Genome instability and somatic retropositional mosaicism in the adult human brain

Dissertation
zur Erlangung des Grades
Doktor der Naturwissenschaften

am Fachbereich Biologie
der Johannes Gutenberg-Universität Mainz

Jonas Möhner
geboren am 14.03.1996 in Worms, Deutschland

Mainz, 2023

Dekan:

Erster Berichterstatter:

Zweiter Berichterstatter:

Tag der mündlichen Prüfung: 07.11.2023

Table of contents

Zusammenfassung 1

Abstract 2

List of abbreviations 3

1. Introduction 4

1.1 Somatic mosaicism 4

1.2 Brain development and mosaicism 4

1.3 Retrotransposons in human genomes 6

1.4 *De novo* transposon insertions as markers for brain mosaics 7

1.5 DNA double-strand break induction and repair 9

1.6 DSBs as markers for mosaicism and genomic instability 11

1.7 Transposon- or DSB-associated diseases 14

1.8 Brain areas of interest and their function 15

1.9 Aim and objectives of the thesis 17

2. Material and methods 18

2.1 Material 18

2.2 Methods: RDA 19
2.2.1 Sample preparation L1-RDA 20
2.2.2 Methodical approach L1-RDA 20
2.2.3 RDA oligos 20
2.2.4 Sequencing 21
2.2.5 Bioinformatic scanning of *de novo* LINE-1 flanks 21
2.2.6 Chromosomal distribution of *de novo* L1 insertions 22
2.2.7 Genomic feature analysis 22

2.3 Methods: BLISS 23
2.3.1 Sample preparation 23
2.3.2 DNA isolation 23
2.3.3 Blunting 24
2.3.4 Ligation of DNA double-strand breaks 24
2.3.5 Gel extraction 24
2.3.6 Covaris fragmentation 24
2.3.7 In vitro transcription of RNA 25
2.3.8 Poly(A) tailing 25
2.3.9 Synthesis of cDNA 25
2.3.10 BLISS-PCR 25
2.3.11 Agarose gel electrophoresis 26
2.3.12 Sequencing 26
2.3.13 Bioinformatics 26

3. Representational difference analysis: retrotransposon mosaic in the human brain 33

3.1 RDA coupled deep sequencing detects somatic SVA-retrotranspositions and mosaicism in the human brain 34
3.1.1 Abstract 34
3.1.2 Introduction 35
3.1.3 Material and Methods 38
3.1.4 Results 42
3.1.5 Discussion 51
3.1.6 Data availability statement 54
3.1.7 Ethics statement 54
3.1.8 Author contributions 54
3.1.9 Funding 55
3.1.10 Acknowledgments 55
3.1.11 Conflict of interest 55
3.1.12 Publisher's note 55
3.1.13 Supplementary material 55
3.2 Subchapter: LINE-1 targeted RDA in the human brain 56
3.2.1 Results 56
3.2.2 Discussion 62
3.3 Conclusion of NGS-coupled SVA- and L1-RDA 65

4. Double-strand break labeling: breakome & brain mosaic 67
4.1 Results 67
4.1.1 DSB statistics and hotspot calling 67
4.1.2 Chromosomal distribution of DSB hotspots 70
4.1.3 Feature analysis of DSB hotspots 70
4.1.4 DSB enriched genes and Recurrent DSB Cluster (RDC) 74
4.1.5 Shared DSB hotspots 78
4.1.6 Motif search in DSB hotspots 79
4.1.7 DSB hotspots and epigenetic marks in prefrontal cortex 81
4.1.8 Association of DSBs and *de novo* SVA insertions 84
4.2 Discussion 86
4.2.1 DSB detection and evaluation of applied methods 86
4.2.2 DSB hotspot identification 87
4.2.3 DSBs and transcription 88
4.2.4 DSBs and retrotransposons 89
4.2.5 Recurrent breaking clusters and associated genes 91
4.2.6 Predictors of DSB hotspots 93
4.2.7 Conclusion of the DSB analysis 95

5. Conclusion of the present thesis 96

References 98

Supplement A 114

Author contributions 123

Ethics statement **124**

Danksagung **125**

Eidesstattliche Versicherung **126**

Lebenslauf **127**

Zusammenfassung

Zellen des Somas, insbesondere des Gehirns, bilden regionenspezifisch unterschiedlich häufig genomische Variationen aus, was zu einem somatischen Mosaizismus führt. Dieses postzygotische Phänomen ist u. a. Folge von DNA-Schädigungen oder fehlerhafter Reparatur und kann zu neurogenetischen Störungen beitragen. Die vorliegende Arbeit präsentiert zwei innovative Ansätze, um die Rolle von Retrotransposons und DNA Doppelstrang-Brüchen (DSBs) bei der Entstehung des somatischen Mosaiks im humanen Gehirn zu untersuchen. Retrotransposons, darunter SVA und LINE-1 (L1), sind mobile genetische Elemente, die sich im Genom mittels des "*Copy-and-Paste*"-Mechanismus vermehren. Aktuelle NGS-basierte Studien haben gezeigt, dass die Retrotranspositionsmaschinerie im humanen Gehirn aktiv ist. Dies wirft die Frage auf, ob SVA und L1 bzw. deren Anwesenheit an orthologen Loci verwendet werden können, um somatische Unterschiede in Gehirnregionen nachzuvollziehen. Hierzu wird eine subtraktive kinetische Anreicherungstechnik namens *Representational Difference Analysis* (RDA) in Verbindung mit NGS etabliert. Zusätzlich werden chromosomale DSB-Hotspots und deren regionale Unterschiede im Gehirn untersucht. Für eine Form der Reparatur dieses DNA-Schadens ist bekannt, dass SINE/LINE-Information im Rahmen eines nicht-homologen end-joinings eingesetzt wird, d. h. es entstehen typische Signaturen von SINE/LINE-Integrationen an DSB-Stellen. Um das "*Breakome*" zu beschreiben, wird ein DSB-Markierungssystem auf der Grundlage von *Breaks Labeling In Situ and Sequencing* (BLISS) eingesetzt. Die RDA liefert Beweise für somatischen Mosaizismus, der durch unterschiedliche Retrotransposition von L1 und SVAs im humanen Gehirn hervorgerufen wird. Dabei können SVAs als „*Presence/Absence*“ Marker die Entwicklung von Telencephalon und Metencephalon widerspiegeln. *De novo* SVA und L1 Insertionen besitzen chromosomenweite Raten und eine bevorzugte Integration in GC- und TE-reiche Regionen und Genen, die tendenziell an neuraler Funktion beteiligt sind. Die "*Breakome*"-Ergebnisse zeigen DSB-Hotspots, welche im gesamten Gehirn oder hirnregionsspezifisch auftreten. Infolgedessen sind mehrere bekannte und neue "*recurrent DSB cluster*" (RDC) assoziierte Gene nachweisbar, die mit neurologischen Krankheiten in Verbindung gebracht werden können. Ergänzend lassen sich (epi-) genetische Prädiktoren für die Ausbildung von DSBs identifizieren, darunter DNA-bindende Proteine, die eine Rolle in der DSB-Reparatur spielen. Interessanterweise treten Retrotransposons und DSBs oft in unmittelbarer Nähe zueinander auf, was auf eine mögliche Beteiligung von mobiler DNA an der Induktion oder Reparatur von DSBs hindeutet. Zusammengefasst bieten die in dieser Arbeit vorgestellten Methoden vielfältige Anwendungsmöglichkeiten, wie z. B. für „*cell lineage tracing*“-Experimente oder die Analyse von möglicherweise pathogenen DNA-Schädigungen in Zusammenhang mit neurologischen oder tumorösen Erkrankungen.

Abstract

Cells of the soma, especially of the brain, generate genomic variations with region-specific differences in frequency, which leads to somatic mosaicism. This postzygotic phenomenon is, among others, a consequence of DNA damage or defective repair and may contribute to neurogenetic disorders. The present work provides two innovative approaches to investigate the role of retrotransposons and DNA double-strand breaks (DSBs) in the formation of somatic mosaicism in the human brain. Retrotransposons, including SVA and LINE-1, are mobile genetic elements that replicate in the genome by the "copy-and-paste" mechanism. Recent NGS-based studies demonstrated that the retrotransposon machinery is active in the human brain. This raises the question of whether SVA and LINE-1, respectively their presence at orthologous loci, can be used to track somatic differences in brain regions. For this purpose, a subtractive kinetic enrichment technique called Representational Difference Analysis (RDA) coupled with NGS is established. In addition, chromosomal DSB hotspots and their regional differences in the brain will be investigated. For one type of DSB repair, SINE/LINE information is known to be used in the context of non-homologous end-joining, i.e. typical signatures of SINE/LINE integrations at DSB sites are generated. To describe the 'breakome', a DSB labeling system based on Breaks Labeling In Situ and Sequencing (BLISS) is implemented. The RDA provides evidence for somatic mosaicism caused by differential retrotransposition of LINE-1 and SVAs in the human brain. In this context, SVAs as 'presence/absence' markers can reflect the development of telencephalon and metencephalon. *De novo* SVA and LINE-1 insertions have chromosome-wide rates and preferential integration in GC- and TE-rich regions and genes that tend to be involved in neural functions. The 'breakome' results show DSB hotspots occurring across the brain or in a brain region-specific manner. As a result, several known and novel recurrent DSB cluster (RDC) associated genes are detectable and can be linked to neurological diseases. Moreover, (epi-) genetic predictors of DSB formation can be identified, including DNA-binding proteins that play a role in DSB repair. Interestingly, retrotransposons and DSBs frequently occur in close proximity to each other, suggesting a possible involvement of mobile DNA in the induction or repair of DSBs. In summary, the methods presented in this work can be applied in various research areas, such as cell lineage tracing experiments or the analysis of potentially pathogenic DNA damage in the context of neurological or tumor diseases.

List of abbreviations

SNV	single nucleotide variants
GW	gestational week
CNV	copy-number variant
LINE-1	long interspersed nuclear element-1
SVA	SINE-VNTR-Alu
TE	transposable element
LTR	long terminal repeat
HERVs	human endogenous retroviruses
RNAPII	RNA polymerase II
ORF2	open reading frame two
VNTR	variable number of tandem repeats
RDA	representational difference analysis
DSB	DNA double-strand beak
SSB	single-strand DNA break
ROS	reactive oxygen species
NHEJ	non-homologous end joining
HR	homologous recombination
BLISS	breaks labeling in situ and sequencing
UMI	unique molecular identifier
RMDR	RT-product-mediated DSB repair
GDP	Genome Decoration Page
TOP	Topoisomerase
TPRT	target site-primed reverse transcription
DMSO	dimethyl sulfoxide

1. Introduction

1.1 Somatic mosaicism

Parental *de novo* mutations in gametes are described as a prezygotic event, where neutral or advantageous mutations of parental germ cells establish and transmit to offspring and are thus represented as a genotype in all descendant cells if no true reversion follows. In contrast, a so called somatic mosaicism occurs when at least two cells or populations of one individual present a unique mutational landscape due to *de novo* mutational events at postzygotic stages (Acuna-Hidalgo et al., 2015; Wright et al., 2019).

The genomes of somatic cells can vary as a result of single nucleotide variants (SNVs), small insertions/deletions, structural variants but also insertions of so called mobile transposable elements (Bodea et al., 2018; Campbell et al., 2015; Y. Wang et al., 2021). Bizzotto et al. (2021) were able to trace somatic mutational patterns to progenitor cells and unveiled their clonal distribution and thus their contribution to somatic mosaicism in distinct germ layers and organs. Consequently, genetic differences potentially converting to distinctive phenotypic patterns can affect multiple cell populations or tissues, depending on the stage at which these somatic mutational events are introduced during development. One major contributor to somatic mosaicism is the developmental segmentation process, namely gastrulation, in which distinctive germ layers such as meso-, endo- and ectoderm are formed, limiting the clonal expansion of mutations at these definitive stages to a single layer and logically later to specific organs. Mutational patterns can be considered at various levels and are not limited to the dissimilarity of organs, thus they can be detected even among genetically unique cells within a single organ. Concordantly, whole genome and targeted sequencing revealed that both healthy and pathological human tissues are indeed a mosaic, showing different somatic patterns that were observed in skin, lung, liver, bladder, cardiovascular system and brain (reviewed in Ogawa et al., 2022).

1.2 Brain development and mosaicism

The brain is of particular interest when it comes to somatic mosaicism because it is an organ that starts early in prenatal development and is among the last to complete postnatal development, thus providing a long period for genetic alterations.

As reviewed by Stiles & Jernigan (2010) neuroectodermal progenitor cells first emerge during gastrulation and later contribute to the neural plate. The neural plate of the embryo forms two ridges and gradually folds to generate the hollow neural tube at the third week of gestation. The rostral part of the neural tube gives rise to the brain, the caudal part forms the spinal column and the hollow cavity forms the ventricular system. The primary brain vesicles emerge in the rostral neural tube region at embryonic day 28 and can be divided into the prosencephalon, which further develops into the telencephalon and diencephalon, the mesencephalon and posterior the rhombencephalon, which further progresses to the metencephalon and myelencephalon. During the embryonic period at gestational week (GW) 8, primitive neural patterning of the brain with specification and organization occurs. Subsequently, at GW9 until the completion of gestation, fetal development gives rise to neuron production, migration and differentiation and forms mature patterns with sulci and gyri.

With the introduction of neurogenesis, other cell types like astrocytes and oligodendrocyte precursor cells, which originate from RG cells as part of gliogenesis, are produced. This process persists in the adult brain, with cycling and migrating populations of these cells (Jakovcevski et al., 2009). Adult brain-resident macrophages, called microglia, originate from the embryonic yolk sac and generate a long-lived population with self-renewal capacity during brain development (Alliot et al., 1999; Ginhoux et al., 2010). According to current research, adult mammalian brain neurogenesis is scaled down and neuronal precursor generation is limited to the dentate gyrus of the hippocampus and the subventricular zone of lateral ventricles, where migration to the olfactory bulb has been observed in rodent models (Lazarini et al., 2014; Merkle et al., 2014; Ramirez-Amaya et al., 2006; Spalding et al., 2013).

Based on the brain development described above, it becomes clear that mutations can occur at many developmental stages and also postnatally in the adult brain, thus making a different contribution to somatic mutational patterns depending on the time given to clonally expand (Figure 1). The occurrence of cortical mosaicism is largely shaped by the presence of SNVs, which were analyzed by single-cell sequencing of cerebral cortex neurons and were demonstrated to be traceable relative to development as a result of mutational SNV patterns (Lodato et al., 2015). Additionally, single cell genomic sequencing strategies revealed mosaic copy-number variants (CNVs) in neurons from normal human brain tissue, e.g. one report observed that 13 – 41% of human frontal cortex neurons contain a *de novo* variant (Cai et al., 2014; McConnell et al., 2013). Recently, transposable elements (TE), which utilize the so called ‘copy-and-paste’ mechanism, are recognized as another source of somatic mosaicism because one study detected somatic insertions of long interspersed nuclear element-1 (LINE-1 or L1), SINE-VNTR-Alu (SVA) and Alu with retrotransposon capture methods in hippocampus and caudate nucleus. Another study utilized whole-genome sequencing of single neurons to highlight LINE-1 somatic retrotransposition in brains (Baillie et al., 2011; Evrony et al., 2015).

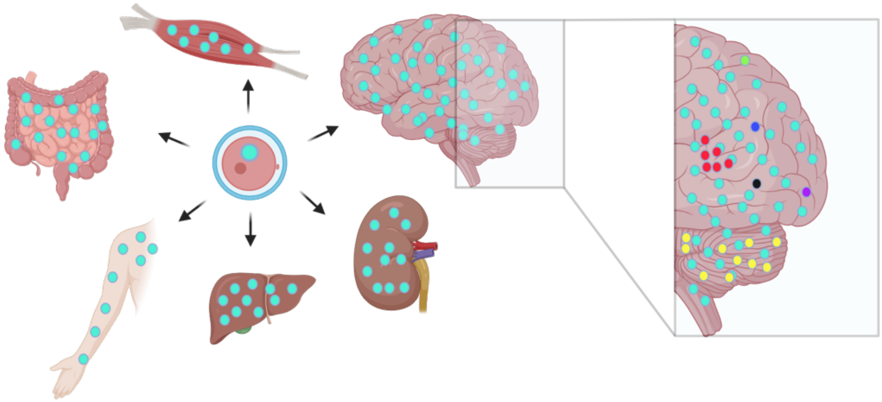


Figure 1: Illustration of the mutational landscape; dots with the same color represent a specific mutation and the respective genotype. A prezygotic mutation is transferred into progeny and all resulting tissues contain the respective mutation when no true reversal occurs (left). Upon closer inspection, a particular tissue, for example the brain (right), does not consist of a single genotype but represents a mosaic. Different mutations with varying degrees of clonal expansion result in a genotype being represented, e.g. in either one postmitotic cell, a cell population or a specific region. Created with BioRender.com.

1.3 Retrotransposons in human genomes

SVA, LINE1 and Alu are TEs that can move within the genome and are therefore also called jumping genes. Jumping genetic elements, which are incorporated at new chromosomal positions, were first discovered and described in the 1950s by McClintock (1956), and with increasing research up to the present day, they are accepted to comprise approximately half of the human genome (Lander et al., 2001). In humans, they can be classified into currently inactive DNA transposons, which made use of the ‘cut-and-paste’ mechanism in the primate lineage and anthropoid ancestors (Pace & Feschotte, 2007), and retrotransposons that duplicate by ‘copy-and-paste’ machinery. Retrotransposons can be further subdivided based on the presence of long terminal repeats (LTRs). Among these LTR-containing elements are the human endogenous retroviruses (HERVs), constituting approximately 8% of the human genome (Lander et al., 2001). SVA, LINE-1 and Alu are lacking LTR elements, thus are classified as non-LTR retrotransposons, accounting for one-third of the human genome (> 500,000 LINE-1, > 1,000,000 Alu and ~ 3,000 SVA copies) (Lander et al., 2001).

As reviewed by Cordaux & Batzer (2009) a full length LINE-1 is approximately 6 kb long and consists of an internal RNA polymerase II (RNAPII) promoter incorporated in the 5´-UTR, two open reading frames and a 3`-UTR with polyadenylation signal. The open reading frame two (ORF2) encodes a protein with reverse transcriptase and endonuclease activity for the ‘copy-and-paste’ machinery, thus acts in cis to accomplish LINE-1 reintegration as an autonomous retrotransposon in new genomic regions. In trans, Alus and SVAs, which lack ORF2, can hijack the LINE-1 machinery as non-autonomous retrotransposons and also integrate at new genomic sites. Alu elements are approximately 300 bp long, consist of components from the 7SL RNA gene and contain a characteristic left and right monomer, which are separated by an A-rich region. SVAs are also non-autonomous retrotransposons that can be divided into six hominoid-specific subfamilies named SVA_A through SVA_F, with SVA_E and SVA_F being exclusive to humans (H. Wang et al., 2005). A full length SVA element consists of a $(CCCTCT)_n$ hexamer repeat, Alu-like region, variable number of tandem repeats (VNTR), SINE-R region and a poly(A) tail (H. Wang et al., 2005). Owing to the variable number of tandem repeats and hexamer repeats, SVA insertions can range from 700 to 4000 bp (Hancks et al., 2009).

1.4 *De novo* transposon insertions as markers for brain mosaics

SVAs have been shown to integrate *de novo* - even as full-length element - in the human genome in the presence of LINE-1 activity, as validated by in vitro cell culture studies (Hancks et al., 2011). Evrony (2015) and Baillie (2011) reported the integration of LINE-1 and SVAs in the human brain, with retrotransposon capture methods specifying those as *de novo* insertions in the soma. Since these transposable elements can contribute to brain mosaicism, LINE-1 and SVAs are of particular interest in the present study and are validated as retropositional fingerprints in distinct areas of the adult human brain from two male donors.

To that end, a kinetic enrichment technique was implemented to trace each somatic retrotransposon insertion to a unique event in a common ancestor of a cell population and thereby compare complex somatic genomes. In contrast to recent methods, the RDA focuses on rare genomic changes and uses retrotransposons as informative clade markers at orthologous loci to describe their occurrence across multiple samples. This method was proposed by Lisitsyn et al. (1993) as representational difference analysis (RDA) and is modified in the present work to specifically amplify the 5′-flanking region of unique transposon insertions in brain samples (tester) compared to ectodermal skin (driver). Therefore, MboI-restricted DNA is ligated with a ‘GATC’-ligatable adaptor, complementary to MboI mediated sticky ends, in driver and tester samples. The 5´ flanking regions of transposon are specifically enriched by PCR, utilizing adapter related primers and outward primer systems complementary to a consensus sequence of the transposon class of interest. The PCR is specifically designed to amplify regions spanning from the ligated restriction sites located 5´-upstream of a transposon to the internal transposon region. The RDA-implemented SVA mosaic approach uses a consensus region of

SVA_A to SVA_F as outward primer system as proposed in the mobile element scanning method for SVAs (Ha et al., 2016). In RDA experiments targeting LINE-1 mosaicism, LINE-1 specific primers were generated based on full-length, 5´ untruncated L1-ORF0 consensus sequences (hg38). ORF0, which is located in the 5´-UTR of LINE-1, was recently characterized by Denli et al. (2015) and may influence LINE-1 mobility through expression of the ORF0 encoding protein. After specific enrichment of retrotransposon-flanked regions by PCR, exclusively the tester PCR products are ligated to an RDA adapter and are given a 100-fold molar excess of driver PCR products for hybridization (Figure 2). During hybridization 3 possible double-stranded fragments can occur: 1) both single strands are of driver origin and therefore do not contain the tester specific adapter, 2) a hybrid of tester-strand and driver-strand hybridizes and contains one tester-derived adapter, 3) both strands originate from the tester sample and contain the tester-derived adapter on each end. After the fill-in reaction of single-strand ends, complementary to tester adapter, a PCR reaction targeting the RDA adapter can be performed. The fragments of driver origin lacking the adapter structure are not amplified during PCR. Tester-driver fragments, marked on one side with an adapter, undergo linear amplification and do not represent a unique transposon event in the tester (brain). Lastly, tester-tester fragments, which contain a *de novo* transposon insertion event absent in the driver, are flanked with RDA adapters on both ends, leading to exponential amplification. In context of the present work, a somatic *de novo* retrotransposition in the human brain can be enriched compared to an insertion that is germline-fixed and present in both brain and skin ectodermal DNA. The enriched, unique retrotransposon-flanks are deep sequenced to evaluate the landscape of *de novo* retrotransposition. The retrieved NGS reads are analyzed with a bioinformatical pipeline, including the scanning of reads with specific RDA-primer sequence, mapping of the respective reads to the human genome and differentiation between hg38-annotated, germline-transmitted retrotransposons and newly formed somatic retrotranspositions in five human brain regions. A more detailed description for the experimental and bioinformatical methods of the RDA is presented in chapter 3.1 and Figure 5.

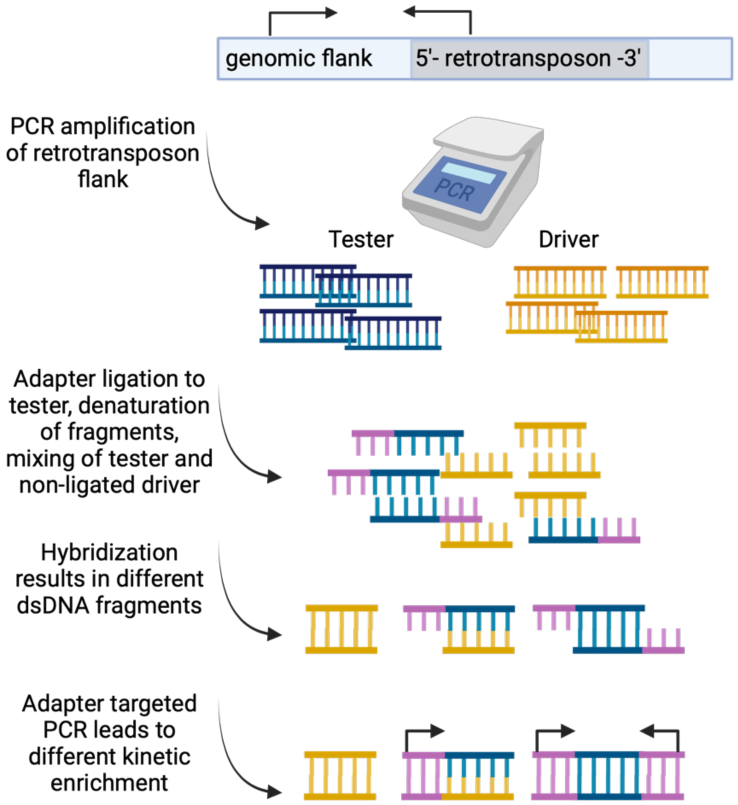

genomic flank
5'- retrotransposon -3'
PCR amplification of retrotransposon flank
PCR
Tester
Driver
Adapter ligation to tester, denaturation of fragments, mixing of tester and non-ligated driver
Hybridization results in different dsDNA fragments
Adapter targeted PCR leads to different kinetic enrichment

Figure 2: General scheme of the subtractive kinetic enrichment process via RDA; a more descriptive illustration is presented in chapter 3.1 (Figure 5). Genomic DNA is fragmented and retrotransposon flanks are amplified via PCR-adapter and TE complementary primer. RDA-adapters can be ligated to the tester sample, while the driver sample remains untreated. The samples are denatured and hybridized with a ratio of 100:1 (driver:tester), resulting in different double-stranded fragments. After a fill-in reaction of single-strand overhangs, a PCR can be performed, resulting in different kinetic enrichment activities. The unique fragments in the tester are exponentially enriched compared to the driver due to the availability of two primer binding sites. This is a modified illustration based on the schematic representation (Figure 1) of Lisitsyn et al. (1993). Created with BioRender.com.

1.5 DNA double-strand break induction and repair

Frederick W. Alt and Bjoern Schwer assessed the existing research on DNA double-strand breaks (DSBs) and proposed in their insightful article (2018) that somatic mosaicism may also arise due to the occurrence of such DNA breakages. Murine models with inactivated non-homologous end joining (NHEJ) repair pathways demonstrated aberrant repair of RAG endonuclease-mediated DSBs, which is essential for the antigen receptor rearrangement processes known as V(D)J recombination. The immune cell receptors are not properly assembled, resulting in the absence of functional B and T lymphocytes. Moreover, in NHEJ-deficient mice, this phenomenon leads to the appearance of lymphomas and primary tumors in murine brains. Therefore, this thesis focuses on the potential link between DNA double-strand breaks (DSBs) and mosaicism to understand their role in genomic instability in the human brain. This means that failed or aberrant DSB repair can cause mutational events such as interstitial deletions or insertions of genomic regions at the DSB site (Varga & Aplan, 2005), so that hotspots of DSBs may generate mutational clusters that differ in tissues or organs.

The genomic DNA of nucleated cells is constantly exposed to exogenous and endogenous damage, including ionizing radiation, chemical agents, reactive oxygen species, replication stress and transcription. Ionizing radiation such as ultraviolet-, gamma- and X-rays can directly damage the helical DNA by collision of high energy particles or indirectly by creating hydrogen and hydroxyl free radicals from H_2O, which react with DNA in close proximity and cause DNA single-strand breaks (SSBs) (Cannan & Pederson, 2016). Reactive oxygen species (ROS) also belong to the class of free radicals and are mostly generated through oxidative stress in a physiological cell when the antioxidant capacity is exceeded. The human brain is a highly metabolic active tissue with elevated energy consumption and thus mitochondrial activity is essential to meet the energy requirements. Consequently, the human brain is exposed to ROS, mainly produced by mitochondrial respiratory complexes, and some neurons as well as areas like hippocampus, amygdala or frontal cortex are more sensitive to oxidative stress, inducing SSBs when excessive ROS production occurs within the cells (Stefanatos & Sanz, 2018). As mentioned earlier, most DNA damaging events result in SSBs and can spontaneously convert to a DSB when more than one SSB is in close proximity to each other. It was estimated that 1% of induced SSBs can convert to a DSB, translating to 10-50 DSBs per cell per cycle (Vilenchik & Knudson, 2003). Another major endogenous DSB contributor, which is initially introduced as SSB, is DNA damage associated with replication stress. The encounter of polymerases and SSB during DNA replication can lead to stalling of the replicative protein machinery, subsequent collapse of the replicative fork and induction of a double-stranded break (Cannan & Pederson, 2016).

Topoisomerases (TOP), which alter the topological state of the DNA double helix, are proposed as an additional contributor to DSB induction during cell cycle and transcription. Depending on the number of DNA strands cleaved by TOP, they can be classified into the single-strand break inducing TOP1 and double-strand break inducing TOP2 (Deweese & Osheroff, 2009). Since TOP2 induces DSBs, they are of particular interest in studying double-strand breaks and repair mechanisms, with the paralogs TOP2A and

TOP2B being present in human. TOP2A is present in cycling cells and involved in processes of DNA replication, whereas TOP2B can be ubiquitously detected and is observed to be a major factor in transcriptional initiation and elongation (Morimoto et al., 2019). The TOP2 homodimer cleaves the DNA, induces a DSB and resides as TOP2 cleavage complex at 5´ ends of DSBs. Under normal conditions the TOP2 cleavage complexes re-ligate to resolve the DSB but the catalysis can fail or abort (Morimoto et al., 2019). If a DSB is not re-ligated, the non-homologous end joining (NHEJ) repair pathway is initiated in the affected cell. Consequently, joining of the DSB ends may result in insertions and deletions due to potential modification of incompatible DNA ends by NHEJ. Since TOP2B is involved in transcription, DSBs can accumulate at active gene sites and pose a potential threat by failed re-ligation of break ends and induction of mutations.

