Rishishwar, L., Mariño-Ramírez, L. & Jordan, I.K. Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements. Nucleic Acids Res 45, 2318-2328 (2017).
25. van Bree, E.J. et al. A hidden layer of structural variation in transposable elements reveals potential genetic modifiers in human disease-risk loci. Genome Res 32, 656-670 (2022).
26. Hancks, D.C. & Kazazian, H.H. Active human retrotransposons: variation and disease. Curr Opin Genet Dev 22, 191-203 (2012).
27. Savage, A.L. et al. An evaluation of a SVA retrotransposon in the FUS promoter as a transcriptional regulator and its association to ALS. PLoS One 9, e90833 (2014).
28. Savage, A.L., Bubb, V.J., Breen, G. & Quinn, J.P. Characterisation of the potential function of SVA retrotransposons to modulate gene expression patterns. BMC Evol Biol 13, 101 (2013).

29. Pontis, J. et al. Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. Cell Stem Cell 24, 724-735.e5 (2019).
30. Trizzino, M., Kapusta, A. & Brown, C.D. Transposable elements generate regulatory novelty in a tissue-specific fashion. BMC Genomics 19, 468 (2018).
31. Trizzino, M. et al. Transposable elements are the primary source of novelty in primate gene regulation. Genome Res 27, 1623-1633 (2017).
32. Patoori, S., Barnada, S.M., Large, C., Murray, J.I. & Trizzino, M. Young transposable elements rewired gene regulatory networks in human and chimpanzee hippocampal intermediate progenitors. Development 149(2022).
33. Hancks, D.C. & Kazazian, H.H., Jr. Roles for retrotransposon insertions in human disease. Mob DNA 7, 9 (2016).
34. Nakamura, Y. et al. SVA retrotransposition in exon 6 of the coagulation factor IX gene causing severe hemophilia B. Int J Hematol 102, 134-9 (2015).
35. Vogt, J. et al. SVA retrotransposon insertion-associated deletion represents a novel mutational mechanism underlying large genomic copy number changes with non-recurrent breakpoints. Genome Biol 15, R80 (2014).
36. Pfaff, A.L., Singleton, L.M. & Kõks, S. Mechanisms of disease-associated SINE-VNTR-Alus. Exp Biol Med (Maywood) 247, 756-764 (2022).
37. Fröhlich, A. et al. CRISPR deletion of a SINE-VNTR-Alu (SVA_67) retrotransposon demonstrates its ability to differentially modulate gene expression at the MAPT locus. Front Neurol 14, 1273036 (2023).
38. Lee, L.V., Pascasio, F.M., Fuentes, F.D. & Viterbo, G.H. Torsion dystonia in Panay, Philippines. Adv Neurol 14, 137-51 (1976).
39. Aneichyk, T. et al. Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. Cell 172, 897-909.e21 (2018).
40. Bragg, D.C., Sharma, N. & Ozelius, L.J. X-Linked Dystonia-Parkinsonism: recent advances. Curr Opin Neurol 32, 604-609 (2019).
41. Makino, S. et al. Reduced neuron-specific expression of the TAF1 gene is associated with X-linked dystonia-parkinsonism. Am J Hum Genet 80, 393-406 (2007).
42. Lee, L.V. et al. The unique phenomenology of sex-linked dystonia parkinsonism (XDP, DYT3, "Lubag"). Int J Neurosci 121 Suppl 1, 3-11 (2011).
43. Falk, A. et al. Capture of neuroepithelial-like stem cells from pluripotent stem cells provides a versatile system for in vitro production of human neurons. PLoS One 7, e29597 (2012).
44. Friedli, M. & Trono, D. The developmental control of transposable elements and the evolution of higher species. Annu Rev Cell Dev Biol 31, 429-51 (2015).
45. Goodier, J.L. Restricting retrotransposons: a review. Mob DNA 7, 16 (2016).
46. Deniz, Ö., Frost, J.M. & Branco, M.R. Regulation of transposable elements by DNA modifications. Nat Rev Genet 20, 417-431 (2019).
47. Imbeault, M., Helleboid, P.Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. Nature 543, 550-554 (2017).
48. Thomas, J.H. & Schneider, S. Coevolution of retroelements and tandem zinc finger genes. Genome Res 21, 1800-12 (2011).
49. Rowe, H.M. & Trono, D. Dynamic control of endogenous retroviruses during development. Virology 411, 273-87 (2011).
50. Jacobs, F.M. et al. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature 516, 242-5 (2014).
51. Emerson, R.O. & Thomas, J.H. Adaptive evolution in zinc finger transcription factors. PLoS Genet 5, e1000325 (2009).
52. Wells, J.N. et al. Transposable elements drive the evolution of metazoan zinc finger genes. Genome Res 33, 1325-1339 (2023).
53. Haring, N.L. et al. deletion in human embryonic stem cells leads to ectopic activation of SVA retrotransposons and up-regulation of KRAB zinc finger gene clusters. Genome Res 31, 551-563 (2021).
54. Garza, R. et al. L1 retrotransposons drive human neuronal transcriptome complexity and functional diversification. bioRxiv, 2023.03.04.531072 (2023).
55. Rowe, H.M. et al. KAP1 controls endogenous retroviruses in embryonic stem cells. Nature 463, 237-40 (2010).
56. Yoder, J.A., Walsh, C.P. & Bestor, T.H. Cytosine methylation and the ecology of intragenomic parasites. Trends Genet 13, 335-40 (1997).
57. Macia, A. et al. Engineered LINE-1 retrotransposition in nondividing human neurons. Genome Res 27, 335-348 (2017).
58. Coufal, N.G. et al. L1 retrotransposition in human neural progenitor cells. Nature 460, 1127-31 (2009).
59. Jönsson, M.E. et al. Activation of neuronal genes via LINE-1 elements upon global DNA demethylation in human neural progenitors. Nat Commun 10, 3182 (2019).