In the presence of DNA lesions, functional physiological cells undergo repair. The predominant repair-pathways of DSBs are reported as non-homologous end joining (NHEJ) and homologous recombination (HR) (Scully et al., 2019). The classical NHEJ initiates repair by binding of the Ku70-80 heterodimer to both ends of a DSB and recruits essential NHEJ pathway proteins to the DNA ends to form a synaptic complex. The DSB ends are processed and ligated by microhomology of limited reference bases in overhanging DNA single-strands. In contrast, HR mediates DNA break-repair through sequence homology between the fragmented DNA and a donor molecule. RAD51 is a key protein enabling the strand invasion and identification of a homologous sequence, thus a templated DNA synthesis and subsequent repair of DSB can be initiated (Scully et al., 2019). Another interesting DNA-repair mechanism, which is of special interest in the present work due to the association of DSBs and retrotransposons, was introduced by Ono et al. (2015) and termed RT-product-mediated DSB repair (RMDR). RMDR uses a pre-existing cDNA that anneals with the DNA ends of a DSB and acts as a ‘bridge’. Alternatively, an RNA anneals with one DSB end and cDNA synthesis is mediated by RT. Similar to NHEJ, the DNA is repaired by microhomologies of template and single-strand ends of a break point.

1.6 DSBs as markers for mosaicism and genomic instability

In contrast to retrotransposon clade markers with undisputable character polarity, DNA double-strand breaks can be dynamic due to DNA repair, requiring an analysis of genomic hotspots to reflect patterns of DSBs and a potential link to mosaicism. This study evaluates the existence of chromosomal DSB hotspots and potential differences in their genomic localization between five human brain regions. By analyzing bulk tissues of distinctive brain regions and their different functionality or activity, we can evaluate the accumulation of DSBs at specific chromosomal positions and consider differences between brain regions as possible precursors of mosaicism. This means DSB-introduced genome instability could cause mosaicism by NHEJ-mediated DNA-end modification, insertions deletions and RMDR with TEs.

The DSB labeling system on the basis of breaks labeling in situ and sequencing (BLISS) (Yan et al., 2017) is introduced in the present work to identify chromosomal DSB hotspots in different human brain regions and to describe the respective ‘breakome’. Besides BLISS, there are several methods reported that identify DSBs, for example BLESS (Crosetto et al., 2013), DSBCapture (Lensing et al., 2016), dDIP (Leduc et al., 2011) and CC-seq (Gittens et al., 2019). BLESS and DSBCapture require a large input, dDIP is used to detect DNA damage including DSBs and SSBs and CC-seq marks only specific features such as covalently bound proteins.

In contrast, BLISS is suitable for direct labeling of DSB-ends by double-stranded adapter, requires low-input because tagged DSBs are amplified by in vitro transcription and DSBs can be quantified through unique molecular identifiers (UMIs). Hence, the present work implements the key features of BLISS to sensitively identify DSB positions in adult human brain regions, including prefrontal cortex, hippocampus, cerebellum, calcarine sulcus and olfactory bulb. DSBs that may contain a single-strand overhang are blunted - to efficiently ligate adapters - using a mix of T4 DNA polymerase, with 3´ → 5´ exonuclease and 5´ → 3´ polymerase activity, and T4 polynucleotide kinase for phosphorylation of the 5´ ends of blunt-ended DNA (Figure 3). The adapter is a double-stranded DNA oligonucleotide that contains a T7 RNA polymerase promotor (blue) and downstream six randomized bases (red) as UMI, which is flanked by two barcodes (green and grey). The barcode sequence adjacent to the promoter also serves as a PCR primer site. The adapter is blocked on one end to limit the ligation of DSB sites to the 3´-end of the adapter-strand recognized by the T7 polymerase and thus set the direction for in vitro transcription towards the genomic DSB flank. To achieve this, one strand is blocked at its 5′-end using Spacer-C3, and the other strand is blocked at its 3′-end with Spacer-C3. The Spacer-C3 is a modified oligonucleotide with an alkyl chain at the 5´- or 3´-end. Next, the labeled DNA and excess adapters are separated on an agarose gel to isolate only the DNA of interest and decrease the adapter-dimer amplification in downstream processes. The genomic DNA is fragmented and linearly amplified via the T7 RNA polymerase-mediated in vitro transcription (5´ → 3´) to increase the sensitivity for detecting uniquely labeled DSBs as well as decrease biases produced by unspecific genomic background. Additional processes include the reverse transcription of DSB-related transcripts and exponential amplification by PCR. The PCR products are deep sequenced and the resulting NGS reads are scanned for adapter specific barcodes. The 6 nt long sequence between the barcodes is extracted as UMI and stored in the corresponding sequence header. The barcode sequences of filtered reads are trimmed and the cleaned reads are stored in FASTQ-files. The paired read file, containing the DSB flank sequences, is mapped to the human reference genome hg38. The DSB coordinates are checked for duplicates, i.e. when a coordinate is extended by +-6 nt and overlaps with another coordinate, which contains the identical UMI, the second coordinate is discarded to deduplicate. A detailed description of the applied bioinformatic workflow is presented in method 2.3.13.1. Overall, the chromosomal locations of DSB hotspots can be compared between the tested brain regions and are analyzed for differences or consistencies.

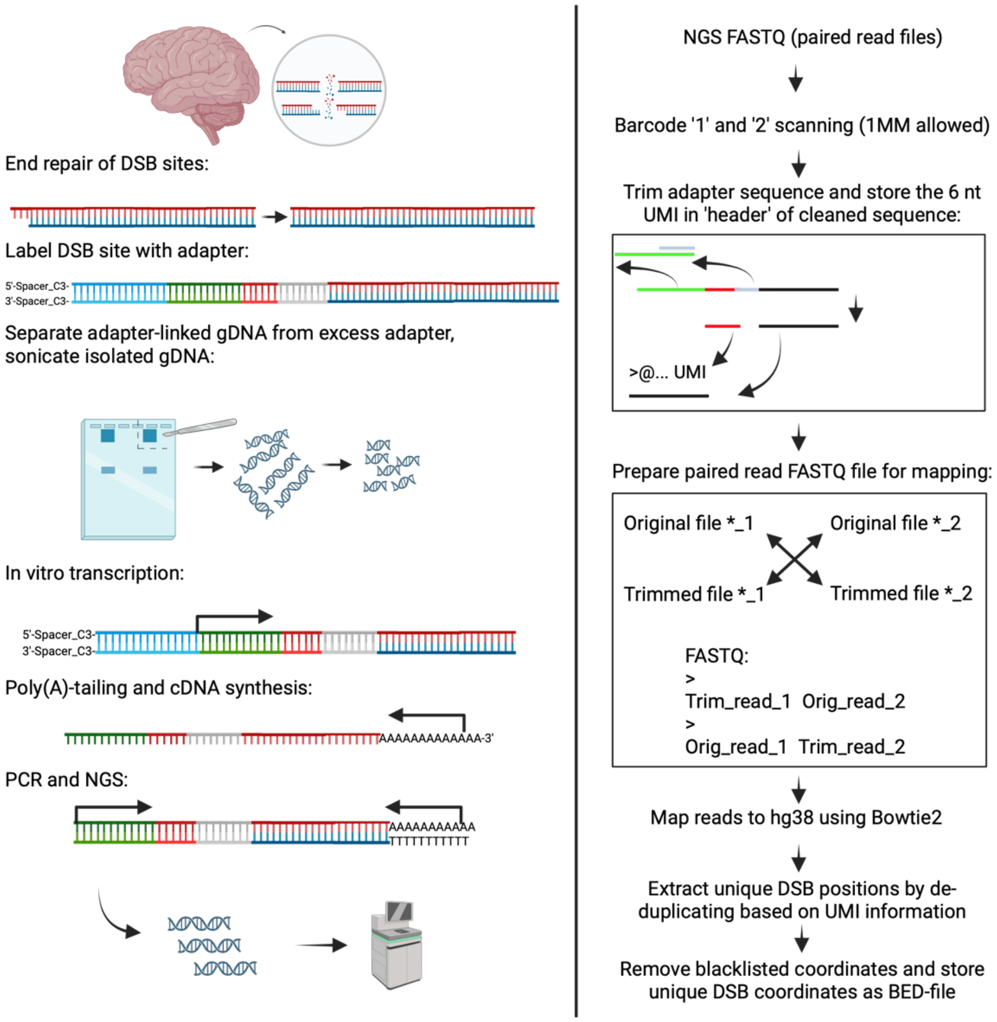

End repair of DSB sites:
Label DSB site with adapter:
5'-Spacer_C3-
3'-Spacer_C3-
Separate adapter-linked gDNA from excess adapter, sonicate isolated gDNA:
In vitro transcription:
5'-Spacer_C3-
3'-Spacer_C3-
Poly(A)-tailing and cDNA synthesis:
AAAAAAAAAAAAA-3'
PCR and NGS:
AAAAAAAAAAA
TTTTTTTTTT
NGS FASTQ (paired read files)
Barcode '1' and '2' scanning (1MM allowed)
Trim adapter sequence and store the 6 nt UMI in 'header' of cleaned sequence:
>@... UMI
Prepare paired read FASTQ file for mapping:
Original file *_1
Original file *_2
Trimmed file *_1
Trimmed file *_2
FASTQ:
>
Trim_read_1 Orig_read_2
>
Orig_read_1 Trim_read_2
Map reads to hg38 using Bowtie2
Extract unique DSB positions by de-duplicating based on UMI information
Remove blacklisted coordinates and store unique DSB coordinates as BED-file

Figure 3: Illustration of experimental and bioinformatic DSB detection. Left panel: Blunted DSBs are directly ligated to the adapter containing a T7 RNA polymerase promotor (blue), six randomized bases (red) as UMI and two barcodes (green and grey). Excess adapters are removed by electrophoretic separation, the adapter-ligated DNA is isolated and fragmented by sonication. In vitro transcription is performed by T7 RNA polymerase and the resulting RNA is polyadenylated. Following reverse transcription, templates can be amplified by PCR targeting the poly(A)-tail and adapter. Finally, the DSB target sequences are obtained using NGS (150 bp paired). The experimental illustration is modified based on Figure 1 of Yan et al. (2017). Right panel: The NGS reads are scanned for the introduced barcodes allowing one mismatch (MM). Reads containing the

barcode are trimmed and the UMI is stored in the sequence header. The headers of the trimmed forward sequence file (TRIM_1) are used to grep identical header plus sequence in the untrimmed/original reverse sequence file (ORIG_2). The same procedure applies to ORIG_1 and TRIM_2 to generate a paired FASTQ file. Only read pairs with one trimmed read are accepted because one molecule should be represented by a single ligated adapter, thus eliminating ligation artefacts. The FASTQ input is mapped to the human reference genome (hg38) and all unique DSB coordinates are obtained by utilizing the UMI information. Created with BioRender.com.

1.7 Transposon- or DSB-associated diseases

The relevance of studying retrotransposon insertions and DSBs in the human brain is clearly demonstrated by their effects on the genomic stability and association with several diseases.

Active retrotransposition of TEs promotes a shift towards genomic instability as a consequence of insertions within genic or other intergenic regions. Retrotransposon reintegrations are often associated with insertions/deletions, chromosomal rearrangements, frameshift mutations or disruption of normal gene expression (reviewed by Bhat et al., 2022). The influence of retrotransposons on genomic stability, including L1, SVA and Alu, is already characterized in disease settings, showing the effects of TE integrations and their association with diseases like lung, colon, pancreatic, breast, and ovarian cancer, hemophilia, and leukemia (Bhat et al., 2022). In addition to the profound arguments regarding the association of TEs and cancer, recent studies provided compelling evidence of retrotranspositions that are linked to major CNS diseases. SVA insertions are characterized in many neurological diseases including Parkinson’s disease or X-linked dystonia parkinsonism (XDP) (Aneichyk et al., 2018; Pfaff et al., 2021) as well as L1 was reported to be associated with schizophrenia (Bundo et al., 2014; Doyle et al., 2017), autism and the rett syndrome (Suarez et al., 2018). Moreover, Baeken et al. (2020) demonstrated increased expression of LINE-1 in Parkinson’s disease models, which are characterized by an inhibited mitochondrial chain complex, and argue for a retrotransposon activation by mitochondrial distress.

Endonucleases, encoded by L1, also affect the genome stability by induction of DSBs and interference with DNA repair mechanisms (Gasior et al., 2006). In addition, as reviewed by Ciccia & Elledge (2010), one major contributor to DSB induction is the high oxygen consumption of the mitochondrial respiration and resulting oxidative stress. Failed or aberrant DNA repair can cause neuronal death and neurodegeneration, which was also described for neurogenerative disorders like Parkinson’s disease, Alzheimer’s and Huntington’s disease that are associated with mutations in mitochondrial DNA and increased ROS levels. Moreover, Suberbielle et al. (2013) demonstrated an increased occurrence of DSBs in Alzheimer’s disease mouse models and observed that Aβ oligomers can cause DSB formation in neuronal cells.

Concluding, the retrotransposon and DSB landscape can shape the genome stability and affect the function of the human brain, thus investigating underlying mechanisms and frequencies of retrotransposon or DSB events in distinctive brain regions is crucial to describe the CNS in health and disease.

1.8 Brain areas of interest and their function

The brain regions subjected to the RDA- and BLISS-based methods in this particular study include the prefrontal cortex, hippocampus, olfactory bulb, calcarine sulcus and cerebellum (Figure 4). The cerebellum is part of the metencephalon, which originates from the rhombencephalon, the third brain vesicle. The other examined brain structures derive from the prosencephalon, the first brain vesicle, which divides into telencephalic and diencephalic structures. The olfactory bulb, as part of rhinencephalon (olfactory brain), is the evolutionary oldest telencephalic structure. The hippocampus belongs to the telencephalic archicortex, which develops earlier than the analyzed neocortical structures. The neocortex serves as the origin for both the prefrontal cortex and the calcarine sulcus.

The prefrontal cortex, often referred to as working memory, is located in the frontal lobe of the human brain, receives input from other cortical regions to process information and is connected with multiple cortical regions to send information for the purpose of adjusting to varying circumstances or situations. This reciprocal connection is necessary for reacting to the individual’s perceptions and thus planning, strategy and executive decisions can be attributed to the prefrontal cortices as their main responsibility (Hathaway & Newton, 2022).

The hippocampus is a part of the limbic system and located in the parahippocampal gyrus inside the temporal horn of the lateral ventricle. The main hippocampal functions include memory processing and consolidation, converts short-term memory into long-term memory and accesses information, including auditory, visual, olfaction and tactile senses, when needed for decision making in the future (Fogwe et al., 2022).

The calcarine sulcus is located on the medial surface of the occipital lobe as a fissure that extends from the parieto-occipital sulcus to the occipital pole and is related to the primary visual cortex. This cortical region (granular cortex) is highly specialized and important for the perception of visual stimuli, including visual components such as orientation and direction (Rehman & Al Khalili, 2022).

The cerebellum, as predominant part of the hindbrain, is located behind the pons and medulla oblongata and harbours 80% of the brain's neurons. The main function of the cerebellum is the coordination of movement and includes the control of posture, muscle tone and muscle activity (Jimsheleishvili & Dididze, 2022).

The olfactory bulb is part of the rhinencephalon and located on the inferior side of the human frontal lobe. The main function of this olfactory system is to receive odor information from the nose through sensory neurons, which are then processed in the bulb as first processing site in the brain (Menini, 2010).

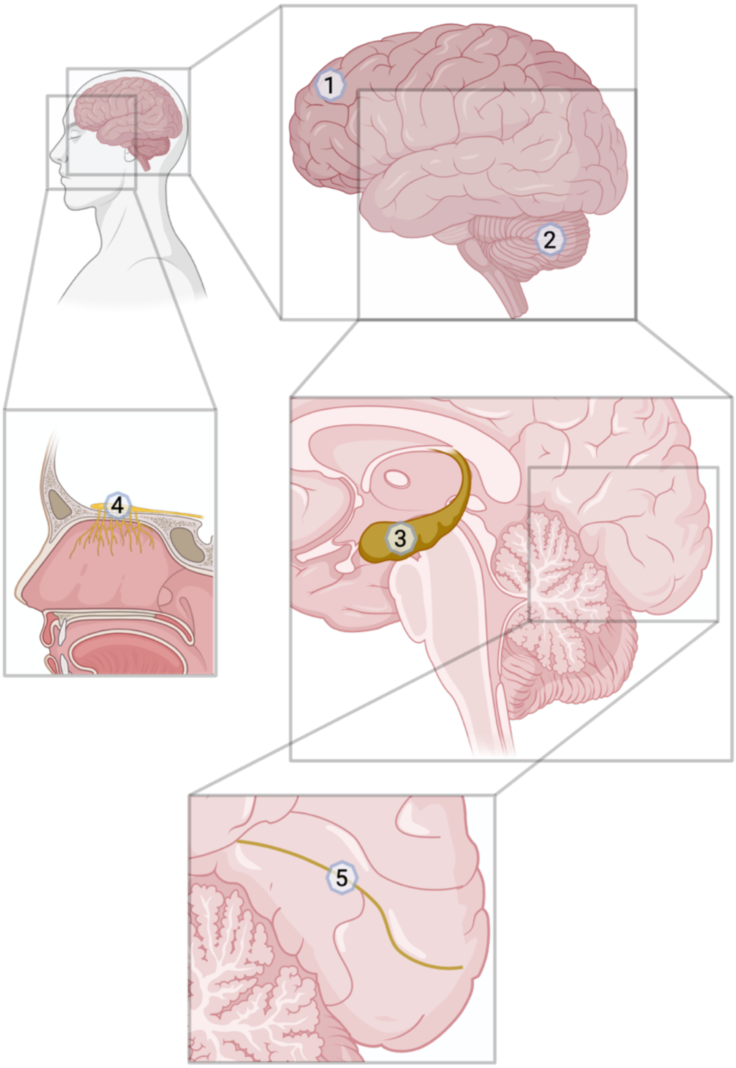


Figure 4: Localisation of the analyzed human brain regions that are isolated for RDA- and BLISS-based methods. (1) prefrontal cortex, (2) cerebellum, (3) hippocampus, (4) olfactory bulb and (5) calcarine fissure. Created with BioRender.com.

1.9 Aim and objectives of the thesis

The present work is organized into two main chapters: the first chapter describes the RDA-detection of retrotransposons and explores the implications of transposon-induced mosaicism in the adult human brain. The second chapter introduces the BLISS centered approach to pinpoint DSBs, accompanied by the evaluation of chromosomal DSB landscapes across distinct brain areas.

Addressing the first objective involves the identification of retrotransposon activity in the human brain. As a result of this, the implemented NGS-coupled RDA method and potential differences in the tested brain regions can be evaluated in a somatic mosaic context. Together with other methods introduced by Baillie et al. (2011) and Evrony et al. (2015), I anticipate to expand the scope of *de novo* retrotransposon detection and provide a new method in the field of cell tracing research. The clear character state of a unique retrotransposon loci, with their definition by RDA, might be well-suited for distinctively tracking cell lineages based on somatic differences.

The second goal involves the confirmation of chromosomal DSB hotspots in the adult human brain and to demonstrate differences in chromosomal localization across the tested brain regions because each area produces different amounts of mtROS and transcriptional patterns that can contribute to DSB induction in the highly metabolic active brain (Stefanatos & Sanz, 2018). Hotspots of DSBs may generate mutational clusters that differ in tissues or organs and may lead to somatic mutations and genomic alterations, which can arise in dividing neural progenitors and contribute to the diversity of neuronal cell types. Moreover, I aim to identify fragile genes that may be related to neuronal diseases, while placing the distribution of DNA double-strand breaks in the context of genome instability.

Finally, with the detection of *de novo* retrotransposon integrations and pinpointing DSB hotpots, I aim to identify overlapping regions of both features to potentially describe their relation because L1 endonucleases induce DSBs (Gasior et al., 2006) and retrotransposon transcripts can be used to ligate breakpoints in RT-product-mediated DSB repair (RMDR) (Ono et al., 2015). The intersection of DSBs and retrotransposons has the potential to reveal the origins of DSB sites; vice versa, could explain the repair of DSBs and the introduction of new transposon regions, thus the two phenomena may complement each other.

2. Material and methods

2.1 Material

RDA:

- Tissue samples (dermis, prefrontal cortex, hippocampus, calcarine fissure, olfactory bulb, cerebellum)
- QIAamp® DNA Mini Kit (Qiagen, 51306)
- Heater with rocking platform
- centrifuge
- RNase A
- Nuclease-free water
- Qubit 2.0 Fluorometer
- Qubit™ dsDNA BR Assay Kit
- MboI (NEB, R0147S)
- ROTI® phenol/chloroform/isoamyl alcohol (25:24:1)
- 100% ethanol
- 3 M sodium acetate
- glycogen (10 µg/µl)
- RDA primer (method 2.2.3)
- 5 M NaCl
- T4 DNA ligase (NEB, M0202S)
- Amicon® Ultra 0.5mL Centrifugal Filters 3K (Merck Millipore, UFC500396)
- Taq PCR Core Kit (Qiagen, 201225)
- PCR thermo cycler
- 10 mM TRIS (pH = 7.9)
- 500 mM EDTA (pH = 8)
- mineral oil
- Exonuclease I (NEB, M0293S)

BLISS:

- Tissue samples (prefrontal cortex, hippocampus, calcarine fissure, olfactory bulb, cerebellum)
- Proteinase K (ThermoFisher, EO0491)
- Lysis buffer (10 mM Tris pH 8.0, 100 mM NaCl, 10 mM EDTA pH 8.0, 0.5% SDS)

- Heater with rocking platform
- centrifuge
- RNase A (20 mg/ml)
- Nuclease-free water
- Qubit 2.0 Fluorometer
- Qubit^{TM} dsDNA BR Assay Kit
- ROTI$^{®}$ phenol/chloroform/isoamyl alcohol (25:24:1)
- 100% ethanol
- 3 M sodium acetate
- glycogen (10 µg/µl)
- 5 M NaCl
- T4 DNA ligase (NEB, M0202S)
- Taq PCR Core Kit (Qiagen, 201225)
- PCR thermo cycler
- Exonuclease I (NEB, M0293S)
- Quick Blunting^{TM} Kit (NEB, E1201S)
- Bliss primer
- 1.5% agarose gel
- Ethidium bromide solution (10 mg/ml) (Carl Roth GmbH, 2218.2)
- Electrophoretic chamber
- Orange DNA loading Dye (6x) (ThermoFisher, R0631)
- QIAquick$^{®}$ Gel Extraction Kit (Qiagen, 28704)
- Covaris S-series
- TUBE AFA Fiber Slit Snap-Cap 6 x 16 mm (Covaris)
- SpeedVac Vacuum Concentrator
- HiScribe™ T7 Quick High Yield RNA Synthesis Kit (NEB, E2050)
- DNase I (NEB, M0303)
- E. coli Poly(A) Polymerase (NEB, M0276)
- Biozym cDNA synthesis Kit (Biozym, 331470L)

2.2 Methods: RDA

The detailed protocols of RDA are described in the published chapter 3.1. The subchapter of the RDA (3.2), focusing on L1, implements the same experimental procedure as described in 3.1, thus only changes are listed in the respective methodical section (2.2).

2.2.1 Sample preparation L1-RDA

The tissue samples (dermis, prefrontal cortex, hippocampus, calcarine fissure, olfactory bulb and cerebellum) of one male adult donor, which is the same individual as described in 3.1 as donor 1, were provided by the Institute of Anatomy, University Medical Center of the Johannes Gutenberg-University Mainz. The tissue samples were immediately snap frozen after dissection and stored at -80°C until further processing.

2.2.2 Methodical approach L1-RDA

The experiments to enrich 5′-flanking regions of *de novo* inserted L1s are carried out as described in methods 3.1.3 with minor changes.

The primer system is changed to specifically target L1 and associated flanking regions during target site PCR (5´-gggagtgacccgattttccag-3´ for initial PCR and 5´-atcggtgatcctcagatggaaatgcagaaatc-3´ for semi-nested PCR; primer design is described in 2.2.3). The primers targeting the ligated adapter sequences are identical to the primer system of the implemented SVA-RDA experiments (see Table 1; YAdAampli, OBam24HZRAD). The target site PCR (see method 3.1.3.6) is carried out according to the cycler program with an annealing temperature of 57°C.

2.2.3 RDA oligos

The primers for amplification of genomic SVA positions are adopted from the Me-Scan-SVA method (Ha et al., 2016) and primers targeting the 5´-UTR of LINE-1 are designed based on a consensus sequence generated with SeaView 4.0 ‘muscle’ alignment option (Galtier et al., 1996; Gouy et al., 2010). To that end, intact LINE-1 coordinates (Human Full-Length, Intact LINE-1 Elements FLI-L1, version 2016-06-01), obtained from L1base2 (Penzkofer et al., 2017; https://l1base.charite.de/l1base.php; accessed on 24 November 2021) are processed with BEDTools v2.30.0 function ‘getfasta’ to retrieve the corresponding sequences of hg38 (Quinlan & Hall, 2010). The consensus sequence is build and the outward-primer of ORF0 (5´-gggagtgacccgattttccag-3´) and a semi-nested primer (5´-atcggtgatcctcagatggaaatgcagaaatc-3´), targeting the 5´-UTR upstream of ORF0 and introducing a new ‘GATC’-restriction site, are generated (Table 1).

Table 1: RDA oligos

Oligo name	Sequence (5´-3´)
YMboAdlong	gcagaagacggcatacgagatggcattccggtct
YMboAdshort	gatcagaccggaatgcc
YAdAampli	gcagaagacggcatacgagat
SVAoutfirst	agaatcaggcagggaggttg
L1outfirst	gggagtgacccgattttccag
SVAoutnested	atctgtgatcagtacmgtccagcttcggct
L1outnested	atcggtgatcctcagatggaaatgcagaaatc
OBam24HZRAD	accgacgtcgactatccatgaacg
OBam12HZRAD	gatccgttcatg

2.2.4 Sequencing

PCR product sequencing with 150 paired-end strategy and resulting raw data as FASTQ files were provided by Novogene Co., Ltd. using the Illumina NovaSeq 6000 platform.

2.2.5 Bioinformatic scanning of *de novo* LINE-1 flanks

The paired sequencing reads of FASTQ-files are merged with PEAR v0.9.6 (J. Zhang et al., 2014) and scanned for the introduced RDA primer. Reads containing the RDA primer are collapsed to non-redundant sequences. The collapsed sequences are scanned for the consensus LINE-1 sequence (gatcctcagatggaaatgcagaaatc) with a ‘GATC’-restriction site, which was introduced in PCR by the semi-nested primer, and are extracted. The adapter sequences and LINE consensus sequence are cut off to retrieve only the genomic flanking region of L1s. The flank sequences are further cleaned with an additional refinement process, which includes the removal of potential L1 related sequences adjacent to the scanned L1 consensus sequence. To that end, the flanks are mapped to complete and intact LINE-1 sequences from the hg38 reference genome (retrieved from L1base2, see method 2.2.3) using BLAT v. 36 as mapping tool (Kent, 2002). L1 mappable parts of the flank reads are eliminated and the cleaned flanks are length filtered (> 30 nt) using SeqKit version 2.0.0 (Shen et al., 2016).

The cleaned flanks are mapped to hg38 with BLAT v. 36 and only reads with one chromosomal alignment are accepted. As internal control, to check for re-alignments of flank reads to known, hg38 annotated L1, the retrieved coordinates of BLAT are intersected with all hg38 annotated L1 coordinates with a 3 kb flank, obtained by UCSC Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables; accessed on 06 January 2022) with RepeatMasker track (Karolchik et al., 2004; Smit et al., 2013; Smit & Hubley, 2008), using BEDTools v2.30.0. The chromosomal positions lacking any overlap with the hg38 annotated L1 coordinates are accepted as *de novo* somatic L1 positions. Additionally, flank reads containing a 'GATC' restriction site between genomic flank region and L1 sequences are eliminated as artefacts because the re-ligation of L1s with a genomic fragment would be seen as a false *de novo* insertion.

The *de novo* L1 coordinates of one sample are extended to a broader chromosomal position when they overlap with each other, thus shared chromosomal positions are collapsed to represent a single L1 integration event. The chromosomal coordinates of *de novo* L1 insertions are stored in BED format. Coordinates that are assigned to one integrational event of LINE-1 can be counted to estimate how many flanks contribute to each event.

The shared L1 positions between brain regions are depicted by Venn diagrams, using Intervene v0.6.5 (Khan & Mathelier, 2017).

2.2.6 Chromosomal distribution of *de novo* L1 insertions

The number of *de novo* L1 insertions per human chromosome (hg38) is calculated by using the 'summary' function in BEDTools v2.30.0 and providing the integration coordinates in BED format. The counts per chromosome are corrected for chromosome size by dividing the chromosome counts by chromosome size and multiply the resulting values by 1 million. The normalized values indicate the number of *de novo* L1 insertions per million chromosomal bp for each chromosome. Additionally, the distribution of *de novo* L1 insertions is depicted by using NCBI's Genome Decoration Page (https://www.ncbi.nlm.nih.gov/genome/tools/gdp; 15.06.2023) with parameter 'Ideogram Cytogenetic: GRCh38.p13 (GCF_000001405.39) Resolution: 850' and using the L1 coordinates in BED format as input.

2.2.7 Genomic feature analysis

To evaluate whether the *de novo* L1 insertions integrate in TE- or gene-rich regions, the human reference genome hg38 is divided into 100 kb windows with 10 kb sliding and all genes or all LINE and SINE occurrences, obtained from the UCSC Table Browser, were calculated for each window. In both approaches, meaning gene or TE counting, the windows exceeding the average occurrence are extracted and the respective coordinates are intersected with the *de novo* L1 coordinates using BEDTools v2.30.0.

To define the genomic features that are associated with *de novo* L1 positions respectively in close proximity, the *de novo* L1 integration positions of each brain region in BED format are used as input for Homer v4.11 function ‘annotatePeaks.pl’ (Heinz et al., 2010) and hg38 as reference.

De novo L1 integration flanks are extended up- and downstream by 2.5 kb and the respective sequences are retrieved from the hg38 genome by the BEDTools function ‘getfasta’. The GC content of the obtained flanks can be calculated and the null hypothesis for consistency of the average GC content of flanks and genomic average GC content (hg38 = 40.9%) can be tested with two-tailed Student's t-distribution.

2.3 Methods: BLISS

2.3.1 Sample preparation

The tissue samples (prefrontal cortex, hippocampus, calcarine fissure, olfactory bulb and cerebellum) of one male adult donor, the same individual as described in chapter 3.1 as donor 1, were provided by the Institute of Anatomy, University Medical Center of the Johannes Gutenberg-University Mainz. The tissue samples were immediately snap frozen after dissection and stored at -80°C until further processing.

2.3.2 DNA isolation

The DNA isolation from bulk tissue of the human brain is carried out by incubation of 100 mg tissue in 500 µl 0.5% SDS lysis buffer and 20 µl Proteinase K (20 mg/ml) at 54°C for 2-3 hours. Afterwards, 20 µl RNase A (20 mg/ml) is added to the mixture, incubated for 10 min at RT and subsequently the aqueous phase is extracted with a standard phenol/chloroform/isoamyl alcohol (25:24:1) isolation. The DNA of the aqueous phase is isolated with addition of 1/10 volume 3 M NaOAc and equal volume of 100% isopropanol, incubated for 15 min at RT and centrifuged at 13 000 x g, 15 min at 4°C. Additionally, the pellet is washed with 70% EtOH and eluted in nuclease-free water. The DNA concentration is estimated using Qubit 2.0 Fluorometer with the Qubit^{TM} dsDNA BR Assay Kit according to the protocol.

2.3.3 Blunting

1 µg genomic DNA (method 2.3.2) of prefrontal cortex, hippocampus, calcarine fissure, olfactory bulb and cerebellum are treated with NEBs Quick Blunting^{TM} Kit according to the manufactures protocol and incubated for 15 min at room temperature.

2.3.4 Ligation of DNA double-strand breaks

To mark double-strand breaks by ligation, 10 µl of adaptor Bliss_adapt_1 (100 pmol/µl) and Bliss_adapt_2 (100 pmol/µl) are mixed with 5 µl 5 M NaCl. The adapter mix is placed on a 100°C heater and gradually cooled to room temperature for hybridization of the complementary oligos. One adapter is blocked at 5′-end by a C3-Spacer, whereas the other adapter has a phosphorylated 5′-end. The sequences contain a T7 RNA polymerase promotor (blue) and downstream six randomized bases (red) as unique molecule identifier (UMI), flanked by two barcodes (green and grey). The barcode adjacent to the promoter also serves as a PCR primer site.

Bliss_adapt_1:

5′-Spacer_C3-taatacgactcactatagggtcagtagcggacnnnnnncatcacgc-3′

Bliss_adapt_2:

5′-P-gcgtgatgnnnnnngtccgctactgaccctatagtgagtcgtatta-C3_Spacer-3′

The ligation with T4 DNA ligase is carried out according to NEBs protocol with 1 µg blunted DNA and 1 µl adapter mix at 16°C overnight.

2.3.5 Gel extraction

To eliminate excess adapter, the ligation mix (method 2.3.4) is separated on 1.5% agarose gel at 100 volt for 10 min and exclusively the genomic DNA band is excised, leaving the adapter remnants on gel. Genomic DNA is isolated from the gel samples utilizing the QIAquick® Gel Extraction Kit according to manufactures protocol and eluted in 130 µl nuclease-free water.

2.3.6 Covaris fragmentation

The ligated and purified DNA is fragmented to approx. 300 bp by Covaris S-series. To that end, the 130 µl DNA (method 2.3.5) is transferred to a TUBE AFA Fiber Slit Snap-Cap 6 x 16 mm and treated with program settings: duty = 10, intensity = 4, cycles/burst = 200, time [s] = 80. Subsequently, the solution is concentrated to 15 µl using a SpeedVac Vacuum Concentrator.

2.3.7 In vitro transcription of RNA

Adapter marked DSB-DNA-fragments with T7 promoter sequence are transcribed as an enrichment step by utilizing NEBs HiScribe™ T7 Quick High Yield RNA Synthesis Kit according to manufactures protocol at 37°C overnight. Subsequently remaining DNA is digested with DNase I at 37°C, 10 min. The remaining RNA is purified with PCI and precipitated as described in 2.3.2.

2.3.8 Poly(A) tailing

In vitro transcribed RNA is polyadenylated to mark the 3´-ends as target for cDNA synthesis, utilizing NEBs E. coli Poly(A) Polymerase according to manufactures protocol at 30°C for 30 min. RNAs are purified according to method 2.3.2.

2.3.9 Synthesis of cDNA

Polyadenylated RNA is reverse transcribed with Biozym cDNA synthesis Kit according to manufactures protocol at 48°C for 1 hour, utilizing d(T)25VN anchored primer.

2.3.10 BLISS-PCR

The DSB loci are amplified using the adapter complementary primer 5′-GGGTCAG-TAGCGGAC-3′ and the d(T)25VN primer of cDNA synthesis. Amplification is performed by standard Qiagen Taq PCR core reaction protocol with 50 ng of template and the cycler program of Table 2. Excess primers in the PCR product samples are eliminated with the addition of exonuclease I (NEB) according to manufacturer’s protocol.

Table 2: Cycler program of the BLISS-PCR

Step	Temperature [°C]	Time [s]	Cycle
Initial denaturation	95	180	1
Denaturation	95	40	35
Annealing	48	60	
Elongation	72	40	
Final elongation	72	300	1
Hold	4	-	-

2.3.11 Agarose gel electrophoresis

The standard gel electrophoresis is prepared by mixing 30 ml of 1.5% agarose in 1 x TBE with 1.5 µl Ethidium bromide solution (10 mg/ml). Samples are loaded with 6 x loading dye, separated on gel in 1 x TBE at 100 V for 10-30 min and validated on the Intas UV system.

2.3.12 Sequencing

PCR product sequencing with 150 paired-end strategy and resulting raw data as FASTQ files were provided by Novogene Co., Ltd. using the Illumina NovaSeq 6000 platform.

2.3.13 Bioinformatics

2.3.13.1 DSB scanning

The bioinformatic scanning of DSB hotspots (Figure 3) of the tested brain NGS datasets is carried out with a custom python and bash script, which were generated and kindly provided by M. Scheuren.

The bioinformatic analysis starts with a bash script executing the python script as well as generate folders and provides the FASTQ-files. The python script starts with the scanning of sequencing reads for barcode 1 (1 mismatch allowed) of DSB-ligated adapter. Reads containing the first barcode are scanned for the second barcode (1 mismatch allowed) and the 6 nt long sequence in-between is extracted as UMI and stored in the corresponding sequence header. The Barcode-UMI-Barcode sequence of filtered reads are eliminated and the trimmed reads are stored in FASTQ-files. The cleaned reads are alphabetically ordered in respect to their sequencing-header and forward and reverse sequence files (paired-sequencing) are prepared for generating a single file in paired read format (related reads in one line). To that end, the headers of the trimmed forward sequence file (TRIM_1) are used to grep identical header plus sequence in the untrimmed/original reverse sequence file (ORIG_2). Same procedure applies for ORIG_1 and TRIM_2. Writing the forward and reverse sequence in the same line is executed when only one of the two reads contains the adapter structure. Read pairs that both contain an adapter are discarded (reads with adapter can be identified by UMI information in header) because a read pair represents a single molecule with only one adapter, otherwise it would display a ligation-artefact. The paired read file is mapped to the human reference genome hg38 using Bowtie2 (Langmead & Salzberg, 2012), allowing 1 mismatch, discarding non-aligned reads and setting ‘local’ for sensitive mapping. The Bowtie2 SAM output is converted to BAM, using SAMtools (Danecek et al., 2021), and subsequently converted to a BED file containing the mapped coordinates and UMI information. Next, the python script extracts unique DSB events. To that end, mapped reads with no mapped partner (paired-read) are eliminated to retrieve only the paired coordinates. Each pair is extended to one single larger position by merging the respective coordinates and store their UMI information. The resulting coordinates are checked for duplicates, i.e. when a coordinate is extended 6+- nt and overlaps with another coordinate and contains the identical UMI (with 1 mismatch still seen as identical), the second coordinate is discarded to deduplicate. This results in unique DSB events that are stored as BED coordinates. The BED coordinates that are listed in a blacklist reference are discarded and the remaining coordinates are accepted as true DSB position and used for the downstream analysis. The ENCODE blacklist (Amemiya et al., 2019) contains coordinates that are anomalous or contain high signals in NGS experiments.

2.3.13.2 Macs peak calling

The DSBs hotspots are identified by using MACS version 3.0.0a7 with 'callpeak', 'nomodel' and 'no control' options and providing the unique DSB positions in BED format as input.

2.3.13.3 DSB hotspot statistics

The DSB counts per MACS peak can be obtained by using BEDTools intersect option with ‘-c’ and providing the peak coordinate and unique DSB coordinate files as input.

To calculate the average DSB density of the human genome, the hg38 reference genome is chopped into windows of the average macs peak length of the corresponding sample. Subsequently, all unique DSB positions can be counted for each window using BEDTools. The DSB density across all hg38 windows with a DSB count greater 0 can be calculated by dividing the DSB number of a window by window size and multiply by 100.

2.3.13.4 Chromosomal distribution of DSB hotspots

The chromosomal distributions (hg38 as reference) of DSB hotspots are counted by using the BEDTools ‘summary’ function and providing the macs peak coordinates as input. To illustrate the comparison of observed and expected distribution, the hotspot coordinates of each sample are used to randomly shuffle new positions with the BEDTools v2.30.0 ‘shuffle’ option on hg38 (Quinlan & Hall, 2010). Following, the random hotspots are counted for each chromosome using the BEDTools ‘summary’ function. The average occurrence (n = 3) of random positions is set as the expected value for each chromosome and the fold change is calculated by dividing the observed chromosome counts by expected mean count.

2.3.13.5 Feature analysis of DSB hotspots

The underlying genomic features of the DSB peaks are annotated using the HOMER v4.11 function ‘annotatePeaks.pl’ with hg38 as reference (Heinz et al., 2010). The HOMER annotated genes are analyzed with Metascape v3.5.20230101 (Zhou et al., 2019).

2.3.13.6 Visualization of chromosomal coordinates

Chromosomal coordinates are depicted with NCBI’s Genome Decoration Page (https://www.ncbi.nlm.nih.gov/genome/tools/gdp; 04.07.2023) with parameter ‘Ideogram Cytogenetic: GRCh38.p13 (GCF_000001405.39) Resolution: 850’ and using the DSB coordinates in BED format as input.

2.3.13.7 Venn diagrams

To visualize shared DSB hotspots between brain regions, Venn diagrams were created using Intervene v0.6.5 (Khan & Mathelier, 2017) and coordinates of DSBs in BED format.

2.3.13.8 Motif search

The genomic sequences of DSB hotspot coordinates are obtained for each sample by using the BEDTools ‘getfasta’ option with hg38 as reference. The sequences are analyzed with the motif enrichment tool SEA v5.5.3 (https://meme-suite.org; accessed on 28 June 23). The default settings are used with type of control sequences ‘Shuffled input sequences’ and ‘Vertebrates (In vivo and in silico)’ as motif database (Bailey & Grant, 2021; PREPRINT).

2.3.13.9 Epigenetic mark analysis

The source and information of datasets like ATAC-seq peaks of dorsolateral prefrontal cortex of the BOCA project (hg38) (Fullard et al., 2018), fragile sites (hg38) (Kumar et al., 2019) or histone ChIP-seq data of the NIH Roadmap Epigenomics Project (Bernstein et al., 2010) are listed in Table 3. The mapped reads of the histone ChIP-experiments (NIH Roadmap Epigenomics Project) are provided as coordinates in BED format on hg19 and therefore are converted to hg38 coordinates using UCSCs hgLiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver; accessed on 17 May 2023). MACS3 was implemented to call peaks using the histone target BED files along with their respective input control. The default settings are selected except for one: the calling mode (narrow or broad) was adjusted based on the type of histone being analyzed. According to the ENCODE guidelines (https://www.encodeproject.org/chip-seq/histone/; accessed on 17 May 2023), the H3K36me3, H3K4me1 and H3K9me3 peaks are called broad and H3K4me3, H3K27ac narrow. The peaks of the same histone marks of both individuals (ID: 112 and 149) are intersected and the overlapping marks are used as the reference BED files.

Additional sequencing datasets derive from the ENCODE Consortium (Dunham et al., 2012; Hitz et al., 2023; Kagda et al., 2023; Y. Luo et al., 2020), downloaded from the ENCODE portal (https://www.encodeproject.org/; accessed on 17 May 2023) and are listed in Table 4. The CTCF ChIP-seq peak files (hg38) of dorsolateral prefrontal cortex (DLPFC) samples (n = 6) are intersected with BEDTools and the shared positions are used as one reference file. The DNase-seq peak files (hg38) of prefrontal cortex samples (n = 8) are also intersected to retrieve the shared peaks as reference file.

The R-loop data of prefrontal cortex (same individual as used for DSB analysis) were collected in our lab in the course of the bachelor thesis of M. B. The peak data were generated with DRIP-seq and bioinformatic methods as described by Scheuren et al. (2023).

After preparing all reference peak files, the chromosomal coordinates of different epigenetic marks in BED format are intersected with the 2538 DSB hotspot coordinates of prefrontal cortex using BEDTools to retrieve overlapping positions. Next, the DSB peaks that overlap with a certain mark are extracted for each epigenetic dataset and a pairwise comparison of the datasets is carried out with Intervene v0.6.5 (Khan & Mathelier, 2017).

The 2538 DSB hotspot of prefrontal cortex are shuffled with BEDTools and hg38 as reference to generate a total of n = 3 random DSB hotspots datasets. The random coordinates of each dataset are intersected with the epigenetic mark coordinates as described for the original DSB dataset of prefrontal cortex. The resulting overlapping peak counts for each shuffled dataset are averaged for n = 3 and used as the expected value. The observed counts (DSB peaks overlapping with epigenetic mark) are divided by expected values (shuffled peaks overlapping with epigenetic mark) to obtain fold change values. Further analysis of certain DSB peaks is carried out with HOMER and Metascape.

Table 3: Dataset information (BOCA, fragile sites, NIH Roadmap Epigenomics Project)

Target	Tissue	Data source	Publication
ATAC-seq	prefrontal cortex	https://labs.icahn.mssm.edu/roussos-lab/boca/ (accessed on 28 March 2022) GEO accession: GSE96949	PMID: 29945882
Fragile sites	Collection	https://webs.iiitd.edu.in/raghava/humcfs/download.html (accessed on 12 June 2023)	PMID: 30999860
Input	frontal lobe	GEO accession: GSM669960 ID:112	PMID: 20944595
H3K9me3	frontal lobe	GEO accession: GSM669965 ID:112	PMID: 20944595
H3K36me3	frontal lobe	GEO accession: GSM669982 ID:112	PMID: 20944595
H3K4me1	frontal lobe	GEO accession: GSM670015 ID:112	PMID: 20944595
H3K4me3	frontal lobe	GEO accession: GSM670016 ID:112	PMID: 20944595
H3K27ac	frontal lobe	GEO accession: GSM1112810 ID:112	PMID: 20944595
H3K9me3	frontal lobe	GEO accession: GSM772834 ID:149	PMID: 20944595
Input	frontal lobe	GEO accession: GSM773010 ID:149	PMID: 20944595
H3K4me3	frontal lobe	GEO accession: GSM773012 ID:149	PMID: 20944595
H3K36me3	frontal lobe	GEO accession: GSM773013 ID:149	PMID: 20944595
H3K4me1	frontal lobe	GEO accession: GSM773014 ID:149	PMID: 20944595
H3K27ac	frontal lobe	GEO accession: GSM773015 ID:149	PMID: 20944595

Table 4: Dataset information (ENCODE Consortium)

Experiment	Dataset	Tissue	Target	Lab of experiments and data production
ENCSR378KET	ENCFF246XGW	DLPFC tissue male adult (83 years)	CTCF	Bradley Bernstein, Broad; ENCODE Processing Pipeline
ENCSR452KYY	ENCFF306BLG	DLPFC tissue male adult (84 years)	CTCF	Bradley Bernstein, Broad; ENCODE Processing Pipeline
ENCSR979PTL	ENCFF816RIB	DLPFC tissue male adult (86 years)	CTCF	Bradley Bernstein, Broad; ENCODE Processing Pipeline
ENCSR813KUE	ENCFF512PKN	DLPFC tissue male adult (78 years)	CTCF	Bradley Bernstein, Broad; ENCODE Processing Pipeline
ENCSR832TWW	ENCFF535ATM	DLPFC tissue male adult (82 years)	CTCF	Bradley Bernstein, Broad; ENCODE Processing Pipeline
ENCSR374PKX	ENCFF132UGC	DLPFC tissue male adult (83 years)	CTCF	Bradley Bernstein, Broad; ENCODE Processing Pipeline
ENCSR006MAW	ENCFF620YIX	DLPFC tissue male adult (83 years)	DNase-seq	John Stamatoyannopoulos, UW; ENCODE Processing Pipeline
ENCSR811OUF	ENCFF750SYW	DLPFC tissue male adult (86 years)	DNase-seq	John Stamatoyannopoulos, UW; ENCODE Processing Pipeline

ENCSR849WGE	ENCFF100UJA	DLPFC tissue male adult (84 years)	DNase-seq	John Stamatoyannopoulos, UW; ENCODE Processing Pipeline
ENCSR880CUB	ENCFF750DUU	DLPFC tissue male adult (83 years)	DNase-seq	John Stamatoyannopoulos, UW; ENCODE Processing Pipeline
ENCSR351FWN	ENCFF351AFQ	DLPFC tissue female adult (88 years)	DNase-seq	John Stamatoyannopoulos, UW; ENCODE Processing Pipeline
ENCSR606QDB	ENCFF521CSC	DLPFC tissue male adult (82 years)	DNase-seq	John Stamatoyannopoulos, UW; ENCODE Processing Pipeline
ENCSR686LOE	ENCFF853PIR	DLPFC tissue female adult (89 years)	DNase-seq	John Stamatoyannopoulos, UW; ENCODE Processing Pipeline
ENCSR386XPD	ENCFF493KKC	DLPFC tissue female adult (82 years)	DNase-seq	John Stamatoyannopoulos, UW; ENCODE Processing Pipeline

2.3.13.10 Analysis of DSB associated SVAs

The unique DSB positions of each brain region that are located in the *de novo* SVA flank of the respective brain region and donor are counted (data collected in experiments of chapter 3.1) by intersecting the coordinates with BEDTools. In addition, 3 random datasets of DSB positions with the same number as the original file are generated with BEDTools and also counted when located in the *de novo* SVA flank positions. The mean (n = 3) of random occurrences in each *de novo* SVA flank is set as the expected value. The observed DSB counts are divided by the corresponding expected count for each *de novo* SVA flank and depicted as Boxplot. *De novo* SVA flanks with a fold change greater 2, when comparing observed and expected, are extracted and the coordinates are analyzed with HOMER.

3. Representational difference analysis: retrotransposon mosaic in the human brain

Retrotransposition contributes to human brain mosaicism and is increasingly considered a possible cause of neurogenetic disorders. Hominoid-specific SVAs and the autonomous mobilizing L1 are of particular interest because they are found to integrate *de novo* in somatic tissues and exhibit high mobility in germline. I asked whether this is reflected in the human brain and used a subtractive and kinetic enrichment technique called representational difference analysis (RDA) coupled with deep sequencing to compare different brain regions with respect to *de novo* TE insertion-patterns. This method introduces SVAs and L1s, and their presence/absence, as clade markers to explain somatic mosaicism in the human brain, and especially SVAs provide new opportunities to explain intra- and inter-individual variations and to reconstruct the phylogeny of cell lineages.

The main chapter of the RDA method (chapter 3.1), focusing on SVAs, is already published under: Möhner J, Scheuren M, Woronzow V, Schumann S and Zischler H (2023) RDA coupled with deep sequencing detects somatic SVA-retrotranspositions and mosaicism in the human brain. Front. Cell Dev. Biol. 11:1201258. doi: 10.3389/fcell.2023.1201258.

The respective supplementary materials of chapter 3.1 are available at: https://www.frontiersin.org/articles/10.3389/fcell.2023.1201258/full#supplementary-material

Citations used within the publication Möhner et al., 2023 (chapter 3.1) are incorporated in chapter ‘References’ of the present thesis and the layout and numbering of the presented figures and tables are adjusted to be in accordance with the thesis.

The subchapter of the RDA (3.2), focusing on L1, implements the same experimental procedure as described in 3.1, thus only changes are listed in the respective methodological section (2.2).

3.1 RDA coupled deep sequencing detects somatic SVA-retrotranspositions and mosaicism in the human brain

Jonas Möhner[1*], Maurice Scheuren[1], Valentina Woronzow[1], Sven Schumann[2], Hans Zischler[1*]

[1]Division of Anthropology, Institute of Organismic and Molecular Evolution, Faculty of Biology, Johannes Gutenberg University Mainz, Mainz, Germany

[2]Institute of Anatomy, University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany

This chapter was published, based on data collected as part of JM's dissertation project, as original research article in Front. Cell Dev. Biol. Sec. Epigenomics and Epigenetics under the title "RDA coupled with deep sequencing detects somatic SVA-retrotranspositions and mosaicism in the human brain" (Möhner et al., 2023; https://doi.org/10.3389/fcell.2023.1201258).

3.1.1 Abstract

Cells of the developing human brain are affected by the progressive acquisition of genetic and epigenetic alterations that have been reported to contribute to somatic mosaicism in the adult brain and are increasingly considered a possible cause of neurogenetic disorders. A recent work uncovered that the copy–paste transposable element (TE) LINE-1 (L1) is mobilized during brain development, and thus mobile non-autonomous TEs like AluY and SINE-VNTR-Alu (SVA) families can use L1 activity in trans, leading to *de novo* insertions that may influence the variability of neural cells at genetic and epigenetic levels. In contrast to SNPs and when considering substitutional sequence evolution, the presence or absence of TEs at orthologous loci represents highly informative clade markers that provide insights into the lineage relationships between neural cells and how the nervous system evolves in health and disease. SVAs, as the 'youngest' class of hominoid-specific retrotransposons preferentially found in gene- and GC-rich regions, are thought to differentially co-regulate nearby genes and exhibit a high mobility in the human germline. Therefore, we determined whether this is reflected in the somatic brain and used a subtractive and kinetic enrichment technique called representational difference analysis (RDA) coupled with deep sequencing to compare different brain regions with respect to *de novo* SINE-VNTR-Alu insertion patterns. As a result, we detected somatic *de novo* SVA integrations in all human brain regions analyzed, and the majority of *de novo* insertions can be attributed to lineages of telencephalon and metencephalon, since most of the examined integrations are unique to different brain regions

under scrutiny. The SVA positions were used as presence/absence markers, forming informative sites that allowed us to create a maximum parsimony phylogeny of brain regions. Our results largely recapitulated the generally accepted evo-devo patterns and revealed chromosome-wide rates of *de novo* SVA reintegration targets and preferences for specific genomic regions, e.g., GC- and TE-rich regions as well as close proximity to genes that tend to fall into neural-specific Gene Ontology pathways. We concluded that *de novo* SVA insertions occur in the germline and somatic brain cells at similar target regions, suggesting that similar retrotransposition modes are effective in the germline and soma.

3.1.2 Introduction

To date, the origin or the genetic and regulatory-epigenetic mechanisms by which the enormous amount of morphological and functional variability of somatic - here mainly neural - cells is generated remains poorly understood. Several experimental analyses have shown that cells of the brain differentially exhibit somatic genomic variation in a brain region-specific manner, which is partly associated with *de novo* retrotranspositions of transposable elements (TEs). Somatic mutations can be used to study the patterns of progenitor proliferation, migration, and differentiation underlying brain developmental processes. To this end, high-throughput sequencing has been performed to determine single-nucleotide variants (SNVs) (Lodato et al., 2015), whereby position-specific mutation rates, resulting in reversals, possibly create interpretational difficulties in the evaluation of mosaicism. On the other hand, an undisputable character polarity is associated with the retrotransposition of mobile elements. These elements amplify and colonize metazoan genomes by a germline ‘copy-and-paste’ mechanism associated with different activities of long interspersed element-1 (LINE-1 or L1). However, LINE-1 is the only active autonomous retroelement in the human genome, and non-autonomous elements rely on the enzymatic machinery provided by L1 for retrotransposition. Insertional mutagenesis and disease are linked with three families, namely, L1, Alu, and SINE-VNTR-Alu (SVA), all of which rely on ‘copy-and-paste’ mechanisms (reviewed by Cordaux and Batzer, 2009).

L1 expression in the human brain suggests that L1 mobilization may also occur during later development, and this assumption was tested with several NGS-based sequencing strategies such as retroposon capture and comparing the germline with the hippocampus and caudate nucleus (Baillie et al., 2011) and single-cell WGS of neurons (Evrony et al., 2015). Concordantly, somatic insertions of L1, Alu, and SVA sequences were found in different comparative settings and brain regions. In contrast, the absolute rates of somatic LINE-1 element retrotransposition in the brain have been discussed intensively. Moreover, it was suggested that there were brain region-specific rates of mobility.

In the context of hominoid brain evolution, SVAs are of special interest, mainly because they represent the ‘youngest’ class of hominoid-specific retrotransposons. SVAs are comprised of a characteristic $(CCCTCT)_n$ hexamer repeat, Alu-like region, variable number of tandem repeats (VNTRs), and the env-gene plus 3′-LTR from HERV-K10 (H. Wang et al., 2005; Cordaux and Batzer, 2009). SVAs are preferentially found in gene- and GC-rich regions and are thus hypothesized to differentially co-regulate nearby genes (Savage et al., 2013; Gianfrancesco et al., 2019; Barnada et al., 2022).

To define the genomic patterns of somatic *de novo* SVA integrations for different brain regions, we used a subtractive and kinetic enrichment technique coupled with deep sequencing to compare complex somatic genomes (Figure 5A). This method was introduced by Lisitsyn et al. (1993) and is termed representational difference analysis (RDA). Our approach was to specifically amplify the 5′-flanking region of SVAs by using ectodermal DNA from skin as a driver and comparing it with five different brain regions of two adult male donors as testers. To this end, MboI-restricted DNA is ligated with a “GATC”-ligatable adapter in driver and tester samples. The 5′-flanking SVA regions of interest are specifically enriched by PCR with an outward primer system as proposed in a mobile element scanning method for SVA (Ha et al., 2016). The driver PCR products are hybridized with the tester PCR products after ligation of RDA primers exclusively to the tester samples. During hybridization, three possible scenarios can occur: 1) both single strands are of driver origin, thus do not contain the tester-specific adapter, 2) a hybrid of tester-strand and driver-strand is generated and possesses one strand with the tester-derived adapter, and 3) both strands are of tester origin and contain the tester-derived adapters. After hybridization, a PCR reaction with primers specific to tester adapters can be performed. Driver–driver fragments without the adapter structure are not amplified, tester–driver fragments with one-sided adapter marking are linearly amplified and do not represent a unique transposon event in the tester (brain), and lastly tester–tester fragments with a *de novo* transposon insertion event are flanked by the RDA adapter on both sides and consequently amplified exponentially. Accordingly, if somatic retrotranspositions have occurred in the brain, the flank of the newly inserted SVA changes compared to that of skin ectodermal DNA and can be enriched and deep sequenced to estimate the full diversity of the heterogeneous PCR products. The NGS reads were then scanned for RDA primers, and a bioinformatics pipeline (Figure 5B) was developed to distinguish between hg38-annotated germline-transmitted SVAs and newly formed somatic SVA retrotranspositions in the olfactory bulb, cerebellum, prefrontal cortex, calcarine sulcus, and hippocampus. Moreover, to obtain an idea about the frequency of somatic *de novo* insertions of SVA, we took advantage of the well-defined character polarity of SVA insertions that allows each somatic insertion to be traced to a unique molecular event in a common ancestor of all cells descended therefrom.