60. Cheetham, S.W., Kindlova, M. & Ewing, A.D. Methylartist: tools for visualizing modified bases from nanopore sequence data. Bioinformatics 38, 3109-3112 (2022).

61. Ewing, A.D. et al. Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling. Mol Cell 80, 915-928.e5 (2020).

62. Feng, S., Jacobsen, S.E. & Reik, W. Epigenetic reprogramming in plant and animal development. Science 330, 622-7 (2010).

63. Reik, W., Dean, W. & Walter, J. Epigenetic reprogramming in mammalian development. Science 293, 1089-93 (2001).

64. Santos, F., Hendrich, B., Reik, W. & Dean, W. Dynamic reprogramming of DNA methylation in the early mouse embryo. Dev Biol 241, 172-82 (2002).

65. Fulka, H., Mrazek, M., Tepla, O. & Fulka, J., Jr. DNA methylation pattern in human zygotes and developing embryos. Reproduction 128, 703-8 (2004).

66. Smith, Z.D. et al. DNA methylation dynamics of the human preimplantation embryo. Nature 511, 611-5 (2014).

67. Walter, M., Teissandier, A., Pérez-Palacios, R. & Bourc'his, D. An epigenetic switch ensures transposon repression upon dynamic loss of DNA methylation in embryonic stem cells. Elife 5(2016).

68. Matsui, T. et al. Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. Nature 464, 927-31 (2010).

69. Rowe, H.M. et al. TRIM28 repression of retrotransposon-based enhancers is necessary to preserve transcriptional dynamics in embryonic stem cells. Genome Res 23, 452-61 (2013).

70. Walsh, C.P., Chaillet, J.R. & Bestor, T.H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nat Genet 20, 116-7 (1998).

71. Wiznerowicz, M. et al. The Kruppel-associated box repressor domain can trigger de novo promoter methylation during mouse early embryogenesis. J Biol Chem 282, 34535-41 (2007).

72. Jönsson, M.E. et al. Activation of endogenous retroviruses during brain development causes an inflammatory response. EMBO J 40, e106423 (2021).

73. Rowe, H.M. et al. De novo DNA methylation of endogenous retroviruses is shaped by KRAB-ZFPs/KAP1 and ESET. Development 140, 519-29 (2013).

74. Turelli, P. et al. Primate-restricted KRAB zinc finger proteins and target retrotransposons control gene expression in human neurons. Sci Adv 6, eaba3200 (2020).

75. Quenneville, S. et al. The KRAB-ZFP/KAP1 system contributes to the early embryonic establishment of site-specific DNA methylation patterns maintained during development. Cell Rep 2, 766-73 (2012).

76. Agostinho de Sousa, J. et al. Epigenetic dynamics during capacitation of naïve human pluripotent stem cells. Sci Adv 9, eadg1936 (2023).

77. Nichols, J. & Smith, A. Pluripotency in the embryo and in culture. Cold Spring Harb Perspect Biol 4, a008128 (2012).

78. Li, E., Bestor, T.H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell 69, 915-26 (1992).

79. Liao, J. et al. Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. Nat Genet 47, 469-78 (2015).

80. Lanciano, S. & Cristofari, G. Measuring and interpreting transposable element expression. Nat Rev Genet 21, 721-736 (2020).