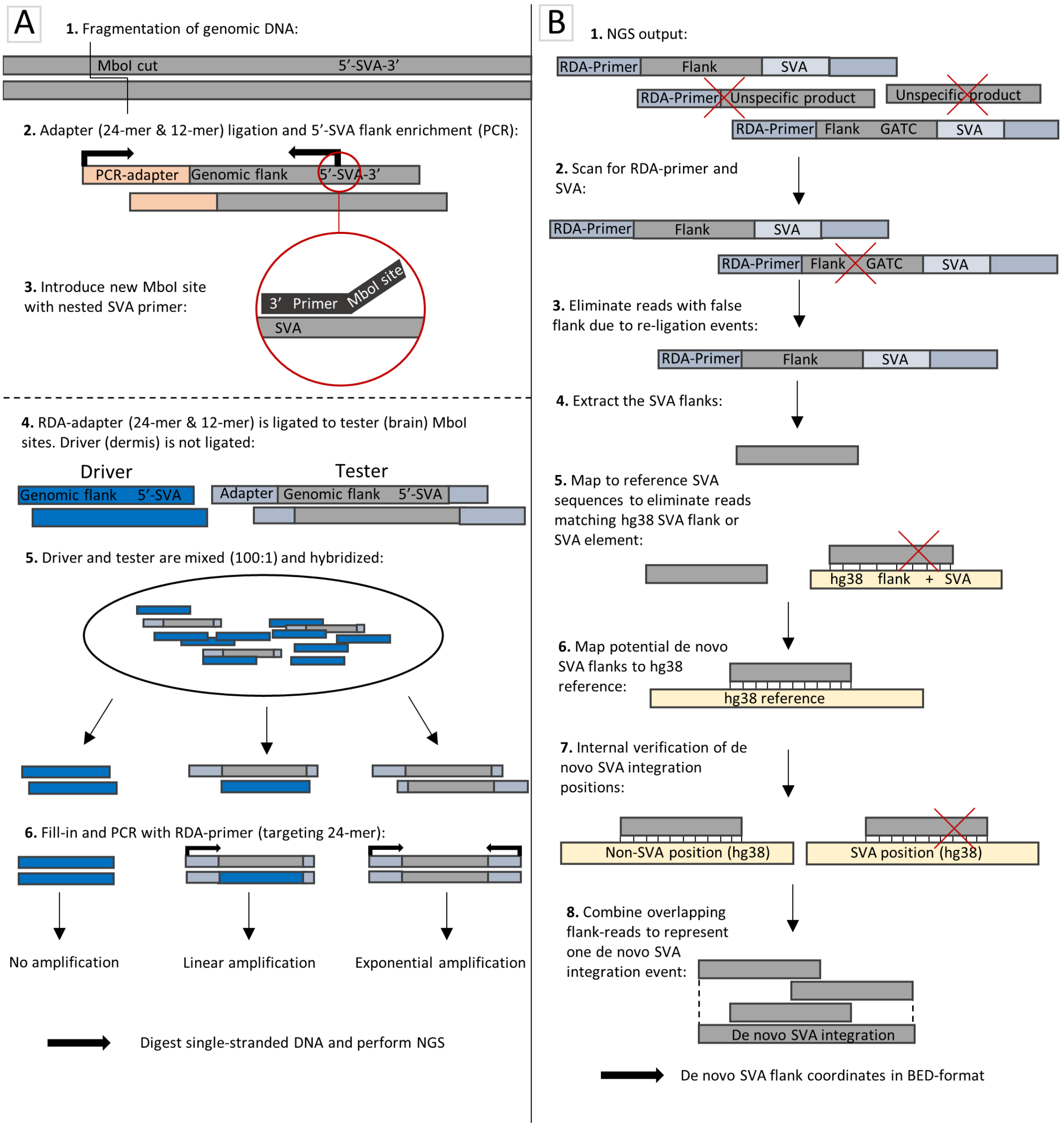

A
1. Fragmentation of genomic DNA:
MboI cut
5′-SVA-3′
2. Adapter (24-mer & 12-mer) ligation and 5′-SVA flank enrichment (PCR):
PCR-adapter
Genomic flank
5′-SVA-3′
3. Introduce new MboI site with nested SVA primer:
3′ Primer
MboI site
SVA
4. RDA-adapter (24-mer & 12-mer) is ligated to tester (brain) MboI sites. Driver (dermis) is not ligated:
Driver
Tester
Genomic flank 5′-SVA
Adapter
Genomic flank 5′-SVA
5. Driver and tester are mixed (100:1) and hybridized:
6. Fill-in and PCR with RDA-primer (targeting 24-mer):
No amplification
Linear amplification
Exponential amplification
Digest single-stranded DNA and perform NGS
B
1. NGS output:
RDA-Primer
Flank
SVA
RDA-Primer
Unspecific product
Unspecific product
RDA-Primer
Flank
GATC
SVA
2. Scan for RDA-primer and SVA:
RDA-Primer
Flank
SVA
RDA-Primer
Flank
GATC
SVA
3. Eliminate reads with false flank due to re-ligation events:
RDA-Primer
Flank
SVA
4. Extract the SVA flanks:
5. Map to reference SVA sequences to eliminate reads matching hg38 SVA flank or SVA element:
hg38 flank + SVA
6. Map potential de novo SVA flanks to hg38 reference:
hg38 reference
7. Internal verification of de novo SVA integration positions:
Non-SVA position (hg38)
SVA position (hg38)
8. Combine overlapping flank-reads to represent one de novo SVA integration event:
De novo SVA integration
De novo SVA flank coordinates in BED-format

Figure 5: (A) SVA–RDA workflow. (A1) Genomic DNA was fragmented using MboI. Fragments were ligated with the “GATC”-ligatable PCR-adapter (double-strand consisting of a 24- and 12-mer sequence with sticky ends). (A2) 5′-SVA flanking regions were amplified by PCR. (A3) A second MboI site was introduced by nested PCR primers targeting the SVA region. (A4) After removing PCR-adapters using MboI, the “GATC”-ligatable RDA-adapter (double-strand consisting of a 24- and 12-mer sequence with sticky ends) can be ligated to both ends of the PCR products of the tester sample, while the driver sample remains untreated and does not contain adapters. (A5) The denatured and hybridized driver and tester samples were mixed at a ratio of 100: 1, resulting in different double-stranded fragments that show different kinetic enrichment activities during PCR depending on the degree of RDA-adapter association. The RDA-primer targets the

24-mer sequence of the RDA-adapter, meaning fragment ends linked to the 12-mer sequence cannot be annealed. Unique fragments in the tester compared to the driver are enriched exponentially. This is a modified illustration based on the schematic representation (Figure 1) of Lisitsyn et al., (1993). (B) Bioinformatical analysis of RDA-SVA NGS data. (B1,B2) Only reads containing the RDA-primer and SVA sequence were extracted and (B3) reads containing an MboI site within the SVA flank were eliminated. (B4) The RDA-primer and SVA sequences were cut off to retrieve the SVA flank. (B5) The cleaned reads were mapped to reference SVAs and associated flanking sequences to exclude germline-fixed positions. (B6) The resulting reads were mapped to the human reference genome to identify *de novo* SVA integration positions. (B7) The positions were checked for re-alignments to known SVAs or SVA flanks as the internal verification step. (B8) The filtered *de novo* SVA flanks that share coordinates are accepted as one SVA integration event and combined into one chromosomal coordinate.

3.1.3 Material and Methods

3.1.3.1 Ethical approval

Human tissue samples were obtained as part of the body donation program of the Institute of Anatomy, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany. The people donated their body voluntarily for medical education and research, and the present study was conducted within the parameters of the written permission we received from the body donor during lifetime. This research on human postmortem tissue was reviewed and approved by the Ethics Committee of Landesärztekammer Rheinland-Pfalz, Mainz, Germany (24/05/2022; Ref.# 2022-16488).

3.1.3.2 Sample preparation

The non-diseased brain and dermis samples of two (Hsa n = 2) male adult donors (dermis, prefrontal cortex, hippocampus, calcarine fissure, olfactory bulb, and cerebellum) were provided by the Institute of Anatomy, University Medical Center of the Johannes Gutenberg University, Mainz. The tissue samples were immediately snap-frozen after dissection and stored at –80°C until further processing.

3.1.3.3 DNA isolation

To isolate DNA from brain regions as well as dermis, the frozen tissues were cut under cooled conditions and aliquots of 20–30 mg were prepared. The genomic DNA of brain samples was isolated utilizing the QIAamp® DNA Mini Kit, following the procedures of “Protocol: DNA Purification from Tissues (QIAamp DNA Mini Kit).”.

3.1.3.4 Fragmentation of genomic DNA

Fragmentation of 1 µg genomic DNA of brain samples (tester) and dermis sample (driver) was carried out with MboI (NEB) according to manufacturer's protocol for 1 h at 37°C. Fragmented DNA was purified using phenol/chloroform/isoamyl alcohol (25:24:1), precipitated with ethanol, and resuspended in nuclease-free water.

3.1.3.5 Flank adapter ligation

Primers 5′-gcagaagacggcatacgagatggcattccggtct-3′ and 5′-gatcagaccggaatgcc-3′ were hybridized by mixing 5 µL (100 pmol/µL) of each primer with 5 µL 5 M NaCl, heating the mixture to 100°C, and gradually cooling down to room temperature to allow annealing. Since the double-stranded adapter contains a "GATC"-overhang, ligation to 1 µg MboI-fragmented DNA can be carried out with NEBs T4 DNA ligase according to manufacturer's protocol with 2 µL adapter solution overnight at 16°C. After ligation, excessive adapter was removed with Amicon® Ultra 0.5 mL Centrifugal Filters (3 K).

3.1.3.6 Target site PCR

PCR amplification of 5′-flanking SVA regions of 50 ng ligated DNA was carried out with 10 pmoles primer 5′-gcagaagacggcatacgagat-3′ (adapter complementary) and SVA consensus outward primer 5′-agaatcaggcagggaggttg-3′ according to the standard Qiagen Taq PCR core reaction, and the details of the thermal profile were as follows: 94°C for 120 s, 30 cycles amplification with 94°C for 40s, 59°C for 40s, and 72°C for 60s and a final elongation at 72°C for 300 s. The semi-nested PCR was performed with the aforementioned thermal profile but amplified for 25 cycles and with the addition of 10 pmoles primer 5′-gcagaagacggcatacgagat-3′ and 5′-atctgtgatcagtacmgtccagcttcggct-3, which introduces a new MboI restriction site at the SVA outward region of the PCR product.

3.1.3.7 RDA adapter ligation

PCR adapters of the tester (brain) and driver (dermis) PCR products were excised with MboI (NEB) at both ends according to manufacturer's protocol and removed using Amicon® Ultra 0.5 mL Centrifugal Filters (3 K) to finally introduce the new RDA adapters. RDA adapters 5′-accgacgtcgactatccatgaacg-3′ and 5′-gatccgttcatg-3′ were hybridized and ligated, only to the tester sample and not driver, as described in Section 3.1.3.5. Finally, the samples were purified using Amicon® Ultra 0.5 mL Centrifugal Filters (3 K).

3.1.3.8 Hybridization of the tester and driver

A 1:100 molar ratio of the tester (50 ng) and driver (5 µg) was set up for hybridization by mixing a total volume of 11 µL DNA, 2 µL 10mM Tris (pH = 7.9), 1 µL 500mM EDTA (pH

= 8), and 2 µL nuclease-free water. The DNA mix and 5 M NaCl were preheated separately at 95°C for 2 min and subsequently 4 µL of 5 M NaCl was added to the DNA. The samples were covered with 30 µL mineral oil, denatured at 95°C for 4 min, and hybridized at 67°C for at least 18 h.

3.1.3.9 RDA PCR

Hybridized samples (250 ng) were prepared according to the standard Qiagen Taq PCR core reaction without the addition of primers and incubated at 72°C for 20 min to fill in overhangs after hybridization. After the addition of primer (10 pmoles) 5′-accgacgtcgactatccatgaacg-3′ to the sample, amplification was carried out with the following thermal profile: 15 cycles amplification with 94°C for 40 s, 60.9°C for 40 s, and 72°C for 60 s and a final elongation at 72°C for 300 s. The samples (2 µL of the PCR product) were reamplified with the standard PCR reaction and primer (10 pmoles) 5′-accgacgtcgactatccatgaacg-3′, and the details of the thermal profile were as follows: 94°C for 120 s, 20 cycles amplification with 94°C for 40 s, 60.9°C for 40 s, and 72°C for 60 s and a final elongation at 72°C for 300 s. Finally, excessive primers in the PCR product samples were eliminated with the addition of exonuclease I (NEB) according to manufacturer's protocol.

3.1.3.10 Sequencing

PCR product sequencing on the Illumina NovaSeq 6000 platform with the 150 paired-end strategy was performed by Novogene Co., Ltd., and resulting raw data were provided as FASTQ files.

3.1.3.11 Bioinformatical scanning of somatic *de novo* SVA positions

Paired-end sequencing reads were merged using PEAR v0.9.6 (J. Zhang et al., 2014) and scanned for RDA primers. All reads containing the adapter were scanned for the SVA characteristic $(CCCTCT)_n$ hexamer repeat and extracted. Reads containing a "GATC" restriction site between the genomic flank region and SVA part were eliminated as artefacts, since re-ligation of SVAs with a genomic fragment would be considered false *de novo* insertion. Hence, only reads with SVA sequences directly flanked by a genomic region were further processed, and the SVA parts were eliminated starting at the $(CCCTCT)_n$ hexamer repeat. The remaining clean flank sequences were collapsed, length-filtered (only extract reads >30 bp) using SeqKit version 2.0.0 (Shen et al., 2016), and mapped to all GRCh38-annotated SVA sequences with 1kb flanking regions, obtained by the UCSC Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables; accessed on 06 January 2022) with the RepeatMasker track (Karolchik et al., 2004; Smit and Hubley, 2008; Smit et al., 2013), using BLAT v. 36 as the mapping tool (Kent, 2002). Reads that were mapped to SVAs or their respective flanking region were discarded, and the remaining reads were mapped to GRCh38 with BLAT v. 36. Only mapped reads with one chromosomal alignment were accepted, and to increase stringency, the obtained chromosomal positions were intersected with chromosomal positions of GRCh38-annotated SVA sequences with 3 kb flank using BEDTools v2.30.0 (Quinlan and Hall, 2010). Chromosomal positions that did not show shared positions with SVA plus flank were accepted as *de novo* somatic SVA positions. *De novo* SVA positions of one sample were extended to a broader chromosomal position when they overlapped with each other; thus, shared/intersecting chromosomal positions were “collapsed" to represent a single SVA integration event (chromosomal coordinates are provided in Supplementary Table S1).

3.1.3.12 Shared SVA position analysis

All *de novo* SVA positions of all brain regions from one person were combined to generate an individual reference BED file. Each brain region’s SVA positions were intersected with the reference file using BEDTools v2.30.0 and denoted as 0 (no overlap of reference position with the analyzed brain region) or >0 (overlap of reference position with a position of the analyzed brain region) to generate a presence/absence list for each brain region, respectively. Hence, each brain area receives a list of all SVA reference positions with character states as: position is present or absent. This list could be converted to a sequence, where one base like “A” is denoted as SVA position present and “T” as absent. The resulting sequences for each brain region were used to generate a maximum parsimony tree with branch lengths as steps and bootstrap resampling (1,000 replicates) utilizing MEGA11 (version 11) (Stecher et al., 2020; Tamura et al., 2021). Additionally, SeaView 4.0 was used to depict informative sites and bootstrap support with 1,000 bootstrap replicates (Galtier et al., 1996; Gouy et al., 2010).

To visualize shared SVA positions between brain regions, Venn diagrams were created using Intervene v0.6.5 (Khan and Mathelier, 2017).

3.1.3.13 Chromosomal SVA density analysis

The “summary” function in BEDTools v2.30.0 was used to estimate chromosomal SVA densities. To this end, chromosomal density of hg38-annotated SVAs was calculated as reference by estimating the average genomic SVA length using hg38 reference positions (UCSC Table Browser), multiplying SVA counts of each chromosome by the average SVA length, and dividing values by the chromosome size. This procedure was also utilized for each brain region by multiplying the SVA counts by hg38-estimated average

SVA length. The arithmetic mean of *de novo* SVA density values for each chromosome was calculated based on the two donors (n = 2), and the correlation coefficient r, as a relation of *de novo* SVA density with hg38 SVA density, was calculated using the "CORREL (array1, array2)" function in Microsoft® Excel v16.70. The *p*-value of correlation interpretation was calculated with the two-tailed Student's t-distribution "T.DIST.2T" function.

3.1.3.14 Genomic feature and Gene Ontology analysis

The *de novo* SVA positions of each brain region in the BED format were used to annotate the related genomic feature using Homer v4.11 with the "annotatePeaks.pl" function (Heinz et al., 2010) and hg38 as reference. Additionally, Homer-annotated genes were analyzed using Metascape v3.5.20230101 (Zhou et al., 2019) for contribution to Gene Ontology pathways.

To validate the number of SVA integrations in SINE/LINE-rich regions, the hg38 reference genome was divided into 100kb windows with a 10kb stagger, and all LINE and SINE occurrences retrieved from the UCSC Table Browser with the RepeatMasker track were counted for each window. The average occurrence of combined SINE and LINE was calculated for all windows and set as a threshold, so that only windows exceeding the average were accepted as TE-enriched. In addition, only the top 25% TE-enriched windows were used as datasets. Both datasets containing the window coordinates were intersected with the *de novo* SVA insertion coordinates, which were annotated using HOMER as LINE- or SINE-associated, using BEDTools v2.30.0.

SVA flanks were extended up- and downstream by 2.5 kb, and the corresponding sequences were obtained from hg38 as reference using the BEDTools function "getfasta". The GC content of the extended flanks was calculated, and the null hypothesis for consistency of the average GC content of flanks and genomic average GC content (hg38 = 40.9%) was tested with two-tailed Student's t-distribution.

3.1.4 Results

3.1.4.1 SVA retrotransposition is active in the human brain and generates somatic mosaicism

The RDA method was applied to enrich unique 5′-SVA-flank templates, precisely *de novo* SVA insertions in brain regions, with template DNA from the dermis of the same individual as the driver sample. This driver sample represents the bulk of 5′-flanks for the germline-transmitted SVAs, both polymorphic and fixated SVA-integrations, and is given in a 100-fold molar excess to the RDA primer-ligated tester during hybridization. Since only tester-sequences are covalently ligated to RDA primers, SVA flanks that are not present in the driver were PCR-enriched, and the RDA-ligated sequences were extracted from the NGS output. As a result, we obtained somatic "SVA fingerprints" for the human brains of two male adult donors. We bioinformatically eliminated all annotated SVA portions of the collapsed reads and mapped the dataset of experimentally enriched SVA flanking sequences to the human genome (GRCh38), thus obtaining the coordinates of SVA insertions. In addition to the experimental reduction of germline SVAs by the RDA-implemented individual germline background (dermis), we eliminated the germline-transmitted SVAs annotated in hg38 with BED files of the respective coordinates. This resulted in the detection of 748–5,540 *de novo* SVA insertions in brain regions including the cerebellum (Cereb), prefrontal cortex (Pfc), olfactory bulb (Bulb), hippocampus (Hippo), and calcarine fissure (Calca) (Figure 6A). With an overall average of 2,307.4 (SEM = 413.22) *de novo* SVA positions, our findings provide ample evidence of active SVA retrotransposition in the human soma. Moreover, the result of SVA mobility rate in the brain is comparable to estimations of another study that counted 1,350 somatic SVA insertions in samples from the hippocampus and caudate nucleus as obtained by using a transposon capture method (Baillie et al., 2011). Next, we tested whether we were able to efficiently reduce the detection of SVA insertions potentially attributable to germline retrotranspositions by the enrichment procedures we applied experimentally. To this end, we counted putative *de novo* SVA insertions that coalesce deeply in ontogenesis and are therefore shared in all tested brain regions, meaning that they could represent the potentially germline-transmitted background of the SVA landscape. As a result, for the two individuals, we could only pinpoint 96 and 98 SVA deeply coalescing integrations, respectively (Figures 6B, C), accounting for only a small fraction of each individual's SVA landscape. In fact, the majority of *de novo* insertions are traceable to take place on the lineages leading to telencephalon and metencephalon, since most of the examined somatic *de novo* SVA integrations are unique to different brain regions under scrutiny (Figures 6B, C). To count the number of reads supporting a *de novo* SVA insertion in the brain region-specific datasets, we initially collapsed the NGS-output after bioinformatically determining the flanks to datasets of non-redundant unique flanking sequences. When counting the non-redundant flanks specific for every unique *de novo* insertion, approximately 75–79% of *de novo* insertions were read one time (Supplementary Table S2; Supplementary Figure S1).

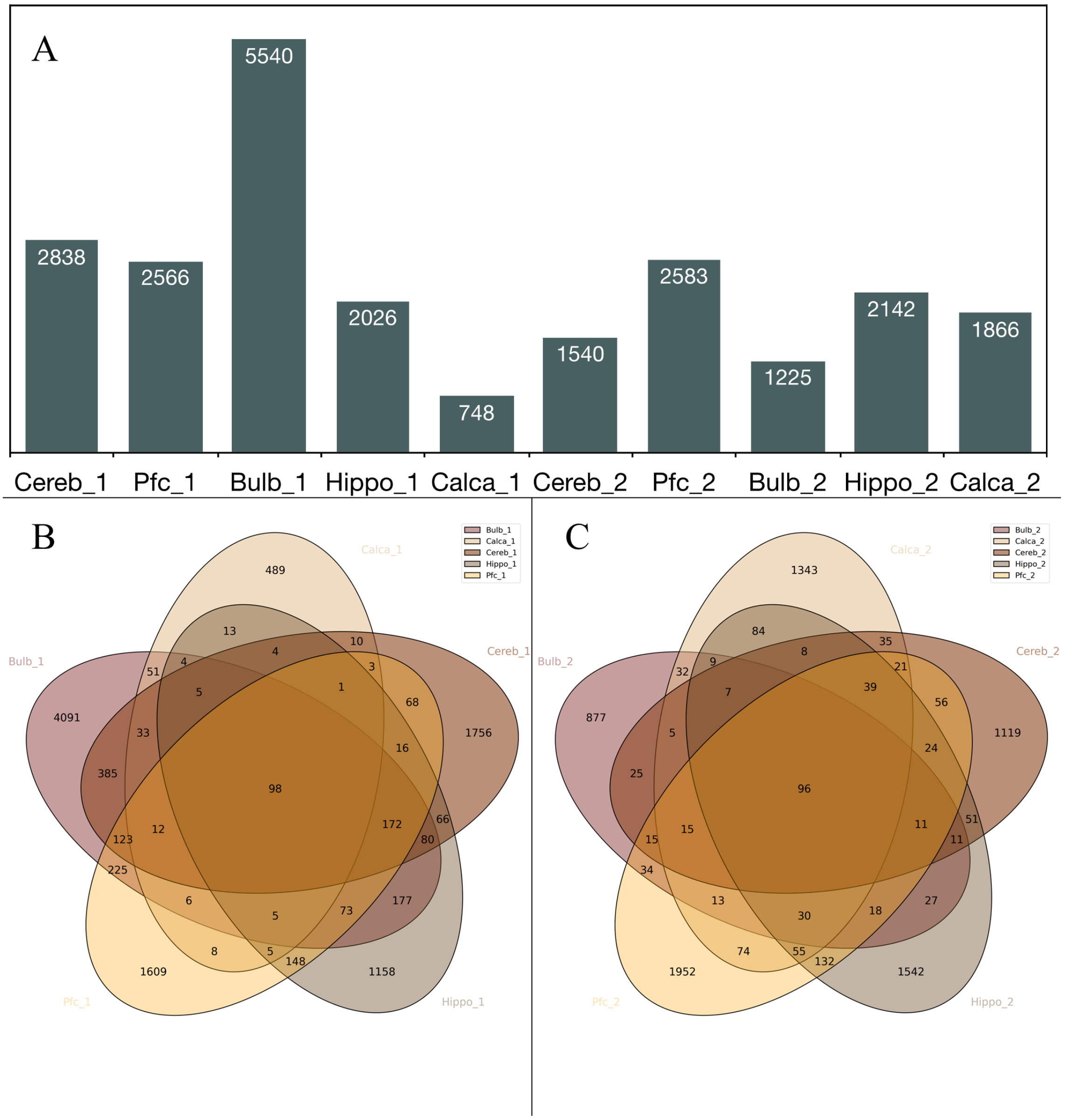


Figure 6: (A) Number of *de novo* SVA integrations for cerebellum (Cereb), prefrontal cortex (Pfc), olfactory bulb (Bulb), hippocampus (Hippo), and the calcarine sulcus (Calca); donors 1 and 2 are denoted as *_1 and *_2. (B,C) Venn diagrams of shared SVA integrations for brain regions depicted for both tested individuals.

3.1.4.2 Somatic SVAs recapitulate general evo-devo patterns

For each individual, all existing *de novo* chromosomal SVA coordinates were combined as the reference file, de-duplicated, and intersected with the SVA *de novo* integration positions of each brain area to generate a presence/absence matrix. We applied an approach in which only overlapping *de novo* SVA positions of multiple brain regions are considered as shared integration, resulting in 1,775 (donor 1) and 918 (donor 2) *de novo* SVA insertions being present in more than one brain region of the tested individuals. Both unique and shared SVA integrations were used to generate a character-based data matrix with the presence/absence markers of SVA as character states to construct a maximum parsimony phylogeny of different telencephalic brain regions (Figure 7). The cerebellar dataset was set as an outgroup because this region is developmentally separate from the telencephalon as part of the metencephalon. In both individuals, the same phylogenetic branching patterns were generated and are well supported by the datasets as mirrored in the bootstrap values (Figures 7B, D). To obtain circumstantial evidence on the number of lineage-specific *de novo* integrations, a phylogenetic reconstruction without resampling was carried out. Altogether 10,888 characters or individual SVA integrations for person 1 and 7,757 for person 2 were analyzed, of which 1,487 and 717 were informative (Supplementary Table S2; Supplementary Figure S2), respectively. Non-informative were all integrations that occurred in one lineage only or are shared between all five brain regions. The respective phylogeny recapitulates the accepted ontogenetic processes of the brain with the longest branches leading to the 'terminal taxa' or brain regions (Figures 7A, C). As previously mentioned, the majority of SVA insertions remain unique to each brain region (Figures 6B, C). At this point, it should be noted that the net length of these edges can be understood as the sum of individual integrations that arose in the bulk of cells, from which DNA was prepared. This argues for extensive somatic SVA mosaicism in the adult human brain, even though in most cases, only a small number of unique integrations occur per cell. Although the internal branches are short, the number of shared integrations that coalesce on a retrotransposition taking place on the lineage leading to the common ancestor of the respective brain region-specific cells strongly supports the presented topology. Moreover, this well-supported topology was generated from two independent datasets.

From an anatomical and ontogenic perspective, the phylogenetic trees correspond well to the ontogenic and phylogenic origin of different brain regions. The cerebellum is part of the rhombencephalon, the third brain vesicle. The rhombencephalon divides into the metencephalon, which is the origin of the pons and cerebellum, and the myelencephalon, which is the precursor of the medulla oblongata. All remaining brain structures in the phylogenetic trees derive from the prosencephalon, the first brain vesicle. The prosencephalon divides into the telencephalon and diencephalon. Evolutionarily, the oldest structure of the telencephalon is the rhinencephalon (olfactory brain). The olfactory bulb derives from this ancient part of the cerebral cortex (paleocortex), and thus the distinctive separation may describe the early branching in the phylogenetic tree compared to the hippocampus as part of the archicortex and the prefrontal cortex and area striata as neocortical structures. The archicortex develops earlier than the neocortex and the localization of the area striata, respectively, calcarine sulcus, in the phylogenetic tree might be explained by the high specialization of this cortical region (granular cortex), which is important for the perception of visual stimuli.

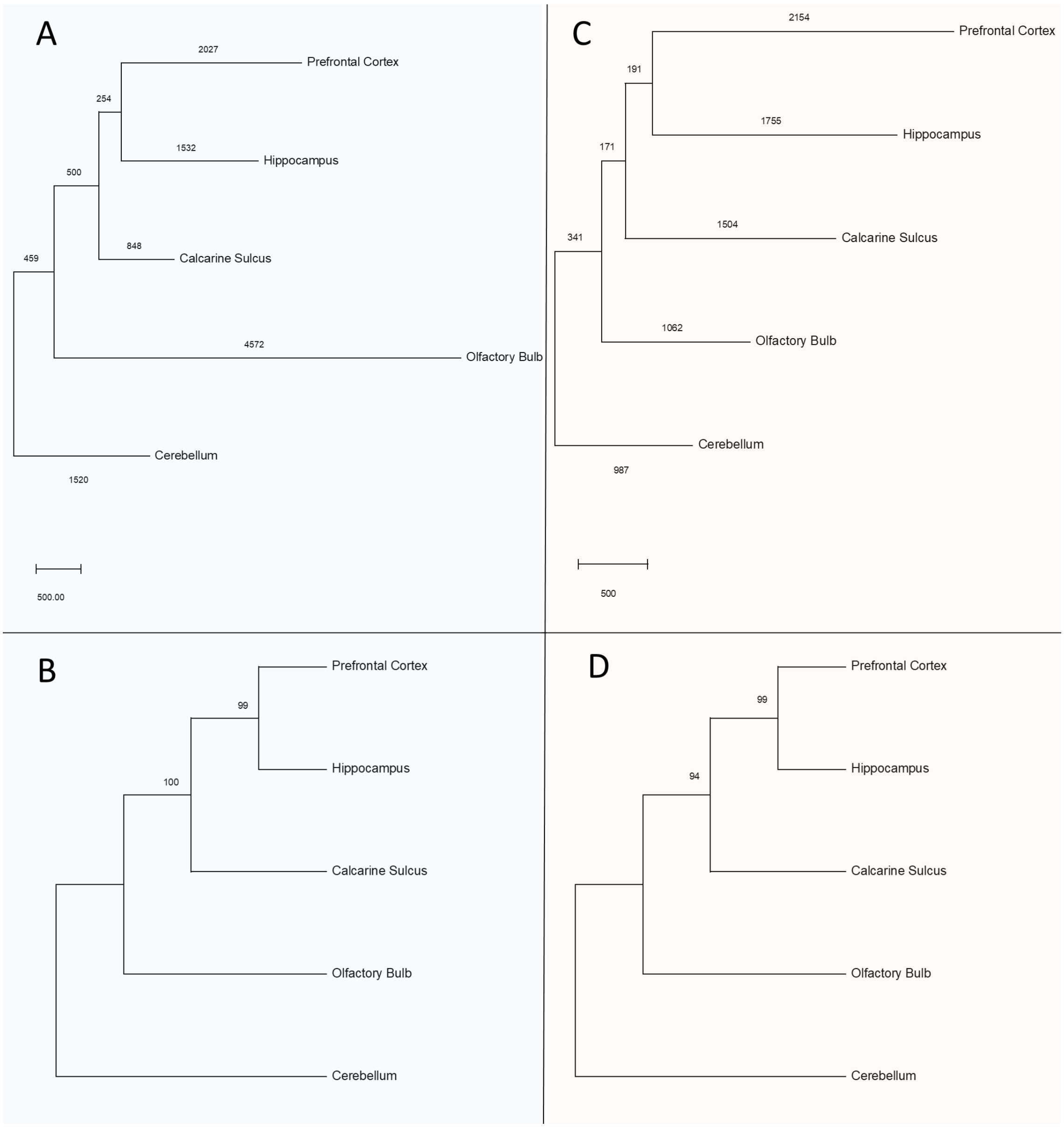


Figure 7: A maximum parsimony tree was constructed to phylogenetically relate the lineages, leading to brain regions based on *de novo* SVA positions. (A) Phylogenetic tree of donor 1 brain regions with values indicating branch length as steps (tree length = 11,710 steps) and (B) bootstrap support values of the phylogenetic tree (1,000 replicates). (C) Phylogenetic tree of donor 2 brain regions with values indicating branch length as steps (tree length = 8,163 steps) and (D) bootstrap support values of the phylogenetic tree (1,000 replicates).

3.1.4.3 Target regions of *de novo* SVA integrations

To compare our ontogenetic data in SVA integration targets with the targets that emerged over evolutionary timescales, we compared the presented *de novo* SVA integration datasets with germline SVA integration coordinates, as depicted in hg38. First, the *de novo* SVA density of each chromosome was estimated by calculating the average genomic SVA length of hg38, which was then multiplied by the SVA counts of each chromosome and finally corrected for chromosome size. The resulting density values are depicted in Figure 8 as the arithmetic mean (n = 2) of SVA bp per million chromosomal bp for each brain region and chromosome. The reference, based on hg38-annotated SVA data, and *de novo* SVA positions in each brain region recapitulate general evolutionary patterns of chromosome-specific SVA density and reveal chromosome-wide rates of *de novo* SVA retrotranspositions. Here, the detected preferences for specific chromosomes, such as Chr. 17 and Chr. 19, are largely consistent with evolutionarily conserved chromosomal SVA patterns.

Concordantly, other reports also found SVA elements to be more frequent than those expected on chromosome 17 and especially chromosome 19, whereas chromosomes 13, 18, and Y exhibited less targets for the reintegration of SVAs (H. Wang et al., 2005; Tang et al., 2018). Taken together for all brain regions under scrutiny, the correlation coefficient r indicates a strong positive correlation of SVA density of reference with all tested brain regions ($p < 0.05$), suggesting similar upward and downward trends of SVA densities on human chromosomes with preferential integration regions.

To quantify a possible enrichment of *de novo* SVA integrations in sites with defined genome features, we applied HOMER software with the *de novo* SVA integration coordinates. In this way, the genomic features of all *de novo* SVA positions were annotated and displayed as fractions of the total annotated features for each brain region (Figure 9). Interestingly, the SVA elements favored integration in genomic positions containing retrotransposon families of LINEs, LTRs, and short interspersed nuclear elements (SINEs), more precisely Alus, as well as intronic and intergenic regions. We checked whether the regions with LINE or SINE association of *de novo* SVA insertions are generally TE-rich regions. To that end, we divided the human genome in 100 kb windows with a 10 kb stagger, extracted all hg38-annotated LINE and SINE positions, and counted the occurrences, i.e., the sum of SINEs and LINEs, for each window. The average count of retrotransposition events within the windows was set as the normal density of TEs, and all windows above average were extracted as TE-rich windows. In addition and for more stringent analysis, we used only the top 25% TE-enriched windows (highest LINE/SINE count windows). We then intersected the two datasets containing the window coordinates with *de novo* SVA integration positions that are associated with SINE or LINE sequences according to the HOMER annotation; 72.83–92.23% of LINE- or SINE-associated *de novo* SVA integrations are located in 100kb windows with higher SINE/LINE count than that in average 100kb windows (Table 5). When only the top 25% TE-rich 100 kb windows are considered, 51.25–62.08% of *de novo* SVAs are still located within the TE-enriched region. In conclusion, *de novo* SVA integrations tend to fall in regions with high count of both SINE and LINE families, which are enriched together. Therefore, *de novo* SVA integration is apparently preferred in regions where previous retrotranspositions occurred, such as HOMER-annotated LINE-2 families (Supplementary Table S2; Supplementary Figure S3), that are mostly truncated remnants mobilized before the mammalian radiation, or LINE-1 as the only remaining autonomous mobilizing element in humans (X. Zhang et al., 2020). In addition to integration near retrotransposon sequences, *de novo* SVA insertions associated with intronic and intergenic regions show that integrations preferentially occur at genomic loci where disruption of gene integrity is less likely.

From chromosome-specific SVA integration target densities and the genome feature analysis, we conclude that reintegrations of SVA target sites with similar characteristics, both in the germline and in the ectodermal brain cells. Therefore, it is reasonable to assume that this causes comparable regulatory consequences, suggesting similar retrotransposition modes being effective in both the germline and soma.

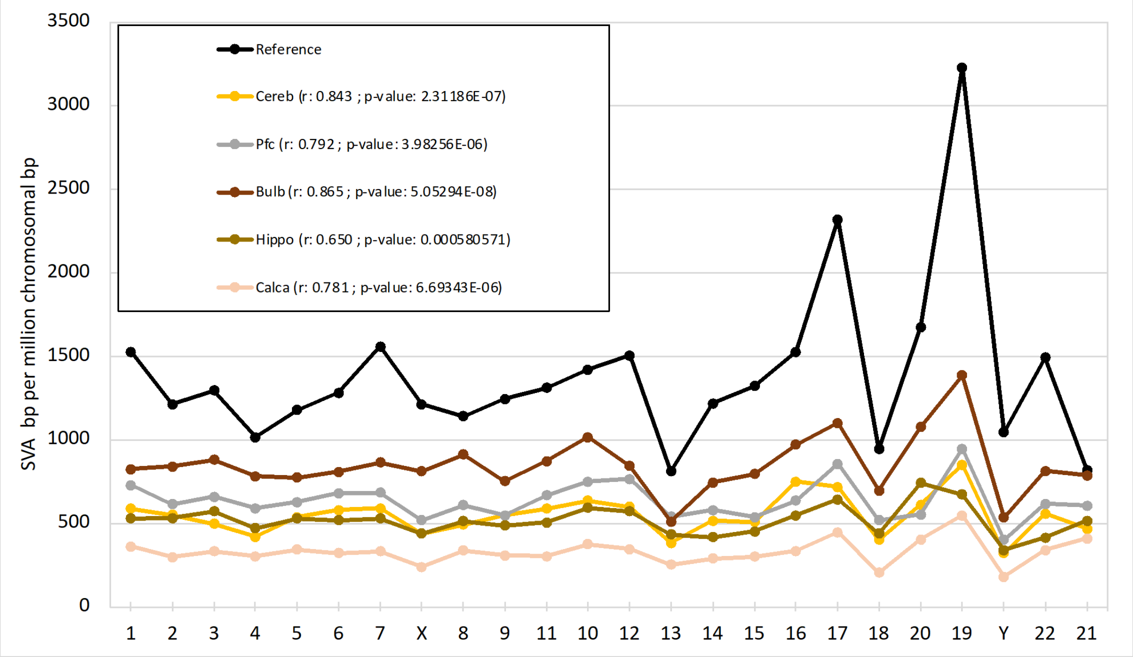


Figure 8: Chromosomal SVA density as SVA bp per million chromosomal bp on y-axis for reference SVAs (hg38) and *de novo* SVA integrations of each tested brain region (density values are calculated as mean values, n = 2). Human chromosomes are listed on x-axis. R-values and significance, depicted as *p*-values, indicate correlation of SVA density of reference with brain samples.

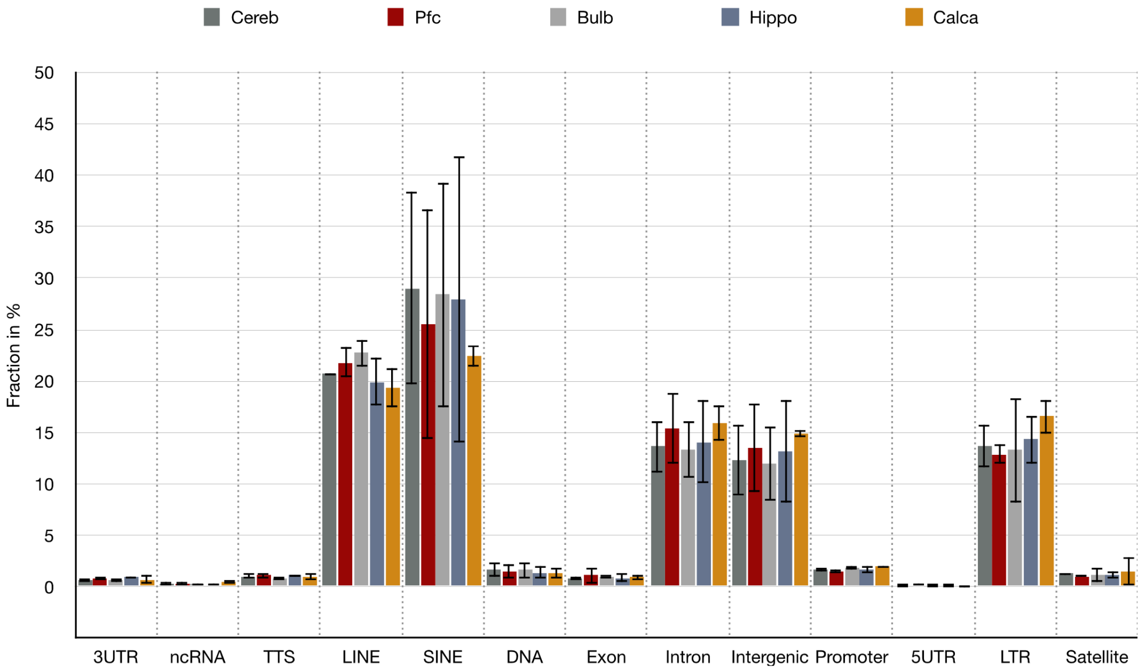


Figure 9: Annotation of genomic features of *de novo* SVA integrations for each brain region. Mean values (n = 2) of fractions of annotated features with respect to sum of all HOMER-annotated features are displayed in % with the standard deviation shown as bars. Features are: 3′-untranslated region (3′-UTR), non-coding RNA (ncRNA), transcription termination site from −100 bp to +1 kbp (TTS), LINE transposons (LINE), SINE transposons (SINE), DNA transposons (DNA), exonic region (Exon), intronic region (Intron), intergenic region (Intergenic), promoter-TSS from −1 kbp to +100 bp (Promoter), 5′-untranslated region (5′-UTR), long terminal repeats (LTR), and satellite region (Satellite).

Table 5: *De novo* SVA integrations associated with retrotransposon-rich regions.

Sample (donor 1 = *_1; donor 2 = *_2)	% of LINE/SINE associated *de novo* SVA insertions in TE-rich 100 kb windows (average TE count as the threshold)	% of LINE/SINE associated *de novo* SVA insertions in TE-rich 100 kb windows (top 25% TE-enriched windows)
Cerebellum_1	78.36	54.83
Prefrontal_cortex_1	88.24	62.08
Olfactory_bulb_1	80.11	53.35
Hippocampus_1	79.70	57.03
Calcarine_sulcus_1	86.71	58.86
Cerebellum_2	88.57	55.64
Prefrontal_cortex_2	90.16	52.19
Olfactory_bulb_2	72.83	51.25

Hippocampus_2	92.23	61.71
Calcarine_sulcus_2	82.05	57.35

3.1.4.4 *De novo* SVA integrations frequently locate in close proximity to neural-specific genes

To elucidate the functional consequences of *de novo* integrations, we examined the genes that could be physically linked to the gene body features extracted from HOMER analysis. To this end, we first removed the intergenic regions that were physically close to *de novo* SVA integrations. We then focused on the related gene names of annotated genomic features linked to *de novo* integrated SVAs like promoters, introns, exons, and 5′-and 3′-UTRs (Figure 9). The gene names were subsequently analyzed for their association using Gene Ontology analysis (Supplementary Table S3). We found that genes physically close to certain somatic SVA integrations fall into neural-specific Gene Ontology pathways and the most significant pathways are linked to regulation of synapse structure or activity (GO:0050803), synaptic signaling (GO:0099536), neuronal system (R-HSA-112316), behavior (GO: 0007610), nervous system development (R-HSA-9675108), and neuron projection morphogenesis (GO:0048812). This supports the assumption that SVA retrotranspositions, related to genes of neural-specific pathways, represent events within the lineages of the somatic brain, where neural-specific gene activity is linked to active chromatin states and provides conditions for active retrotransposition. Similar prevalent conditions, where SVA insertions are favored in active chromatin with genic regions, were also reported by Savage et al. (2013).

Since gene density correlates with GC density (Lander et al., 2001; Versteeg et al., 2003) and SVAs can insert in proximity to genes (Figure 9), we determined whether SVA *de novo* insertions are established in GC-rich regions as detected by Raiz et al. (2012) in 5-kb- and 30-kb-long SVA flanking regions. To this end, we calculated the average GC content of 5-kb extended SVA flanks for each brain region (Table 6) and compared the values to the average GC content of the genome hg38. With an average of 42.82% GC in all tested brain samples, the GC content was tested to be significantly different ($p < 0.05$) from the 40.9% genome average (Piovesan et al., 2019). Consequently, with an approximately 2% increase in the GC content, *de novo* SVA insertions tend to prefer GC-rich regions over AT-rich regions.

Table 6: Average GC content of all SVA flanks (5 kb extended) for each tested sample.

Sample (donor 1 = *_1; donor 2 = *_2)	Average GC content of all SVA flanks (%)	*p*-value (testing H_0: GC content of SVA flanks and genome hg38 are consistent)
Cerebellum_1	42.94	6.85205E-73
Prefrontal_cortex_1	43.06	2.4634E-72
Olfactory_bulb_1	42.79	1.3077E-126
Hippocampus_1	42.89	2.73278E-50
Calcarine_sulcus_1	42.99	6.44965E-22
Cerebellum_2	42.82	7.9799E-35
Prefrontal_cortex_2	42.24	3.57767E-29
Olfactory_bulb_2	42.80	2.46387E-29
Hippocampus_2	42.51	2.02962E-35
Calcarine_sulcus_2	43.18	1.12244E-55

3.1.5 Discussion

The proposed method of RDA-implemented enrichment of *de novo* SVA insertions provides further evidence of active SVA retrotransposition in the human brain. We were able to detect 748 somatic SVA insertions in the calcarine sulcus to 5,540 in the olfactory bulb. Based on the primer system adopted from the ME-Scan-SVA method (Ha et al., 2016) and the authors' estimation of the fractions of different SVA families that could be amplified with these primers, we believe that our results are composed of conservative estimations. The quantitative estimations of *de novo* SVAs in our enrichment method are in the same range as those proposed by applying other capture methods, for example, demonstrated by Baillie et al. (2011). In contrast to other methods, RDA focuses on rare genomic changes and utilizes informative clade markers with distinct character polarity as indicators of the frequency of independent insertions. With the RDA-implemented technical reduction of the individual SVA background (driver = same individual dermis) and by excluding hg38-annotated SVAs as a bioinformatical reduction of germline-transmitted SVAs, we were able to decrease the detection of potential *de novo* SVA insertions attributable to retrotranspositions in the germline or outside the brain during early embryogenesis. The result is that only 96 and 98 SVA positions occur in all tested brain regions, thus representing only a small portion of each person's SVAs.

Additionally, we report 1,775 (donor 1) and 918 (donor 2) shared *de novo* SVA insertions in more than one brain region of the tested individuals, with prefrontal cortex, hippocampus, and calcarine fissure being grouped as regions with the highest similarity as obtained from phylogenetic analysis. Because the brain is among the organs that start to emerge early in prenatal development and among the last to complete postnatal development, genetic alterations such as SVA insertions are difficult to attribute to the developmental timing or progenitor cell population that contribute to the similarity of the aforementioned regions. Nonetheless, the majority of SVA insertions are unique to each brain region and thus can indeed be attributed to brain lineages, confirming the observation of distinct somatic mosaicism in the adult human brain in agreement with Baillie et al. (2011) and Evrony et al. (2015) who demonstrated active SVA, Alu, and L1 retrotransposition in the human brain. We hypothesize that the fact that the observed proportion of unique *de novo* integrations is high, could be explained by their preferential occurrence in many postmitotic neurons; thus the respective *de novo* integrations are not transmitted into progeny cells. In contrast, the smaller proportion of multiple-read *de novo* integrations might occur in mitotic brain cells, e.g., glial cells.

Although we can assign a unique SVA insertion to a specific brain area, such as the prefrontal cortex or hippocampus, our bulk DNA preparation does not allow further assignment to a defined neuronal cell type because we did not use a method for appropriate differentiation, such as cell sorting. Our motivation to start with bulk cell preparations of brain areas to detect *de novo* SVA insertions was based on the assumption that brain neurons and all resident cell types form a functional unit that contributes to a physiologically functional brain. Several brain diseases can be associated with pathological changes in specific cell types, such as interneurons and microglia, and autism, schizophrenia, and Alzheimer’s disease can be associated with changes in all major brain cell types (Skene and Grant, 2016). Thus, the demonstrated somatic mosaicism in the brain may have functional consequences for health and disease, regardless of the cell type.

We examined target region preferences of *de novo* SVA insertions at multiple levels, including the chromosomal location and gene features. First, we found that *de novo* integration preferentially targets transposon-rich regions. We demonstrated that *de novo* SVA insertions occur in regions with high numbers of L1, Alu, and LTRs. When comparing SVAs transmitted across evolutionary timescales, we find a striking similarity. Thus, both germline and somatic brain cells tend to have similar target regions in terms of frequencies of SINE/LINE families as annotated by HOMER analysis. This suggests that similar retrotransposition modes associated with the in trans effects of the autonomous mobilizing LINE-1 are operative in both the germline and soma. In addition, our *de novo* SVA density data suggest similarities with evolutionarily conserved SVA patterns, with chromosomes 17 and 19 showing higher SVA frequencies and chromosomes 13, 18, and Y showing lower SVA frequencies, comparable to our hg38 reference data and the reports of Tang et al. (2018) and Wang et al. (2005). Chromosome 19 appears to be particularly notable in terms of high SVA integration rates, thus confirming previous data from Grimwood et al. (2004) who described chromosome 19 as a chromosome with both high transposon content and gene density. Overall, SVA retrotransposition is thought to occur preferentially in gene-rich and active chromatin regions, as observed by Savage et al. (2013), reflecting the situation in the germline and providing ample opportunities to fine-tune gene expression patterns. Barnada et al. (2022) also reported that the epigenetic repression of active SVAs results in differential gene expression of genes near SVAs. Based on our HOMER results, we also detected *de novo* SVA positions near genes, particularly in association with intronic, promoter, and other gene-related regions. In addition, we were able to confirm the results of Barnada et al. (2022) showing the same mode of preferred retrotransposition in close proximity to gene bodies and that a fraction of the genes associated with these SVA positions can be assigned to neural-specific Gene Ontology pathways. Another result of our study shows that the intersection of the same target *de novo* integrations is low in the two individuals studied, and this shared portion could be the cause of probabilistic target region preferences, i.e., the frequency of SINEs/LINEs and neural genes that are more active with an open chromatin state in the human brain (Supplementary Table S2; Supplementary Figure S4). Finally, the GC content within the 5-kb flanking regions of *de novo* SVA insertions was higher than the average of the human genome, suggesting that SVAs in general tend to insert in genic and GC-rich regions, besides TE-rich regions, in agreement with the results of Raiz et al. (2012) and Wang et al. (2005).

To summarize, our data on somatic SVA mosaicism in the brain demonstrate the mobility of a class of retrotransposons that is highly mobile in the human germline, too. Moreover, there is a striking similarity of retrotransposition modes between the germline and soma, as suggested by similar target regions and gene regulatory potential. We hypothesize that transcribed brain genes trigger chromatin states to be amenable for retrotransposition, as suggested by the correlation of physical distances between brain gene loci as uncovered by GO analysis and somatic SVA integrations. Therefore, somatic mosaicism of SVAs in the human brain is of particular interest, since brain disorders such as Parkinson’s disease can be associated with the presence or absence of SVAs at orthologous loci, along with altered gene expression (Pfaff et al., 2021). We were able to obtain data on the level of multilocus SVA mobility in all tested brain areas, resulting in many lineage specific *de novo* SVA insertions that are frequently associated with genes in close proximity and thus possibly associated with differential gene expression as described by Pfaff et al. Moreover, we described *de novo* SVA insertions that take place at earlier stages of brain development in cells that are still mitotic, giving rise to cell lineages phylogenetically linked to the presence/absence of SVA clade markers. The temporally and spatially ubiquitous *de novo* SVA integrations in the brain could be used as clade markers to study the origin and evolution of brain tumors, that is, to reconstruct intratumor heterogeneity and the tumor cell lineages' phylogeny. The proposed RDA–NGS method to define *de novo* SVA integrations is able to detect unique SVA integrations in tester as compared to driver genomes. Thus, the mutation catalog of a brain tumor can be supplemented with the non-reversible presence/ absence of SVA markers at orthologous loci with that - besides obtaining information on tumor heterogeneity - tissue-specific tumor origin, lineages of cell populations harboring cancer promoting mutations, or primary sites of metastasis could be pinpointed. Furthermore, since there are some limitations in defining and naming cell types with dynamic markers, Domcke and Shendure (2023) recommended establishing a data-driven 'consensus ontogeny' to differentiate cell lineages, e.g., in fetal hematopoiesis or intra- and inter-individual variations. As part of an attempt to order cells based on differences in molecular states and lineage history in a tree-based approach, SVAs as stable clade markers, together with their definition by RDA, could provide an additional tool in the field of cell lineage tracing.

3.1.6 Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: NCBI BioProject under PRJNA949405.

3.1.7 Ethics statement

The studies involving human participants were reviewed and approved by Landesärztekammer Rheinland-Pfalz, Deutschhausplatz 3, Mainz, ethik-kommission@laek-rlp.de (24/05/2022; Ref.# 2022-16488). The patients/participants provided their written informed consent to participate in this study.

3.1.8 Author contributions

Conceptualization: **JM** and HZ; methodology: **JM**, HZ, and VW; tissue preparation: SS; bioinformatics analysis: **JM**; data interpretation: **JM**, MS, and HZ; anatomical data interpretation: SS; figures, tables, and graphics: **JM** and MS; manuscript writing: **JM** and HZ. All authors contributed to the article and approved the submitted version.

3.1.9 Funding

This work was financially supported by the Johannes Gutenberg University of Mainz (Inneruniversitäre Forschungsförderung Stufe I “Differentielle SVA-Retroposonmuster als Indikatoren somatischer Genommosaike des Menschen” to HZ).

3.1.10 Acknowledgments

This publication is based on data collected as part of JM’s dissertation project at the Department of Biology, Johannes Gutenberg University, Mainz, Germany. The authors sincerely thank those who donated their bodies to science so that the present research could be performed, whose results can potentially increase mankind’s overall knowledge. Therefore, these donors and their families deserve our highest gratitude.

3.1.11 Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

3.1.12 Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

3.1.13 Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2023.1201258/full#supplementary-material

3.2 Subchapter: LINE-1 targeted RDA in the human brain

The applied SVA-RDA experiments offer evidence of somatic reintegration of non-autonomous SVAs, which hijack the LINE-1 machinery in trans. Previous studies by Baillie (2011) and Evrony (2015) demonstrated the activity of LINE-1 in the human brain.

To follow up on these results, I am utilizing LINE-1 as a marker to assess the suitability of the RDA for enriching somatic L1 integrations and to investigate the presence of L1 mosaicism in the brain, similar to the demonstrated SVA experiments. Moreover, 5′-truncations of reintegrated LINE-1 are frequently occurring, thus the RDA-method evaluates the existence of 5´-intact integrations. To that end, I have implemented an outward primer system that specifically targets 5′-UTR intact L1s. These intact L1s contain an open reading frame (ORF0) that potentially enhances L1 mobility and consequently may contribute to the introduction of somatic mosaicism.

By utilizing this approach, the dynamics of LINE-1 retrotransposition in the brain and its potential role in generating somatic mosaicism can be evaluated.

3.2.1 Results

The RDA method was applied to enrich unique 5´ genomic flanks of LINE-1 in adult human brain regions (tester samples) compared to dermal tissue from the same individual (driver sample). This experiment aims to identify somatic mosaicism of L1, similar to the approach employed for detecting *de novo* SVA integrations (chapter 3.1). L1 5´-flanking regions are amplified with a consensus outward primer system targeting the intact 5´-UTR of L1, precisely an ORF0 consensus sequence as well as a sequence upstream in the untranslated region. The driver sample was introduced as L1 germline background and is provided in a 100-fold molar excess to the RDA-adapter ligated tester during hybridization, resulting in the enrichment of unique L1 integrations of tester samples by RDA-PCR. Following NGS of the PCR products, the *de novo* LINE-1 flank coordinates of the tested brain regions are extracted and hg38 annotated background is removed. The *de novo* L1 integration positions of one sample are extended to a broader chromosomal position when they overlap with each other, thus shared chromosomal positions are ‘collapsed’ to represent a single integration event.

Overall, this results in the identification of 7 to 174 *de novo* LINE-1 integrations in tested regions like cerebellum, prefrontal cortex, olfactory bulb, hippocampus and calcarine sulcus (Figure 10A). After identifying *de novo* retrotranspositional L1 events within the datasets, I tested whether the introduced RDA-method efficiently reduces the individuals' L1 germline background. Similar to the results presented by SVA-RDA (chapter 3.1.4.1, Figure 6), the germline background is reduced because only one insertion is shared in all tested brain regions as potential retrotransposition in early ontogenesis outside the brain (Figure 10B). In fact, most of the detected retrotranspositions are unique to the respective brain area under scrutiny.

Next, I counted for every brain dataset the flank-read number supporting each *de novo* L1 transposition event. To that end, non-redundant flanks that can be assigned to an unique L1 insertion event are counted for each sample. 78.1 – 87.5% of *de novo* L1 insertions are comprised of one non-redundant flank (Figure 10C). Interestingly, several non-redundant flanks, ranging from 15 in prefrontal cortex to 197 flank sequences in cerebellum, can be assigned to the *de novo* integration position, which is common in all brain regions (extended from all positions to chrX:46277981-46278356).

The *de novo* L1 integrations are counted for each human chromosome (hg38) and normalized based on the chromosome size, resulting in values describing the number of *de novo* L1 integrations per million chromosomal bp for each chromosome and sample (Figures 10D & 11). Favoring of L1 integration - across the majority of brain regions - can be observed for chromosome 10, 17, 19 and 22. In addition, there can be individual preferences for certain chromosomes, for example Chr. 9 in olfactory bulb or 4 in cerebellum, that differ when comparing all brain regions.

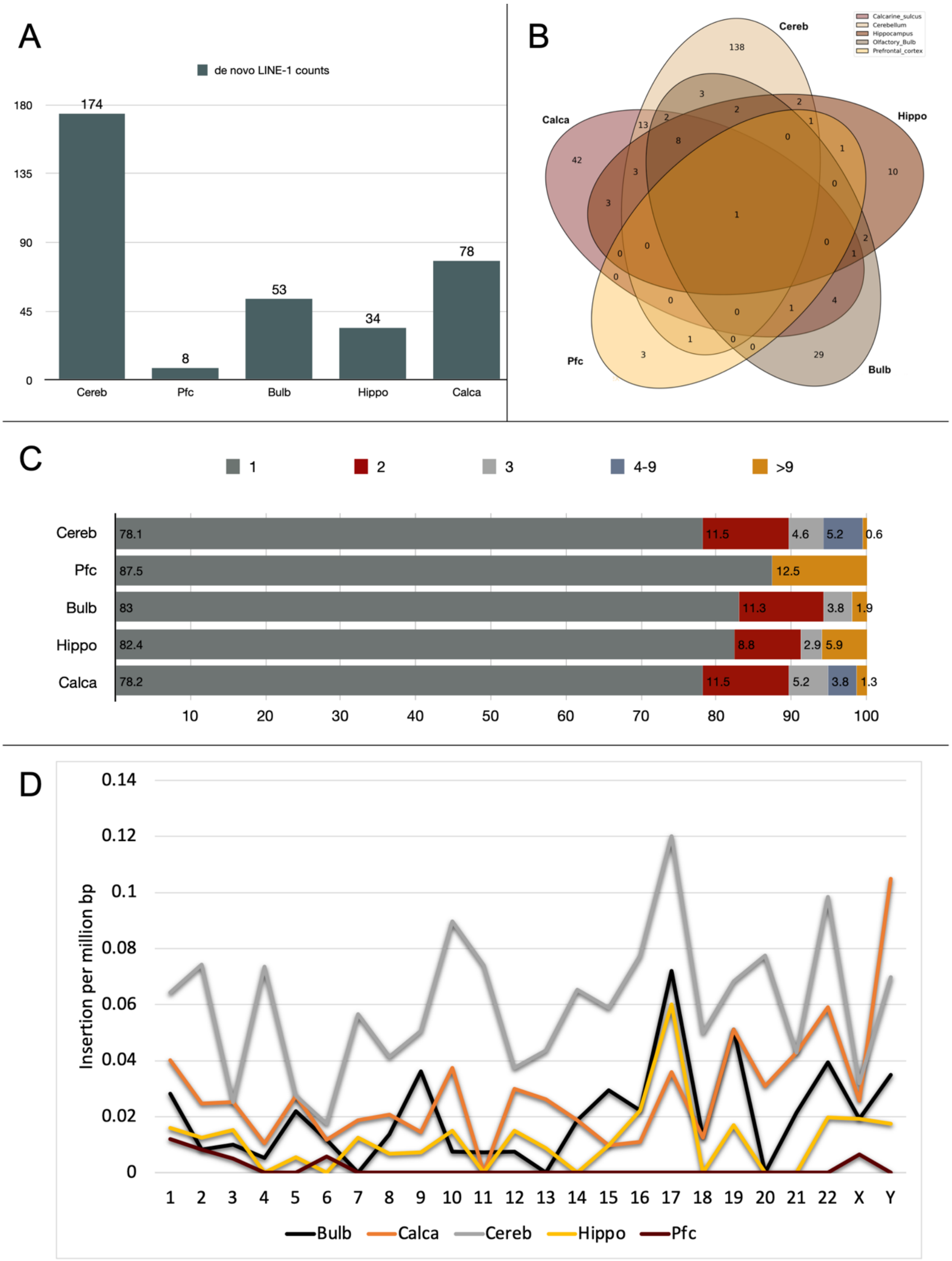


Figure 10: (A) Number of *de novo* L1 integrations for cerebellum (Cereb), prefrontal cortex (Pfc), olfactory bulb (Bulb), hippocampus (Hippo) and calcarine sulcus (Calca). (B) Venn diagram of shared *de novo* L1 integrations between the tested brain regions. (C) Number of non-redundant flanks assigned to each unique L1 insertion event. X-axis shows percent of L1 integrations represented by 1, 2, 3, 4-9 or > 9 flank counts. (D) Number of *de novo* L1 insertions per million chromosomal bp (hg38) on y-axis for each brain region across all chromosomes (x-axis).