81. Shen, M.R. et al. The KCl cotransporter isoform KCC3 can play an important role in cell growth regulation. Proc Natl Acad Sci U S A 98, 14714-9 (2001).

82. Hiki, K. et al. Cloning, characterization, and chromosomal location of a novel human K+-Cl- cotransporter. J Biol Chem 274, 10661-7 (1999).

83. Rosmarin, A.G., Resendes, K.K., Yang, Z., McMillan, J.N. & Fleming, S.L. GA-binding protein transcription factor: a review of GABP as an integrator of intracellular signaling and protein-protein interactions. Blood Cells Mol Dis 32, 143-54 (2004).

84. Sharrocks, A.D. The ETS-domain transcription factor family. Nat Rev Mol Cell Biol 2, 827-37 (2001).

85. Sulovari, A. et al. Human-specific tandem repeat expansion and differential gene expression during primate evolution. Proc Natl Acad Sci U S A 116, 23243-23253 (2019).

86. Hancks, D.C., Ewing, A.D., Chen, J.E., Tokunaga, K. & Kazazian, H.H., Jr. Exon-trapping mediated by the human retrotransposon SVA. Genome Res 19, 1983-91 (2009).

87. Prasad, R. & Jho, E.H. A concise review of human brain methylome during aging and neurodegenerative diseases. BMB Rep 52, 577-588 (2019).

88. Hernandez, D.G. et al. Distinct DNA methylation changes highly correlated with chronological age in the human brain. Hum Mol Genet 20, 1164-72 (2011).

89. Numata, S. et al. DNA methylation signatures in development and aging of the human prefrontal cortex. Am J Hum Genet 90, 260-72 (2012).

90. Bruno, M., Mahgoub, M. & Macfarlan, T.S. The Arms Race Between KRAB-Zinc Finger Proteins and Endogenous Retroelements and Its Impact on Mammals. Annu Rev Genet 53, 393-416 (2019).

91. Sanchez-Luque, F.J. et al. LINE-1 Evasion of Epigenetic Repression in Humans. Mol Cell 75, 590-604. e12 (2019).

92. Pandiloski, N. et al. DNA methylation governs the sensitivity of repeats to restriction by the HUSH-MORC2 corepressor. bioRxiv, 2023.06.21.545516 (2023).

93. Muotri, A.R. et al. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 435, 903-10 (2005).

94. Ito, N. et al. Decreased N-TAF1 expression in X-linked dystonia-parkinsonism patient-specific neural stem cells. Dis Model Mech 9, 451-62 (2016).

95. Grassi, D.A., Jönsson, M.E., Brattås, P.L. & Jakobsson, J. TRIM28 and the control of transposable elements in the brain. Brain Res 1705, 43-47 (2019).

96. Calvo-Garrido, J. et al. Protocol for the derivation, culturing, and differentiation of human iPS-cell-derived neuroepithelial stem cells to study neural differentiation in vitro. STAR Protoc 2, 100528 (2021).

97. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21 (2013).

98. Liao, Y., Smyth, G.K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923-30 (2014).

99. Jin, Y., Tam, O.H., Paniagua, E. & Hammell, M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics 31, 3593-9 (2015).

100. Robinson, J.T. et al. Integrative genomics viewer. in Nat Biotechnol, Vol. 29 24-6 (United States, 2011).

101. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15, 550 (2014).

102. Skene, P.J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. Elife 6(2017).

103. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357-9 (2012).

104. Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res 42, W187-91 (2014).

105. Johansson, P.A. et al. A cis-acting structural variation at the ZNF558 locus controls a gene regulatory network in human brain development. Cell Stem Cell 29, 52-69.e8 (2022).

106. Zufferey, R., Nagy, D., Mandel, R.J., Naldini, L. & Trono, D. Multiply attenuated lentiviral vector achieves efficient gene delivery in vivo. Nat Biotechnol 15, 871-5 (1997).

107. Reyes, C.J., Asano, K., Todd, P.K., Klein, C. & Rakovic, A. Repeat-Associated Non-AUG Translation of AGAGGG Repeats that Cause X-Linked Dystonia-Parkinsonism. Mov Disord 37, 2284-2289 (2022).

108. Loman, N.J., Quick, J. & Simpson, J.T. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods 12, 733-5 (2015).

109. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-2 (2010).

110. Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094-3100 (2018).