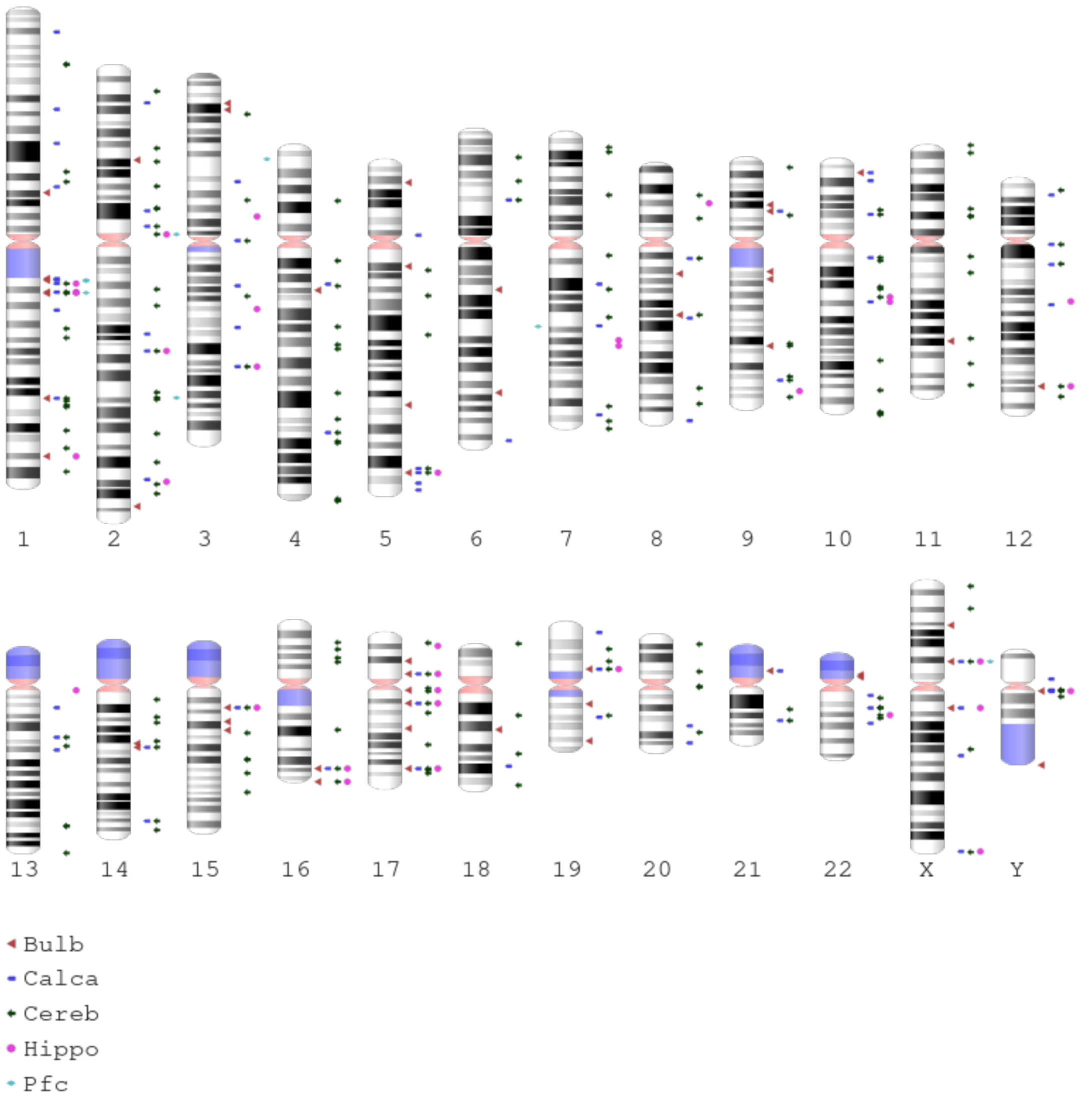


Figure 11: Ideogram (GRCh38.p13) depicting the chromosomal positions of *de novo* L1 integrations for olfactory bulb (red triangle), calcarine sulcus (blue rectangle), cerebellum (green arrow), hippocampus (pink dot) and prefrontal cortex (turquoise rhombus), generated with NCBI's Genome Decoration Page (GDP).

The landscape of chromosome specific *de novo* L1 integrations describes target regions at the chromosome level and can be further refined by analyzing the specific regions where L1 integration preferentially takes place. To that end, the reference genome hg38 is divided into 100 kb windows with 10 kb sliding and the hg38 annotated LINE and SINE positions are counted for each window. The same procedure is implemented for hg38 annotated genes. The 100 kb windows exceeding the average count are extracted as TE- or gene-rich windows. Finally, the coordinates of *de novo* L1 integrations are intersected with the aforementioned windows and counted. On average, 82.02% of *de novo*

L1s integrate in TE-rich regions as well as 55.07% integrate in gene-rich regions. 93.33% of the 55.07% gene-rich L1s are in addition TE-rich (Figures 12A, B).

Since most of the *de novo* L1 integrations are detectable in TE-rich regions that can also be gene-rich regions, I checked whether the integrations are in close proximity to specific genomic features. To that end, the identified *de novo* L1 coordinates are used to define the underlying genomic feature with the HOMER tool. The genomic features of the assigned L1 positions are illustrated as fraction of annotated feature with respect to sum of all HOMER-annotated features in % (Figure 12C). Similar to the presented SVA-RDA results, L1 elements favor integration in genomic regions containing retrotransposon elements like LTRs, SVAs and Alus. Other positions include intergenic regions as well as the gene body, with preferred integration at intronic regions.

Since the HOMER annotated genomic features exhibit an association of *de novo* L1 insertions with other TEs and the gene body - regions known for their high GC content (Lander et al., 2001; Raiz et al., 2012; Versteeg et al., 2003) - the GC content of all *de novo* L1 flank regions with a 5 kb extension is calculated. The GC content of 4 out of 5 datasets is significantly different ($p < 0.05$) respectively higher compared to the average genomic GC content of 40.9% (Piovesan et al., 2019) as depicted in Table 7.

Table 7: Average GC content of all LINE-1 flanks (5kb extended) for each tested sample

Sample	Average GC content of all LINE-1 flanks (%)	p-value (testing H0: GC content of LINE-1 flanks and genome hg38 are consistent)
Cerebellum	43.81	4.11442E-08
Prefrontal cortex	38.52	0.40582653
Olfactory bulb	43.67	0.007708454
Hippocampus	46.46	3.52344E-05
Calcarine sulcus	43.38	0.002516696

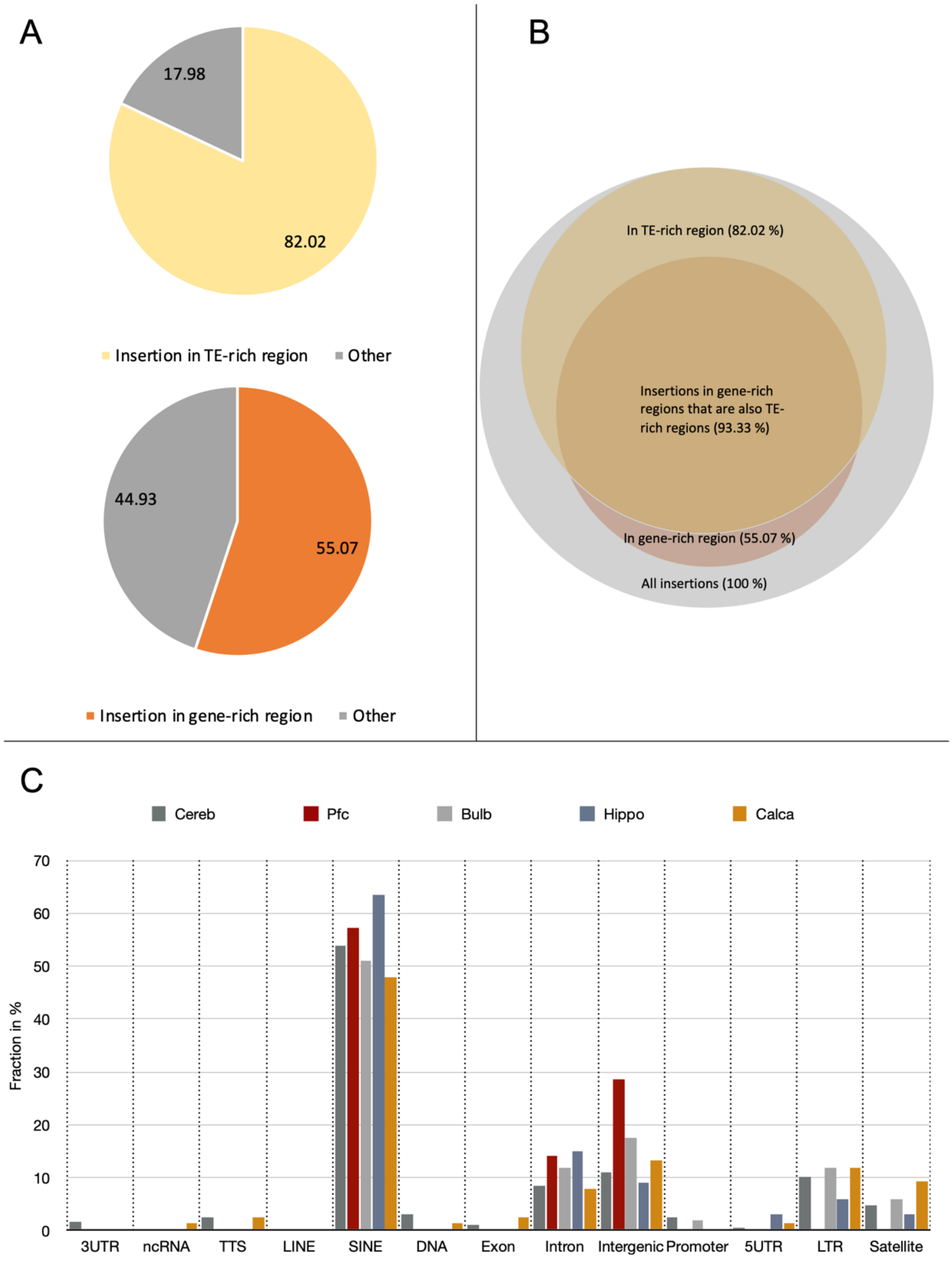


Figure 12: (A) Pie-charts display the percent of *de novo* L1 integrations in 100 kb chromosomal windows that are TE-rich (LINE and SINE) or gene-rich compared to average 100 kb windows, percent values depict the mean of all tested brain regions. (B) Illustration of *de novo* L1-integration targets: 82.02% of *de novo* L1 integrations are within a TE-rich chromosomal site, 55.07% integrated in a gene-rich region, of which 93.33% are in addition a TE-rich region. (C) Annotation of genomic features of *de novo* L1 integrations for each brain region as fractions of annotated feature with respect to sum of all HOMER-annotated features in %. Features are: 3´ untranslated region (3UTR), non-coding RNA (ncRNA), transcription termination site from −100 bp to +1 kbp (TTS),

LINE and SINE transposons, DNA transposons (DNA), Exon, Intron, intergenic region (Intergenic), promoter-TSS from −1 kbp to +100 bp (Promoter), 5´ untranslated region (5UTR), long terminal repeats (LTR), satellite region (Satellite).

3.2.2 Discussion

The RDA is based on an outward primer system and was designed to target the 5´-UTR, specifically ORF0 within the 5´-UTR, and the adjacent 5´-flanking region through PCR. The dermis templates are given in 100-fold excess to the brain templates during hybridization to introduce the germline L1s and early ectodermal background. Hence, amplified L1 templates in brain that originate from germline or early ectoderm, hybridize complementary to the dermis templates and cannot be amplified exponentially in subsequent PCR. This is based on the introduced adapter system, which is only present at both DNA ends in unique tester-tester dsDNA fragments. Consequently, only the brain-specific 5´-untrancated UTRs respectively ORF0 comprising LINE-1 elements and their corresponding flanking sequence are significantly enriched with this method.

3.2.2.1 Detection of *de novo* L1 integrations in the human brain

The bioinformatic analysis resulted in the detection of 7 to 174 *de novo* LINE-1 insertions in the tested human brain regions (Figure 10A) and further provides evidence of active retrotransposition in the human brain in accordance with findings of Zhao (2019), Evrony (2015) and Baillie (2011).

The limited number of detected *de novo* L1 insertions may be explained by the implemented approach, which exclusively targets ORF0 comprising *de novo* integrations but the majority of LINE-1 do not reintegrate as a full-length element. The L1 endonuclease nicks a 5′-TTTT/AA-3′ sequence motif and utilizes the 3′-hydroxyl to initiate the reverse transcription of a L1 mRNA, often resulting in 5´-truncated L1 insertions. In addition to 5´-truncations, inversions or deletions and 3´-truncations are frequently present during reintegration of L1 (Richardson et al., 2014). Therefore, the RDA method with a primer outward system might not be able to present the full scope of L1 insertions. The primary objective behind implementing the L1-RDA was to assess the presence of ORF0 comprising L1 insertions in soma, especially in the somatic human brain, because the human genome contains 780 5´-untrancated L1s with a primate-specific ORF0 (Denli et al., 2015). This ORF is oriented in antisense within the 5´-UTR of LINE-1, contains a region with promoter activity and expresses a capped, polyadenylated ORF0 mRNA. The translation of ORF0 mRNA from an ORF0 integration within human introns was also detected by Denli. As a result, ORF0 of the detected L1 *de novo* integrations - observed to be associated with intronic regions as of HOMER annotation (Figure 12C) - could potentially be transcribed and translated from an intronic position. Since the detected L1 *de novo* integrations contain an intact 5´-region, the antisense promoter activity as well as aforementioned intron associated expressions could have an effect on somatic integrations, because Denli and colleagues reported that overexpression of ORF0 protein in cell models like HEK293T cells and human NPCs has an influence on L1 mobility.

Moreover, the detected L1 *de novo* integrations contribute to the generation of somatic mosaicism within the human brain soma because most of the detected insertions are unique to the respective brain region, with only one position being shared between all tested regions (Figure 10B). The latter supports the effective reduction of germline L1 background by introducing RDA-implemented L1 background (dermis as driver) and bioinformatically eliminating hg38 L1s. 78.1 – 87.5% of *de novo* L1 insertions were read one time when counting the non-redundant flanks specific for each unique *de novo* L1 integration (Figure 10C) and thus may represent retrotranspositional events within postmitotic cells or in late development respectively adolescence. Since these TE integrations might not be transferred to progeny, they remain unique occurrences within a specific brain region and contribute to the somatic mosaicism. In contrast, the *de novo* LINE-1 position, which is detectable in all regions with 15 to 197 assigned non-redundant flanks, may represent an insertion that expanded in progeny and is thus detectable in multiple cells. The various flank counts provide support for the assumption that this LINE-1 integration represents an event in proliferating cells of early embryogenesis, which expanded and is detectable in cells of all tested brain regions.

3.2.2.2 Target regions of *de novo* L1 integrations

Analyzing the *de novo* L1 integration target regions demonstrates a preferences for chromosomes 10, 17, 19 and 22. In a more detailed approach, the data indicate that 82.02% of *de novo* L1 integrations are within a region containing high frequency of LINEs and SINEs combined (Figures 10D & 12), suggesting a preference of L1 reintegration in regions that are gene-rich and especially TE-rich. Based on data of Tang et al. (2018), germline fixed respectively hg38 reference L1s tend to show a homogenous density across all chromosomes. In contrast, SVAs and Alus demonstrate a preference for certain chromosomes, e.g. Chr. 17, 19 and 22 (Grover et al., 2004; H. Wang et al., 2005), similar to what is observed in the presented *de novo* L1 datasets. Since specific chromosomes exhibit a higher TE density and the *de novo* L1 insertions favor TE-rich regions, Chr. 17, 19 and 22 may represent a probabilistic target region for *de novo* integration. Therefore, it can be assumed that similar retrotransposition modes are active and conditions that are especially facilitated by open chromatin states introduce a genome susceptibility to retrotransposition. As a consequence of similar retrotransposition modes, it can be assumed that TE-rich regions - consisting of patterns of evolutionary fixed and *de novo* integrations - are generated.

Engineered L1 elements disclosed the mechanisms and target regions of L1 retrotranspositions in several cell lines and are in accordance with the present HOMER annotated results of *de novo* L1 insertions in the human brain. Flasch et al. (2019) reported that 21-26% of L1 insertions occur in genomic L1s and 6-7% in genomic Alus, indicating that retrotransposon elements are a preferred region of *de novo* integration, which was also demonstrated for SINE containing regions in the present work. The HOMER annotation represents only the close relation to SINEs because mapped L1 flanks located within a distance of 3 kb from a hg38 L1 were marked as germline background and bioinformatically excluded.

Genomic regions containing genes - especially active genes - tend to be associated with an open chromatin state, thus provide conditions for a favored *de novo* integration of retrotransposons. The preference for genic regions was reported by Baillie (2011), Jacob-Hirsch (2018) and Upton (2015), which demonstrated that L1s integrated into exons and predominantly into introns of observed genes. This association with the gene body, meaning favored integration of *de novo* L1s at intronic regions, was also detected by the RDA-method (Figure 12C). The reintegration of TEs at intronic sites - and also the detected intergenic integrations - display regions where disruption of gene integrity is less likely.

Additionally, the presented results show LINE-1 preference for GC-rich over AT-rich regions. The detected *de novo* L1 insertions were identified in GC-rich TE and gene associated regions, meaning regions where gene density correlates with GC density (Lander et al., 2001; Versteeg et al., 2003), explaining the enrichment of GC at LINE-1 flanking regions.

To summarize, the herein presented L1-RDA data discover the presence of active L1 retrotransposition in the human brain with an introduction of a somatic mosaicism due to the clear character polarity of TEs. The detected *de novo* insertions are 5´-untrancated and therefore can potentially express ORF0 as contributor to overall L1 mobility. Based on the L1 target feature analysis it can be concluded that L1 target regions with similar characteristics to polymorphic and fixated germline-transmitted L1 integrations are present in somatic ectodermal brain cells. Therefore, it can be assumed that similar retrotransposition modes are effective in germline and soma.

3.3 Conclusion of NGS-coupled SVA- and L1-RDA

Retrotransposition is increasingly considered a possible cause of neurogenetic disorders and contributes to the human mosaicism with patterns that may be similar to those in germline, based on similar retrotranspositional modes being effective. I asked whether this is reflected in the human brain and used the RDA as subtractive and kinetic enrichment technique, coupled with deep sequencing, to compare different brain regions with respect to *de novo* SVA and L1 insertion-patterns.

Scanning the NGS data resulted in the identification of LINE-1 and SVA activity respectively mobility within the human brain. The *de novo* integrations contribute to somatic mosaicism because most of the detected TE integrations are unique to the brain regions and represent lineage specific *de novo* insertions. Logically, observing active LINE-1 retrotransposition in the human brain implies that the underlying machinery for retrotransposition of the autonomous L1s is present and can be used in trans. Hence, SVAs can 'hijack' the reverse transcriptase and endonuclease activities and contribute to the brain mosaicism. Moreover, both L1 and SVA tend to integrate at similar target regions, suggesting that retrotransposition modes similar to those in the germline are effective. This results in favored integration at TE-rich regions, where previous retrotranspositions occurred and may pose an open chromatin region susceptible to retrotranspositions in general. The *de novo* TE insertions, identified by RDA, can be assigned to a specific brain region but the bulk DNA preparation of the present work does not allow a further differentiation of cell types because methods like cell sorting were not implemented. Brain diseases like autism, schizophrenia and Alzheimer's disease are associated with pathological alterations in all major brain cell types (Skene & Grant, 2016), and thus the brain was assumed to be a functional system, assembled by multiple cell types. This implies that somatic mosaicism may have consequences for health and disease, regardless of the affected cell types.

Focusing on L1s with an intact ORF0 proposes difficulties because insertional events can result in 5′-truncations and the applied RDA-method depends on a L1-primer-outward system. The *de novo* L1 insertions are detectable in low quantity but still generate a mosaicism with most insertion being unique to each tested brain region. However, this was not sufficient to obtain informative sites for constructing the 'phylogenetic' relationships of each brain region because the implemented method is not able to represent the full scope of LINE-1 retrotransposition.

The original intention behind using the RDA method was to reconstruct the phylogeny of cell lineages and explain intra- and inter-individual variations based on the clear character state of TEs and the resulting informative positions. As a result, I conclude that SVAs, which are detectable in high quantity and generate sufficient informative sites, are suited best for the RDA method and yield better results in contrast to LINE-1. Moreover, SVAs are of particular interest because studies of brain disorders report that presence or absence of certain SVAs at orthologous loci, along with changes in gene expression, can be linked to Parkinson’s disease (Pfaff et al., 2021). The presence or absence of *de novo* SVAs at orthologous loci in different lineages, as well as association with genes in close proximity, can be detected by RDA and may help to evaluate a possible link to brain diseases.

Finally, when should the RDA-method be preferred over other proposed methods? The RDA in combination with NGS can detect somatic *de novo* insertions by elevating such events in comparison to the introduced germline-transmitted background. Unlike other methods including TE-capture or WGS methods that estimate frequencies of insertion events, RDA focuses on genomic changes between different areas or on the level of unique cell subpopulations. Thus, utilizes informative clade markers with distinct character polarity as indicators of independent insertions in a somatic context. The motivation for performing RDA is not to estimate frequencies, but to introduce TEs and their presence/absence as clade markers to explain somatic mosaicism in the human brain. Moreover, it provides new opportunities to explain intra- and inter-individual variations and to reconstruct the phylogeny of cell lineages. The field of cell lineage tracing is currently very popular, e.g. Liu et al. (2023) introduced an approach to classify individual stem cells and their close relatives of human brains with protein and RNA analysis. In addition, Domcke & Shendure (2023) recommend to establish a data-driven ‘consensus ontogeny’ to differentiate cell lineages or intra- and inter-individual variations in tree based approaches. SVAs as stable clade markers together with RDA for the definition of a molecular state and lineage history of cells, could provide an additional tool in cell lineage tracing fields in order to sort cells.

4. Double-strand break labeling: breakome & brain mosaic

In contrast to retrotransposon clade markers with undisputable character polarity, DNA double-strand breaks can be dynamic due to DNA repair, requiring an analysis of genomic hotspots to reflect patterns of DNA double-strand breaks and a potential link to mosaicism as precursors for mutational clusters.

Cells are constantly exposed to exogenous and endogenous damage such as radiation, reactive oxygen species, replication stress and transcription, which can induce DSBs. The aberrant repair of damaged DNA leads to mutational events that, together with differences in genomic localization and brain-regional differences, can initiate a somatic mosaicism. By implementing a DSB labeling system based on Breaks Labeling In Situ and Sequencing (BLISS) the ‘breakome’ of each human brain region can be analyzed. This method is able to detect common and region-specific DSB hotspots and associated fragile genes that can be linked to neurological disorders. The presented methods also provide new opportunities respectively targets for predicting DSBs including DNA-binding proteins that are associated with DNA damage pathways.

4.1 Results

4.1.1 DSB statistics and hotspot calling

The NGS data in FASTQ format are bioinformatically scanned for barcode sequences to obtain reads with a DSB site, which is directly marked by the ligated adapter. Next, the reads are mapped to the human reference genome hg38 to retrieve the chromosomal coordinates of all unique DSBs. The term unique refers to the bioinformatic validation of UMIs, which are part of the ligated adapter, thus only reads with an unique identifier at a specific chromosomal position are accepted to eliminate PCR-amplification bias. This is necessary to perform peak calling and retrieve hotspots, where multiple unique DSB events accumulate within a chromosomal region.

Overall, the bioinformatic analysis provides 6,241,600 – 22,695,802 unique DSB events in samples of hippocampus, calcarine sulcus, cerebellum, olfactory bulb and prefrontal cortex.

The human genomic DNA content is circa 6 pg per diploid cell, which means that for an input of 1 µg of genomic DNA in the present experimental procedures (method 2.3), approximately 166,67 diploid cells are used to label newly formed DSB sites.

This translates to an approximate estimation of 37.45 to 136.17 DSBs per cell (Table 8), when dividing the sum of all unique DSBs by total number of input cells. The average of DSBs per cell is 76.60 across all tested samples.

Table 8: DSB numbers and estimations

Sample	Number of unique DSBs	DSBs per cell
Prefrontal cortex	22695802	136.17
Cerebellum	13552555	81.32
Olfactory bulb	15095482	90.57
Hippocampus	6241600	37.45
Calcarine sulcus	6254459	37.53

Next, the unique DSB coordinates, ranging from an average length of 66 – 112 bp in all tested samples (Table 9), are used to call peaks with MACS. This tool detects significant coverage compared to the background and therefore is useful to obtain enriched ChIP regions, binding sites and in general evaluates data that are tested for DNA enrichment. Similar to the bioinformatic analysis in DSBCapture methods of Lensing et al. (2016), significant accumulation of DSB tags can be identified while random positions are discarded using the MACS default for the statistical threshold ($q \leq 0.05$) and the options 'no-model' and 'no control'.

Overall, 423 – 2,538 peaks respectively DSB hotspots can be identified in all tested samples, with prefrontal cortex, cerebellum and hippocampus containing the highest number of DSB hotspots (Table 9). The average length of DSB hotspots ranges from 240.74 – 293.41 bp. When counting the average DSB occurrence per peak, 12.59 – 23.05 DSB events can be assigned to the hotspots. Since a DSB breakpoint can be converted into a single labeled bp position, the DSB density of the peaks is calculated in % by dividing the average DSB count per peak by the average peak length multiplied by 100. In addition - as an internal control - the average hotspot length of each sample is used to chop the hg38 reference genome into windows of the corresponding length and the unique DSB positions are counted for each window. The average DSB density across all hg38 windows with a DSB count greater 0 is calculated as aforementioned. The DSB density of DSB hotspots ranges from 5.19 – 8.16% compared to the genomic average of 0.63 - 1.34%, indicating a fold change greater 6 for the DSB density of hotpots in contrast to the overall distribution (Figure 13).

Table 9: Analysis of MACS peaks (DSB hotspots)

Sample	Pfc	Cereb	Bulb	Hippo	Calca
Average DSB read length	112	89	98	74	69
Peak count	2538	1178	423	1782	1003
Average peak length	282.59	258.13	293.41	242.61	240.74
Average DSB count per peak	23.05	17.59	22.99	12.59	12.99

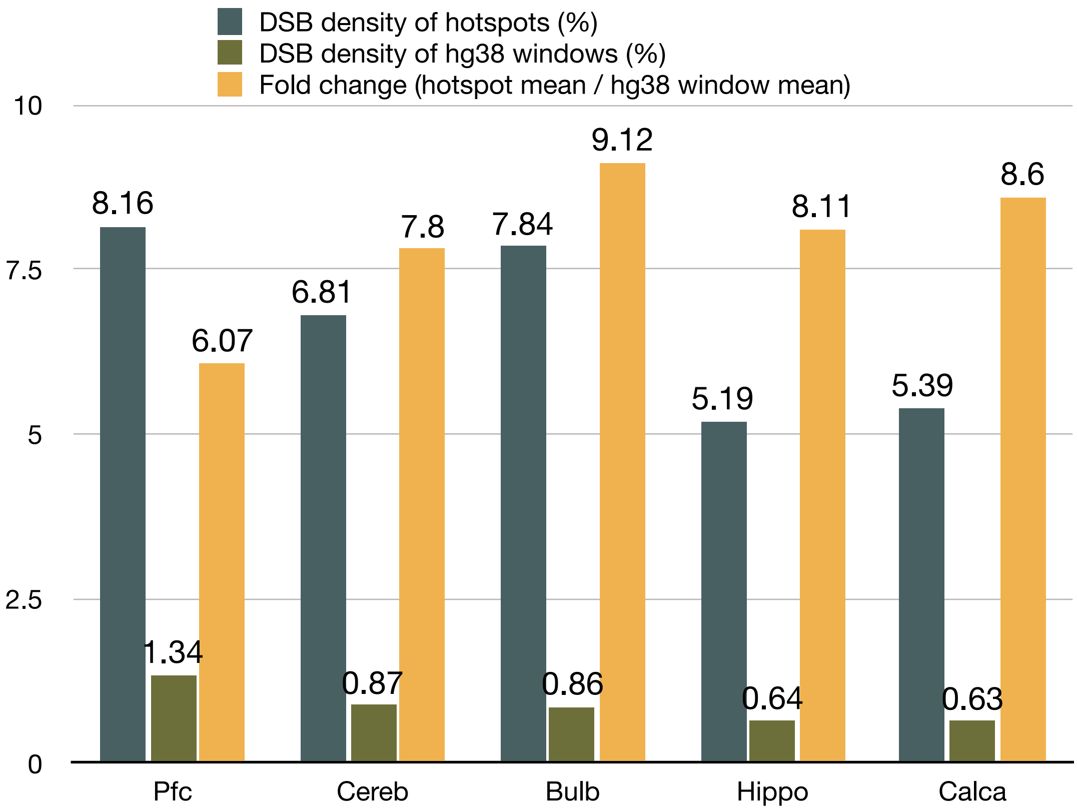


Figure 13: Depicted are the average DSB densities of MACS-generated DSB hotspots and average DSB density across hg38 for each tested sample in %. In addition, the fold change values (DSB density of hotspots divided by the average hg38 density) are provided.

4.1.2 Chromosomal distribution of DSB hotspots

The chromosomal distribution of DSB hotspots is demonstrated by counting the occurrences of hotspots per human chromosome. Next, I evaluated whether the observed occurrence is greater than the expected occurrence of randomly generated DSB hotspot positions to take the chromosome size into account. To that end, the hotspot coordinates of each sample are used to randomly shuffle new positions using BEDTools on hg38, thus generating 3 random datasets. The average occurrence (n = 3) of random positions is set as the expected value for each chromosome and the fold change is calculated by dividing the observed chromosome counts by the expected mean count. The majority of chromosomes, e.g. Chr. 1 to Chr. 12 and Chr. 16 contain approximately the same number of DSB hotspots as expected (Figure 14). Chr. Y and especially Chr. 13 contain less hotspots than the expected number. The number of DSB hotspots on Chr. 17 is enriched compared to the expected value and certain chromosomes like Chr. 18 - Chr. 22 show individual enrichment for specific brain areas.

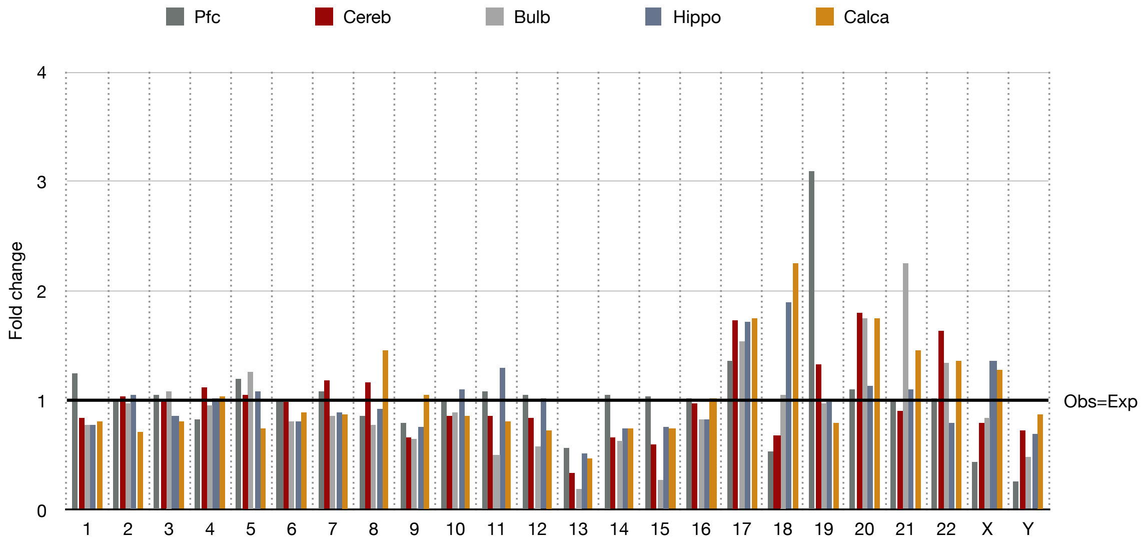


Figure 14: Chromosomal distribution of DSB hotspots depicted as foldchange of observed DSB hotspots divided by expected occurrences of DSB hotspots for each chromosome (x-axis) and sample.

4.1.3 Feature analysis of DSB hotspots

The DSB hotspot coordinates are used to annotate the underlying genomic feature with HOMER, thus I can evaluate which regions of the genome are prone to double-strand breaks.

The annotated features are depicted as fractions of the total annotated features for each brain region and demonstrate that DSBs accumulate at regions associated with retrotransposons like LINE, SINE and LTR, gene body features like TTS, promoter and intron, intergenic regions and satellite regions (Figure 15).

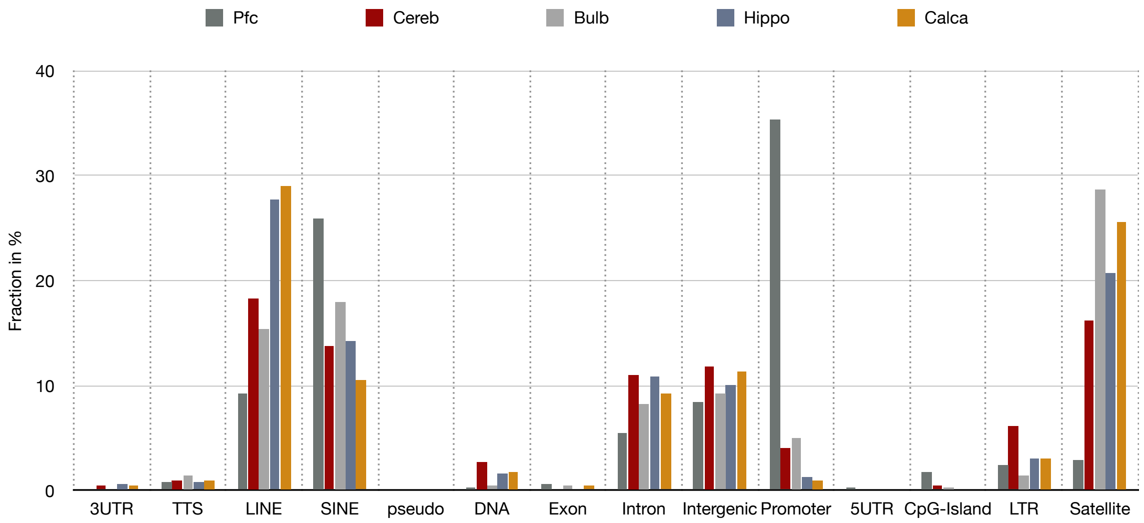


Figure 15: Annotation of genomic features of all DSB hotspots for each brain region as fractions of annotated feature with respect to sum of all HOMER-annotated features in %. Features are: 3´ untranslated region (3UTR), transcription termination site from −100 bp to +1 kbp (TTS), LINE and SINE transposons, pseudogene (pseudo), DNA transposons (DNA), Exon, Intron, intergenic region (Intergenic), promoter-TSS from −1 kbp to +100 bp (Promoter), 5´ untranslated region (5UTR), CpG-Island, long terminal repeats (LTR), satellite region (Satellite).

When focusing on the DSB hotspots located at satellite regions, the annotation of (GAATG)n, BSR/Beta, SAR and HSATI satellites can be observed. In addition, the majority of satellite related DSB hotspots are located in alpha-satellites of centromeric regions (Table 10).

Table 10: DSB hotspots located at satellite regions

Sample	Number of DSB hotspots at satellite region	Number of satellites assigned to centromeric alpha-satellite
Pfc	74	46
Cereb	191	154
Bulb	121	95
Hippo	370	335
Calca	257	231

Next, the genes of the annotated gene body features such as UTR, TTS, intron, exon and promoter, are extracted to analyze which genes are affected by the DSB accumulation and whether these gene sets are related to GO pathways. To that end, the gene datasets are analyzed with Metascape (Figures 16 – 20). Interestingly, the majority of enriched pathways, and thus DSB affected genes, are related to neural pathways like ‘protein localization to synapse’ (GO:0035418), ‘synaptic signaling’ (GO:0099536), ‘retrograde axonal transport’ (GO:0008090), ‘axo-dendritic transport’ (GO:0008088) etc. but are also related to transcriptional activity like ‘Metabolism of RNA’ (R-HSA-8953854) and ‘Ribosome’ (hsa03010).

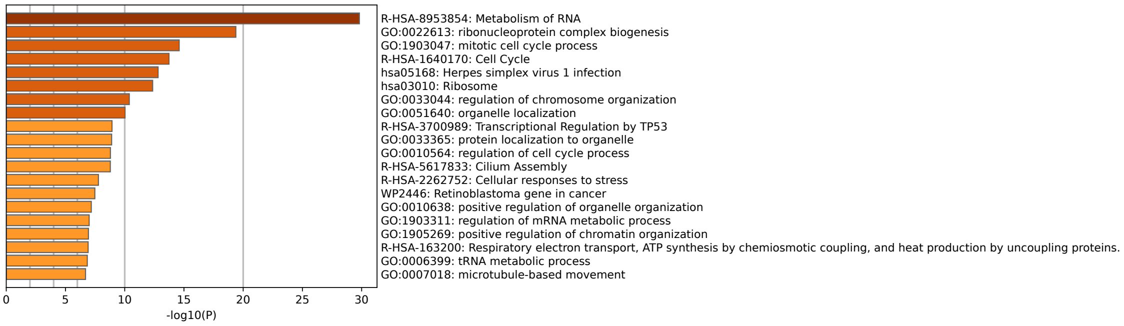


Figure 16: Metascape gene ontology pathways of genes related to DSB hotspots in prefrontal cortex.

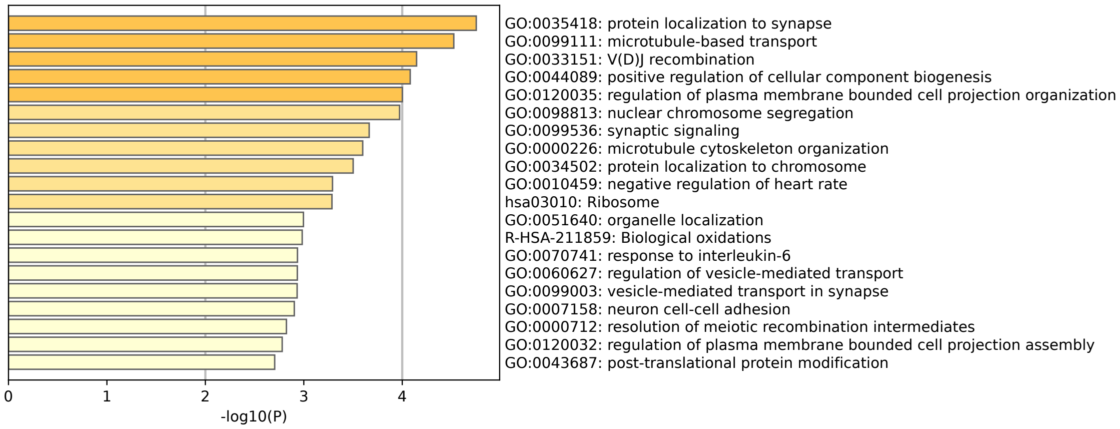


Figure 17: Metascape gene ontology pathways of genes related to DSB hotspots in cerebellum.

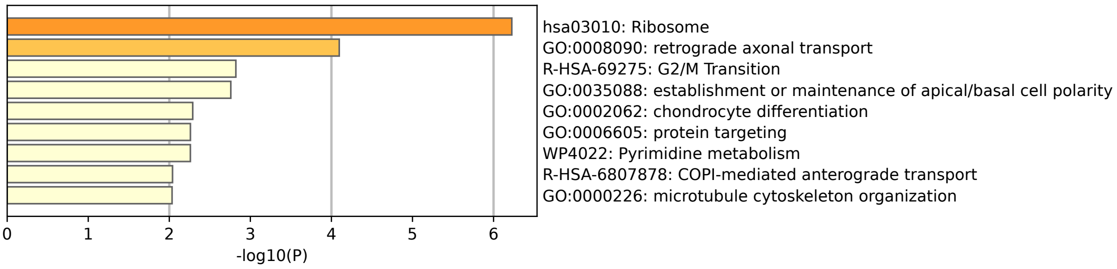


Figure 18: Metascape gene ontology pathways of genes related to DSB hotspots in olfactory bulb.

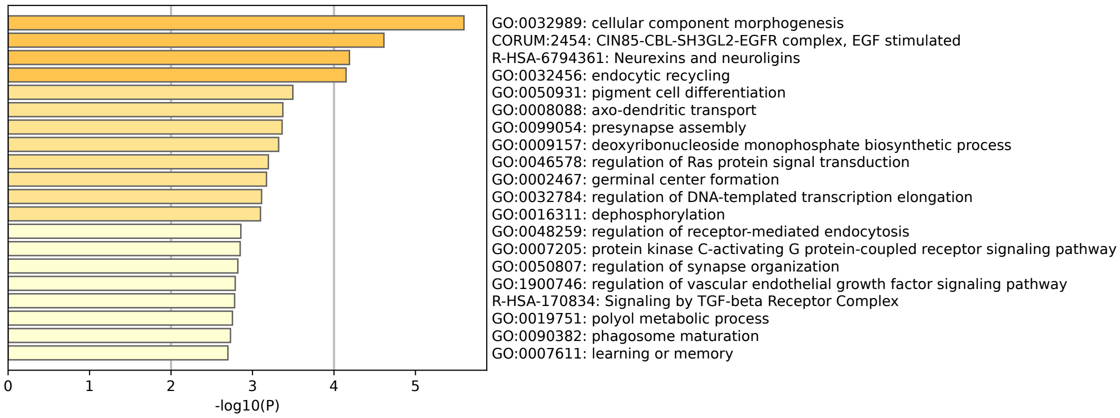


Figure 19: Metascape gene ontology pathways of genes related to DSB hotspots in hippocampus.

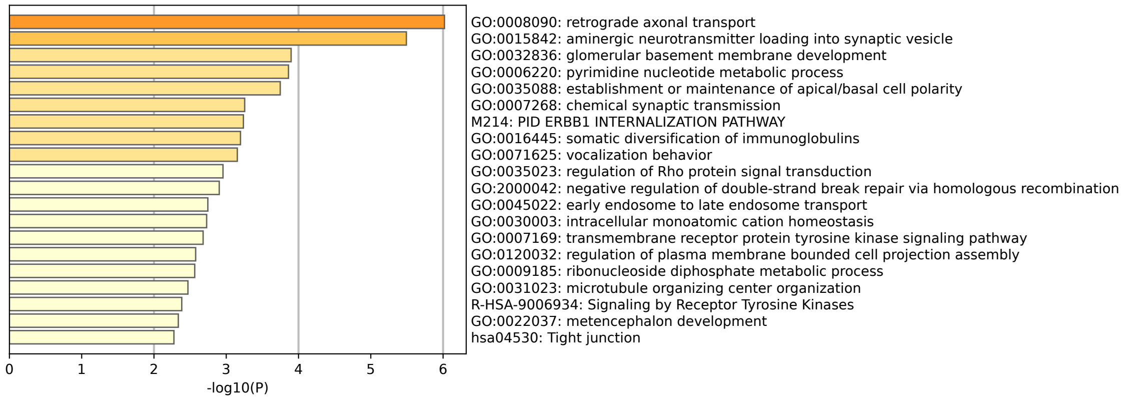


Figure 20: Metascape gene ontology pathways of genes related to DSB hotspots in calcarine sulcus.

4.1.4 DSB enriched genes and Recurrent DSB Cluster (RDC)

4.1.4.1 RDC related genes mouse and human

Previously Wei et al. (2016) identified recurrent DSB clusters (RDC) in murine neural stem/progenitor cells (NSPCs) from frontal brain and provided 27 genes that are localized within the break clusters. Moreover, in a new study, they identified additional clusters and genes commonly associated with neuronal function and disease (Wei et al., 2018). I asked whether these recurrent DSB clusters respectively the affected genes are prone to DSBs in human, too. To that end, genes, which were annotated by HOMER when providing the DSB hotspots, are extracted and compared with the dataset of genes related to RDCs of the mouse brain studies (Wei et al., 2018) (list of reference genes in Supplement A, I). Interestingly, several genes related to mouse RDCs can be identified as highly DSB affected genes in the human brain. For example, the calcarine sulcus and olfactory bulb share 5 gene breaking clusters with mouse, cerebellum 12, prefrontal cortex 18 and hippocampus 19 (Table 11). SOX5 is found in all tested brain regions but the majority of the genes studied are either region specific or shared in a subset of tested human brain regions. When combining and deduplicating the occurrences of RDC genes, a total of 40 mouse RDC genes can be found in the human brain.

The 40 identified RDC genes (Supplement A, II) of the tested human brain regions are evenly distributed across the major chromosomes and not clustered (Figure 21A) on specific chromosomes. Metascape results of the predicted or known diseases related genes reveal that the majority of the 40 RDC genes are potentially involved in neurological and mental diseases (Figure 21B), including Alzheimer´s disease, schizophrenia, bipolar disorder, autism spectrum disorder, cerebellar ataxia, etc. but also various types of cancer.

Table 11: Mouse RDC related genes shared with DSB hotspots of the human brain

Sample	RDC related genes
Bulb, Calca, Cereb, Hippo, Pfc	SOX5
Bulb, Cereb, Pfc	CSMD1
Bulb, Cereb, Hippo	CHRM3
Calca, Cereb, Hippo	DIP2C
Cereb, Pfc	NRXN1, CADM2
Hippo, Pfc	CENPP, AUTS2, DGKI, MDGA2

Bulb, Hippo	SOX6
Calca, Hippo	DGKB, CSMD3
Pfc	NAALADL2, PARD3B, SEMA6D, PTK2, DST, NR3C2, DMD, RBFOX1, PACRG, FGF14
Bulb	EXOC4
Calca	NAV2
Cereb	MAP3K4, VAV3, PID1, ASTN2, PTPRG, GRID2
Hippo	GRIP1, NBEA, DOCK1, CHD6, NPAS3, PTPRD, LRRC4C, CTNND2, CCSER1

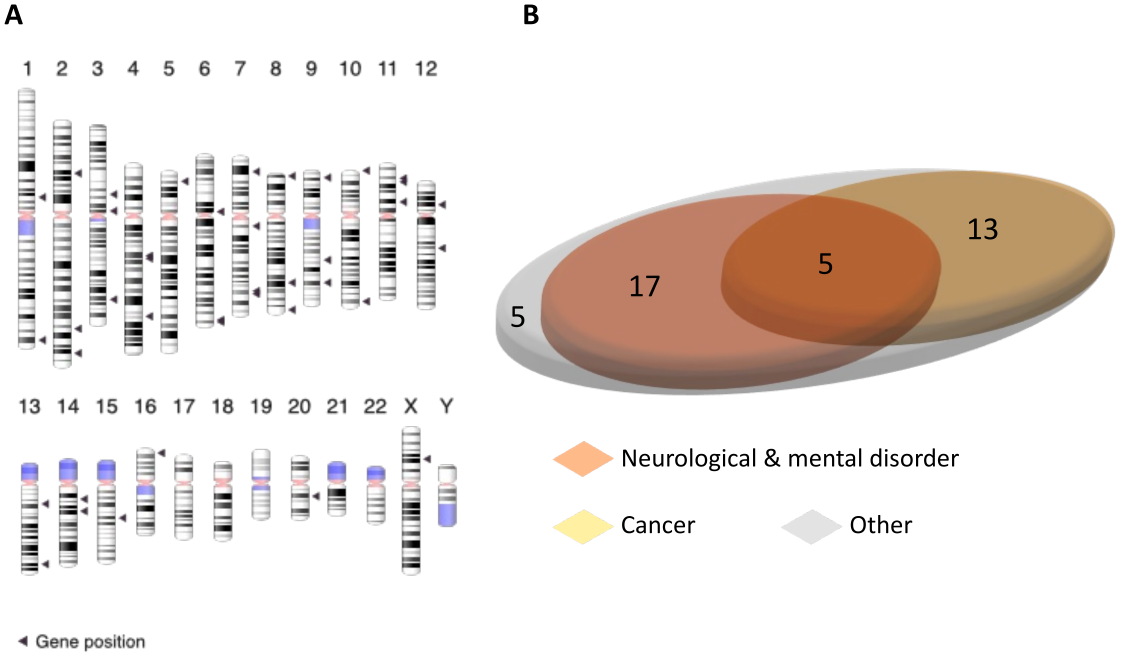


Figure 21: (A) Localization of all RDC genes, identified in human brain regions, on human chromosomes (GRCh38.p13); generated with NCBI's Genome Decoration Page (GDP). (B) Disease association of the 40 identified RDC genes, predicted by Metascape.

4.1.4.2 Recurrent *de novo* DSB enriched genes

Alongside demonstrating the presence of known RDC genes respectively genes that are frequently prone to DSBs, I also examined the occurrence of *de novo* RDC genes within the human brain. The genes of DSB peaks, that are annotated with HOMER, are extracted and compared between all brain samples. Overall, 48 genes (Supplement A, III) can be identified that are shared in all tested brain regions and are categorized as DSB hotspot based on MACS peak calling. The DSB related genes are evenly distributed across the human chromosomes (Figure 22A), similar to the genes shared with mouse datasets (Figure 21). Again, based on Metascape predictions, a total of 27 genes found in all human brain regions are associated with neurological or cancer diseases (Figure 22B). 21 genes are not characterized or belong to other pathways. The most interesting DSB susceptible genes that are involved in neural pathways and described as potentially disease related are listed in Table 12.

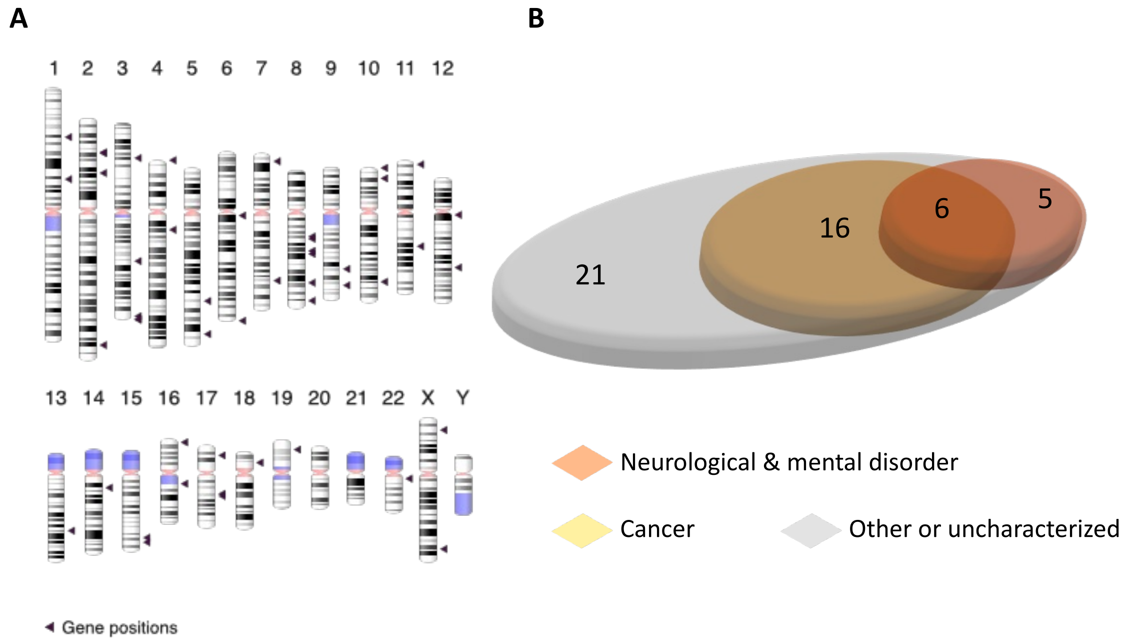


Figure 22: (A) Localization of *de novo* RDC genes, identified in human brain regions, on human chromosomes (GRCh38.p13); generated with NCBI's Genome Decoration Page (GDP). (B) Disease association of the identified RDC genes, predicted by Metascape.

Table 12: Top *de novo* RDC genes and their neural function and disease association

Gene name	Function	Disease
AGBL4	GO:0098958 retrograde axonal transport of mitochondrion	breast cancer and glioblastoma
DLG2	GO:0099642 retrograde axonal protein transport;GO:0099640 axo-dendritic protein transport;GO:0099641 anterograde axonal protein transport	schizophrenia and autism spectrum disorder
FBXW11	GO:0008090 retrograde axonal transport	cancer, neurodegenerative disorders, and cardiovascular diseases

ULK4	GO:2001222 regulation of neuron migration	hypertension and psychiatric disorders
VCX3B	GO:0007420 brain development;GO:0060322 head development;GO:0007417 central nervous system development	male infertility

4.1.4.3 Region specific DSB enriched genes

When examining all MACS provided DSB peaks as Venn diagram (Figure 23C), it becomes evident that numerous DSB hotspots are region specific, thus certain genes may be more susceptible to DSBs in one brain region compared to another. As a result, the region specific, DSB related genes are analyzed in the next step. The MACS peaks, which are annotated by HOMER as gene-related and are unique for a brain region, are analyzed with Metascape to predict disease pathways.

Since I am focusing on the impact of DSBs in the human brain, only genes associated with neurological diseases like Alzheimer´s disease, Parkinson’s disease and mental disorders as well as glioblastoma related genes are considered.

Numerous DSB prone genes (Supplement A, IV), which are associated with neurological diseases or glioblastoma, can be identified (Figure 23B) and are distributed across the major human chromosomes (Figure 23A). Moreover, the identified genes are DSB hotspots that are exclusively found in the respective brain area.

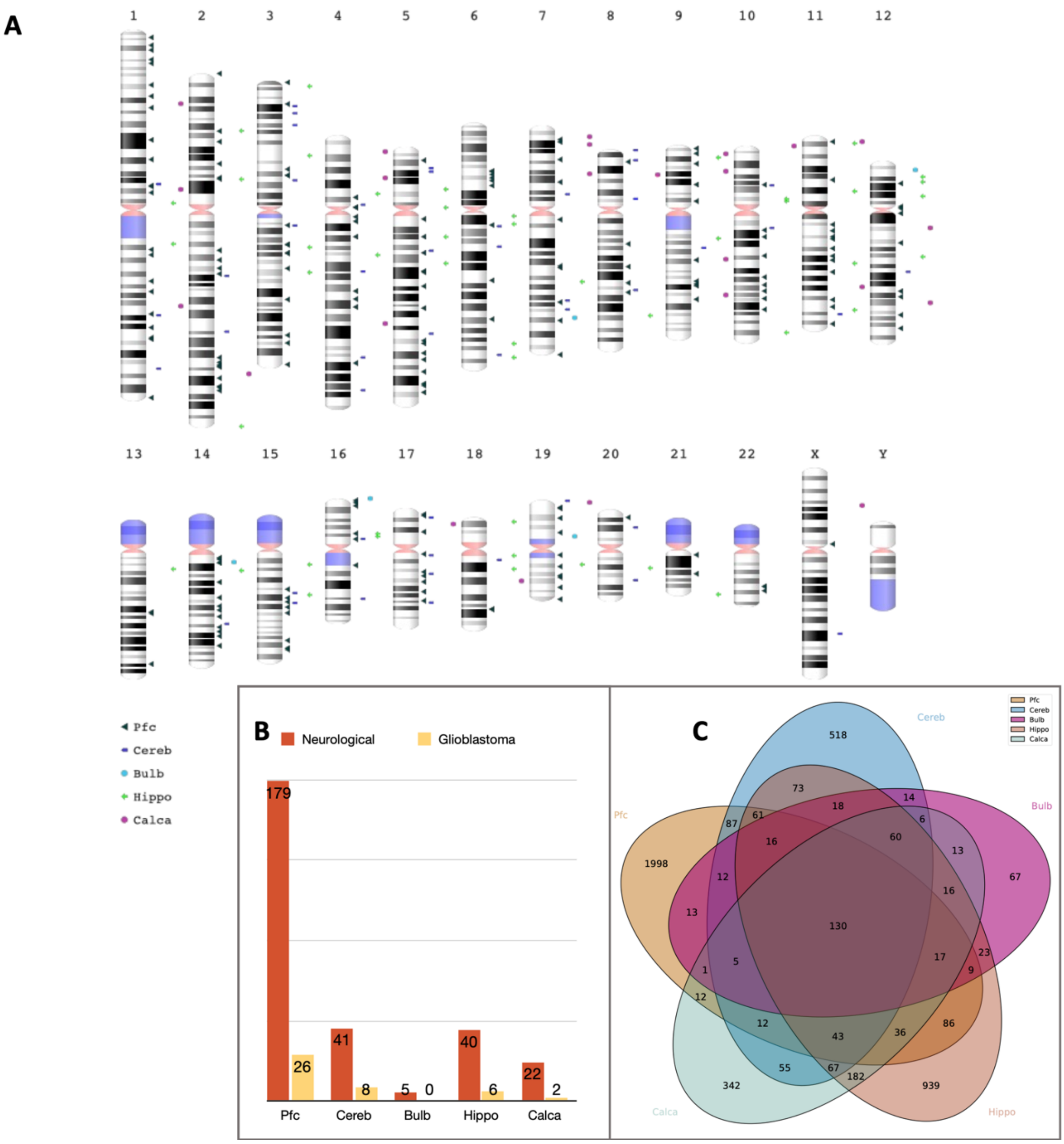


Figure 23: (A) Chromosomal distribution (GRCh38.p13) of brain region specific DSB prone genes that are associated with neurological disease or glioblastoma; generated with NCBI's Genome Decoration Page (GDP). (B) The number of fragile genes categorized as related to cancer or neurological disease in each brain region, predicted by Metascape. (C) Venn diagram of all MACS-provided peaks (not restricted to the gene related peaks), depicting the shared DSB hotspots of tested brain regions.

4.1.5 Shared DSB hotspots

Since recurrent DSB hotspots are detectable (Figure 23C), which are shared within all brain regions, the corresponding peaks are annotated with HOMER. Interestingly, the majority of peaks can be assigned to a region containing SINEs or Satellites (Figure 24A). More specifically, the SINEs are of Alu origin, including AluS, AluY and AluJ and most of the satellites are assigned to alpha-satellites of the centromeric regions (Figures 24B, C).

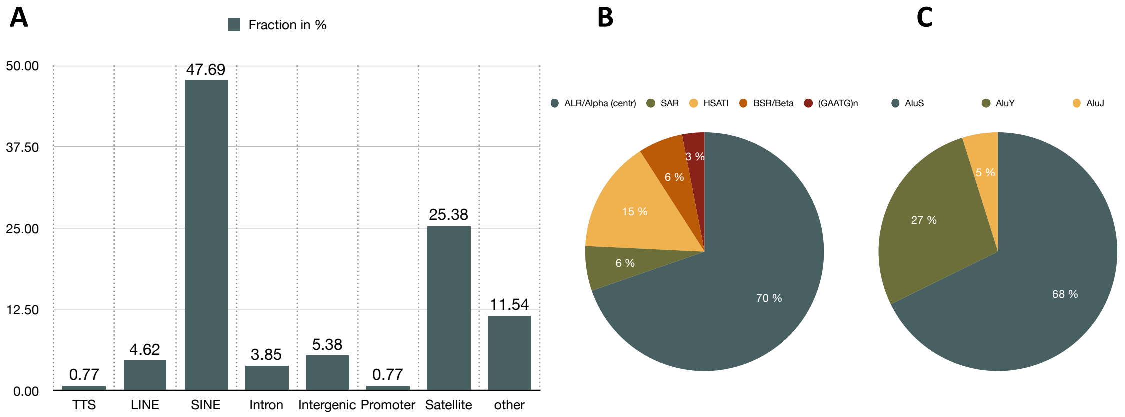


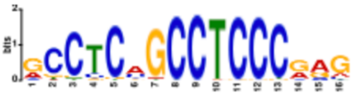
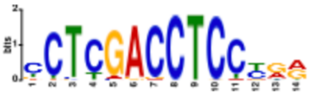
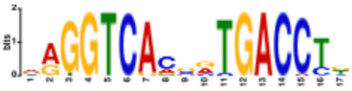
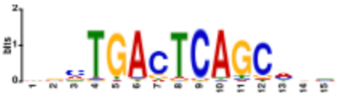
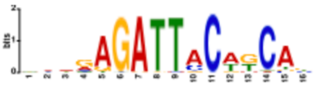
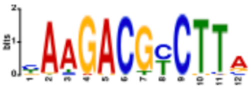
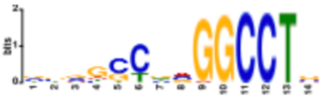
Figure 24: (A) Annotation of genomic features of DSB hotspots that are shared in all tested brain regions as fractions of annotated feature with respect to sum of all HOMER-annotated features in %. Features are: transcription termination site from −100 bp to +1 kbp (TTS), LINE and SINE transposons, Intron, intergenic region (Intergenic), promoter-TSS from −1 kbp to +100 bp (Promoter), satellite region (Satellite). (B) Classes of HOMER annotated satellites as % of all annotated satellites. (C) Classes of HOMER annotated SINEs as % of all annotated SINEs.

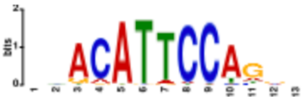
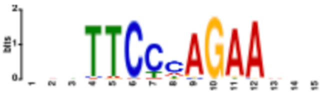
4.1.6 Motif search in DSB hotspots

Next, the DSB hotspots are analyzed for specific enrichment of certain motifs respectively DNA-binding sites of proteins. For this purpose, the underlying sequences of DSB hotspot coordinates of each sample are obtained by using BEDTools with hg38 as reference and are analyzed with the motif enrichment tool SEA. The SEA results are compared between the different brain regions and common motifs are further evaluated because they may be a predictor of DSB sites.

In total, 9 motifs with a relative enrichment greater than 1 (ratio of the motif in the primary sequences vs. control sequences; control = primary sequences shuffled) and a significant enrichment of the motif according to the statistical testing ($p < 0.05$) can be detected in the five brain regions (Table 13 & Supplementary Table 1, Supplement A). Moreover, the DNA-binding proteins associated with the predicted motifs can be identified and their functions are predominantly linked to transcriptional regulation based on UniProtKB/Swiss-Prot annotation (Bateman et al., 2023; https://www.uniprot.org; accessed on 10 July 2023).

Table 13: Enriched motifs of DSB hotspots shared in all samples

Logo	Motif sequence	Motif binding Protein	Function or related pathways
	GCCTCMGCCTCCCRAG	ZNF460	Transcriptional regulation; UniProtKB/Swiss-Prot: Q14592
	CCTCGACCTCCYRR	ZNF135	Transcriptional regulation; UniProtKB/Swiss-Prot: P52742
	ARGGTCACSRTGACCTK	ESR1	regulation of eukaryotic gene expression, affect proliferation and differentiation in tissues. UniProtKB/Swiss-Prot: P03372
	NGMTGACTCAGCMNH	MAFK	Transcriptional regulation; UniProtKB/Swiss-Prot: O60675
	DHDGAGATTACWKCAK	ZNF85	Transcriptional regulation; UniProtKB/Swiss-Prot: Q03923
	YAAGACGYCTTA	PROX1	Transcriptional regulation; UniProtKB/Swiss-Prot: Q92786
	SSSGCCBVGGCCTS	Zfx	Probable transcription activator; UniProtKB/Swiss-Prot: P17010

	NNACATTCCAGSN	TEAD1	Transcriptional regulation; UniProtKB/Swiss-Prot: P28347
	NNNTTCCCAGAANNN	Stat5b	Signal transduction and activation of transcription; UniProtKB/Swiss-Prot: P51692

4.1.7 DSB hotspots and epigenetic marks in prefrontal cortex

Next, the overlap between DSB hotspots and various marks is analyzed, including literature supported DSB marks like DNase sensitive sites and CTCF binding sites, to assess the methods applied in the present work. Moreover, certain overlaps with epigenetic marks may provide further insight into the localization of DSB hotspots.

Given the prefrontal cortex's significance as one of the most interesting areas of the human brain and the availability of sufficient datasets for several epigenetic marks, the following section will focus on prefrontal cortex as an example.

The chromosomal coordinates in BED format of different (epi-) genetic and DSB peaks/marks are intersected with the 2538 DSB hotspot coordinates of prefrontal cortex using BEDTools. The peak datasets are publicly available, including encode and epigenetic roadmap projects, and are listed in method 2.3.13.9 (Tables 3 & 4). The R-loop peak data (generated with DRIP-seq methods as described by Scheuren et al., 2023) were collected in our laboratories and originate from tissue samples of prefrontal cortex from the same individual as used for the DSB analysis.

The DSB hotspot coordinates that overlap with coordinates of specific marks are extracted and each dataset is compared pairwise to evaluate whether multiple marks are localized at the same DSB hotspot. When performing pairwise comparison of the DSB hotspot related marks, several datasets share common DSB hotspots (Figure 25A). The most significant overlaps is observed for DNase-seq data along with ATAC-seq and ChIP-seq of H3K4me3 and H3K27ac (Figures 25A, B).

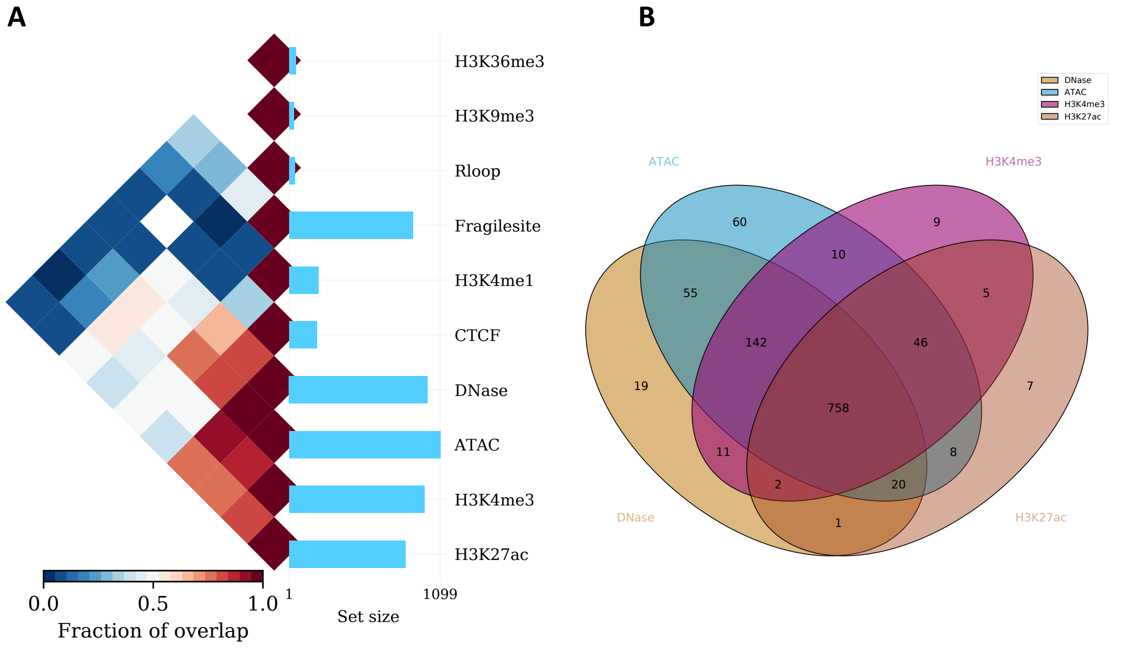


Figure 25: (A) Illustration of pairwise comparison of prefrontal cortex DSB hotspot coordinate datasets that are associated with a specific epigenetic mark, e.g. comparing H3K27ac and H3K4me3 related DSB coordinates for intersection of the same DSB hotspots. The heatmap reflects the degree of shared DSB hotspots when comparing all datasets pairwise. (B) The Venn diagram illustrates the number of prefrontal cortex DSB hotspots that coincidently are associated with ATAC-seq, DNase-seq, H3K4me3 and H3K27ac coordinates.

The question arises whether the increase in overlaps of DSB hotspots with DNase-seq, ATAC-seq and ChIP-seq of H3K4me3 and H3K27ac are detectable based on the set size alone (indicated by blue bars in Figure 25A). To that end, the 2538 DSB hotspot are shuffled, resulting in the same number of peaks and peak size but the peak coordinates are different. This procedure is carried out to generate a total of n = 3 random DSB hotspots datasets. The random coordinates of each dataset are intersected with the epigenetic mark coordinates using BEDTools. The resulting overlapping peak numbers of each random dataset are averaged for n = 3 and used as the expected value when assuming random overlaps. The observed counts (DSB peaks overlapping with epigenetic mark) are divided by mean expected values (shuffled peaks overlapping with epigenetic mark). The resulting values of observed/expected can be interpretated as fold change (Table 14). Overall, marks of DNase-seq, ATAC-seq and ChIP-seq of H3K4me3 and H3K27ac, which were demonstrated to be frequently together at DSB hotspots (Figure 25B) are more enriched at DSB hotspots as expected (Figure 26A). DNase-seq positions contain a fold change of 168 in observed dataset in contrast to random datasets and thus represent the highest enrichment at DSB sites, followed by H3K4me3, CTCF, H3K27ac and ATAC-seq. The fragile sites, H3K36me3 and R-loops do not show significant enrichment at DSB sites.

Table 14: Number of observed overlaps of DSB hotspots and epigenetic marks vs. expected overlap of randomly generated DSB hotspots and epigenetic marks in prefrontal cortex.

Mark	Observed dataset (sample)	Random dataset 1	Random dataset 2	Random dataset 3	Mean of random datasets	Observed/expected
ATAC	1099	48	66	61	58.33	18.84
CTCF	202	6	7	11	8	25.25
DNase	1008	6	6	6	6	168
fragile	902	849	853	827	843	1.07
H3K4me1	212	101	103	117	107	1.98
H3K4me3	983	26	30	37	31	31.71
H3K9me3	35	20	19	11	16.67	2.1
H3K27ac	847	38	41	47	42	20.17
H3K36me3	48	85	107	119	103.67	0.46
R-loop	42	13	12	8	11	3.82

Since the data demonstrate that DNase sensitive, ATAC-seq, H3K4me3 and H3K27ac marks are frequently together and enriched at DSB hotspots, only the DSB coordinates exhibiting the four epigenetic marks are extracted for further analysis. Coordinates that are additionally in a common fragile site are discarded to focus on the positions described by the active markers (frequently at open chromatin, transcription). The resulting pool of DSB hotspots includes 454 coordinates, which are analyzed with HOMER and the obtained genes with Metascape. The majority (91%) of DSB hotspots is assigned to a position associated with the promoter-TSS from −1 kbp to +100 bp of genes (Figure 26B). The GO analysis, obtained from Metascape, demonstrates that the HOMER-annotated genes are predominantly related to transcriptional and translational pathways (Figure 26C), including ‘Metabolism of RNA’, ‘ncRNA metabolic process’, ‘Translation’ but also pathways involved in cell cycle processes.

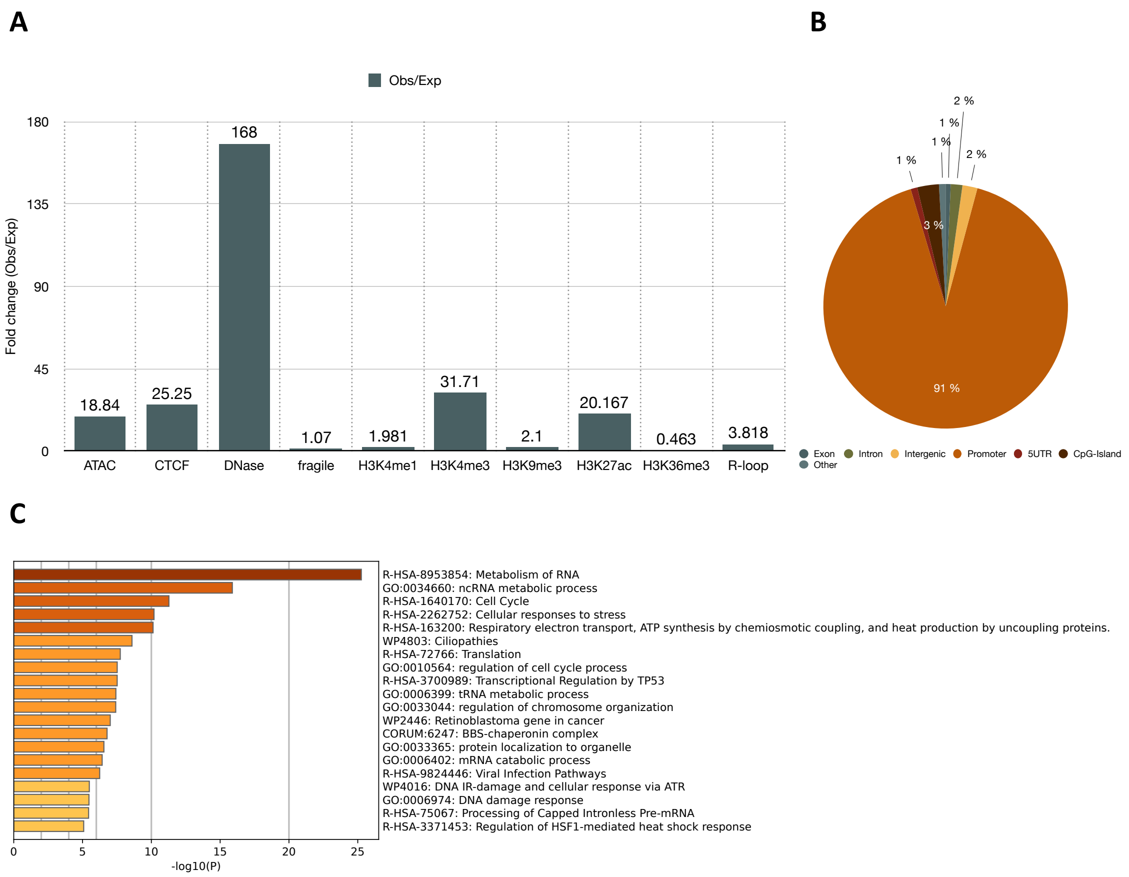


Figure 26: (A) Diagram of DSB hotspot association with specific epigenetic marks in prefrontal cortex. The fold change values are calculated by dividing observed values (number of DSB hotspot associated with epigenetic mark) by expected values (average number of random DSB hotspots associated with epigenetic mark). (B) Homer annotation of DSB hotspot coordinates that are associated with all active epigenetic marks (DNase sensitive, ATAC-seq, H3K4me3 and H3K27ac marks). (C) Metascape gene ontology pathways of genes related to the analyzed DSB hotspots with active epigenetic mark.

4.1.8 Association of DSBs and *de novo* SVA insertions

To put the presented DSB results into retrotransposon context (chapter 3.1), I checked whether certain *de novo* SVA flanks (*de novo* SVA integrations of the same person used for DSB analysis) contain an enrichment of DSBs. To that end, the unique DSB positions obtained from the implemented bioinformatics pipeline are counted in each *de novo* SVA flank by intersecting the coordinates with BEDTools. In addition, 3 random datasets of DSB positions with the same number as in the original file are generated and also counted in the *de novo* SVA flank positions. The mean value of the random occurrence in *de novo* SVA flanks is set as the expected value. The observed DSB count is divided by the expected count for the corresponding *de novo* SVA flank and plotted as a boxplot. The majority of *de novo* flanks contain the expected number of DSBs, but 28 – 226 *de novo* SVA flanks have a fold change of more than 2 in observed datasets compared to expected data (Figure 27A & Table 15). The coordinates of DSB enriched SVA flanks (foldchange value greater 2) are analyzed with HOMER to illustrate the underlying genomic feature (Figure 27B). The DSB rich SVA flanks are predominantly located in retrotransposons like LINE, SINE and LTR, but also in other intronic, intergenic or satellite regions.

Table 15: *De novo* SVA flanks and DSB association

Sample	*De novo* SVA flank count	*De novo* SVA flanks with DSB count greater than expected value (Fc > 2)	% of *de novo* SVAs
Pfc	2566	65	2.53
Cereb	2838	80	2.82
Bulb	5540	226	4.08
Hippo	2026	81	4.00
Calca	748	28	3.74

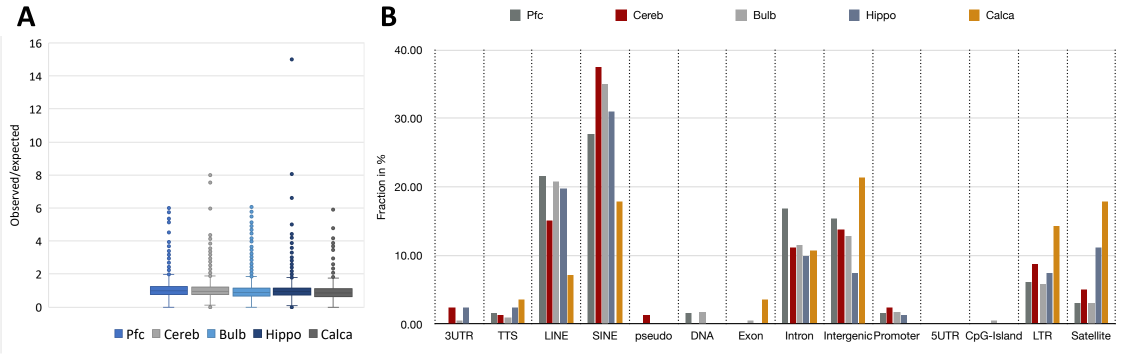


Figure 27: (A) Boxplot of fold change values of DSB counts per *de novo* SVA flank compared to random DSB counts per *de novo* SVA flank (observed/expected). (B) Homer annotation of *de novo* SVA flanks with a fold change value greater 2, as fractions of annotated feature with respect to sum of all HOMER-annotated features in %. Features are: 3´ untranslated region (3UTR), transcription termination site from −100 bp to +1 kbp (TTS), LINE and SINE transposons, pseudogene (pseudo), DNA transposons (DNA), Exon, Intron, intergenic region (Intergenic), promoter-TSS from −1 kbp to +100 bp (Promoter), 5´ untranslated region (5UTR), CpG-Island, long terminal repeats (LTR), satellite region (Satellite).

4.2 Discussion

The human brain is a highly metabolic active tissue with elevated energy consumption and mitochondrial activity. Consequently, the human brain is exposed to ROS but also to replication stress, L1 endonuclease activity, topoisomerases that alter the topological state of DNA double helix during cell cycle, and transcription, i.e. factors related to the induction of DNA double-strand breaks. In the present work, I investigate the existence of chromosomal DSB hotspots and possible differences in their genomic localization among five human brain regions. To identify chromosomal DSB hotspots in distinct human brain regions and describe the respective ‘breakome’, a DSB labeling system based on Breaks Labeling In Situ and Sequencing (BLISS) is introduced (Yan et al., 2017). The double-strand break sites of samples from the prefrontal cortex, cerebellum, olfactory bulb, hippocampus and calcarine sulcus are directly labeled with an adapter and enriched by in vitro transcription and PCR. The PCR products are deep sequenced and the DSB related reads are identified by bioinformatically scanning the reads for the adapter introduced barcode sequence. Reads resembling the DSB flanking sequence are mapped to hg38 to determine the chromosomal coordinate of a genomic DSB. By analyzing the adapter related UMI sequences, only the unique DSB coordinates are obtained and further examined.

4.2.1 DSB detection and evaluation of applied methods

Overall, this experimental and bioinformatical procedure can detect 6,241,600 – 22,695,802 unique DSB positions in the studied brain regions. The number of DSBs corresponds to an estimate of 37.45-136.17 DSBs per cell (Table 8), which is in the range of values observed in other studies (Lensing et al., 2016; Vilenchik & Knudson, 2003). However, we must keep in mind that this is an approximate estimate because the number of cells for the 1 µg of genomic DNA input may vary and some DSBs could be the result of random fragmentation. Since only 1 µg of input is sufficient to detect a number of DSBs per cell comparable to methods such as DSBcapture that require 50 µg of DNA input, the BLISS method tends to be advantageous when samples are precious and DNA quantity is limited (Lensing et al., 2016; Yan et al., 2017).

Multiple DSB detection methods are important because they complement the overall profile of DSBs. Nevertheless, certain methods have a drawback, such as ChIP-seq of γ-H2AX, which is generated as a cellular response to DSBs by histone H2AX phosphorylation, but has low resolution (Turinetto & Giachino, 2015). Moreover, there are certain methods that cannot directly detect DSBs like CC-seq, which only labels specific features such as covalently bound proteins (Gittens et al., 2019), or require large input of cells including BLESS (Crosetto et al., 2013) and DSBCapture (Lensing et al., 2016). BLISS, on the other hand, is sensitive due to the in vitro transcription of labeled DSBs, the high resolution of direct labeling of DSBs, and the quantitative use of UMIs as controls for PCR amplification bias (Yan et al., 2017). Concordantly, the herein presented results demonstrate the properties of BLISS, indicated by low input but still similar detection of DSBs per cell as seen in other proposed methods, and high resolution of DSB hotspots, as indicated by MACS peak calling. In contrast to the original BLISS method, the present experiments involved the labeling of DSBs in isolated DNA from bulk tissue - as proposed in dDIP and DSB-Seq methods (Baranello et al., 2014; Leduc et al., 2011) - rather than in situ within fixated cells. The dDIP, or damaged DNA immunoprecipitation method, demonstrated that the detection of enriched breakage sites in extracted DNA remains sensitive when the integrity is maintained during extraction methods (Leduc et al., 2011). However, labeling of breaks induced by extraction methods may occur but the present protocol implements mild isolation of DNA using SDS-Proteinase K isolation methods to reduce shearing compared to column based approaches. Moreover, the NEB blunting enzyme was incubated for 15 min, which is appropriate for fragmented DNA similar to restriction enzyme digested conditions, as opposed to the 30 min incubation time that favors blunting of sheared/nebulized DNA.

4.2.2 DSB hotspot identification

To overcome the potential problem of incorrectly determine DSBs of random fragmentation as true endogenous DSB, only the accumulation of DSBs respectively genomic hotspots are obtained using MACS peak calling. MACS proves to be an ideal tool for calling DSB hotspots, which was observed by Lensing et al. (2016) and can also be observed in the present experiments because the obtained peaks show a fold change of more than 6 for the DSB hotspot density compared to the normal distribution of DSBs in the human genome (Figure 13). In total, 423 DSB hotspots were identified in the olfactory bulb, 1,003 in calcarine sulcus, 1,178 in cerebellum, 1,782 in hippocampus and 2,538 in prefrontal cortex (Table 9). The hippocampus and prefrontal cortex are the two regions with the highest number of DSB hotspots, which may be related to the increased detection of oxidative stress in these regions. Venkateshappa et al. (2012) demonstrated increased oxidative stress and a progressive decline in antioxidant function with age in human frontal cortex and hippocampal tissue, arguing that such regions exposed to ROS are susceptible to oxidative damage. Moreover, the susceptibility of hippocampus and frontal cortex to oxidative stress has been investigated in several studies and may lead to functional impairment and progression of behavioral and neurological diseases (Salim, 2017; Stefanatos & Sanz, 2018). The increase in detectable DSB hotspots in prefrontal cortex and hippocampus may therefore be directly related to increased oxidative stress, suggesting that ROS, which are generally acknowledged to damage dsDNA, may drive the induction of DSBs in the human brain. Interestingly, it has been reported that oxidative stress and ROS at biologically relevant levels can induce clustered DNA lesions as closely spaced lesions like single-strand breaks that subsequently introduce DSBs, thus may represent one factor in the development of DSB hotspot (Sharma et al., 2016). Moreover, such events are thought to induce NHEJ-mediated mutagenesis. Therefore, the differences in DSB hotspot localization and DSB quantity in the tested samples (Figure 23C) could have an impact on somatic mutational heterogeneity (mosaicism) in the human brain.

4.2.3 DSBs and transcription

Another major factor for DSB formation is transcription, where e.g. R-loops can be formed and accumulated at transcriptional termination regions, inducing nicks as well as DSBs when not properly removed (Ui et al., 2020). In addition, TOP2 forms DSBs through strand-cleaving activity at active transcription sites and Pol II promoter-proximal pausing sites are frequently enriched in DSBs (Singh et al., 2020). Taken together, these observations clearly indicate that genic regions are susceptible to DSBs, with several studies having already characterized specific genes and clusters that are affected (Tchurikov et al., 2022; Wei et al., 2016). In accordance with the aforementioned findings, it can be suggested that chromosomes with higher gene density tend to be more affected by DSBs, as demonstrated in the present work for the majority of brain regions with DSB accumulation on gene-dense Chr. 17 and Chrs. 20 – 22 (Figure 14). Chromosomes with lower gene density like Chr. 13 and Chr. Y are less affected (the gene density of human chromosomes is depicted in Figure 1B of Mayer et al., 2005).

Moreover, we can further review the relationship between DSB hotspots and genes by retrieving the underlying genomic feature annotation from HOMER (Figure 15).The DSB hotspots are located in regions related to the gene body, like 3´-UTR, TTS, promoter, exon and intron in all brain regions, further supporting the induction of DSBs at genic regions. Analyzing the related GO pathways of the HOMER annotated genes reveals the association of DSB affected genes and neural pathways such as ‘protein localization to synapse’ (GO:0035418), ‘synaptic signaling’ (GO:0099536), ‘retrograde axonal transport’ (GO:0008090), ‘axo-dendritic transport’ (GO:0008088) but also pathways of transcriptional activity like ‘Metabolism of RNA’ (R-HSA-8953854) and ‘Ribosome’ (hsa03010). Since many genes are assigned to neural pathways, they support the assumption of an accumulation of DSBs in genes that are transcriptionally active and contribute to the function of the brain. Similar results, in which DSBs occurred predominantly in active genes that control differentiation, development, and morphogenesis, were presented by Tchurikov et al. (2022) and further demonstrate that transcriptionally active genes within open chromatin regions can be susceptible to DSBs.

When combining the results of chromosomal DSB density, feature analysis and GO pathways, the DSB-rich Chr. 19 of prefrontal cortex poses an interesting target for further analysis. Chr.19 is the human chromosome with the highest gene density (Mayer et al., 2005) and the prefrontal cortex has the highest DSB enrichment on this particular chromosome. This also suggests that active genes are affected by DSBs and over 30% of all DSB hotspots in prefrontal cortex are located in promoter regions. When extracting all DSB related promoters respectively their associated genes from Chr. 19, which are affected exclusively in prefrontal cortex, 83 promoter related genes can be analyzed. The gene related GO pathways (Supplementary Figure 1, Supplement A) reveal the involvement of prefrontal cortex genes in the regulation of transcription by RNA Polymerase II, DNA-templated transcription and ribosome assembly. RNA Polymerase II is required for the synthesis of mRNAs, thus GO pathways, including ribosomal pathways, generally reflect the synthesis of proteins. Harris et al. (2009) performed whole genome microarray experiments on normal prefrontal cortex tissue and observed altered gene expression during adolescence, i.e. switch from expressed genes involved in neuronal development and plasticity to genes associated with energy metabolism, including protein and lipid synthesis. This may explain the differences in DSB affected genes when comparing the prefrontal cortex with the other brain regions studied because the transcription patterns of prefrontal cortex can change (Figures 16 – 20). The DSB affected genes of Chr. 19 of the prefrontal cortex are involved in processes of protein metabolism similar to those of the pathways annotated for the total DSB hotspots (Figure 16) and Tchurikov et al. (2022) provided evidence for a similar frequent DSB occurrence in genes associated with metabolism.

4.2.4 DSBs and retrotransposons

In addition to gene association, the feature analysis provides information on other DSB associated features and demonstrates that, for example, retrotransposons like LINE-1, SINEs and LTRs as well as satellite regions are DSB enriched (Figure 15). When analyzing the DSB hotspots that are shared in the five brain regions, retrotransposons, especially Alus, and satellites, here mainly centromeric alpha-satellites, remain the most important DSB affected regions (Figure 24).

The DSB hotspots are detectable at sites where reintegrations of retrotransposons have occurred, i.e. regions that tend to be susceptible to L1-mediated reintegrations. The TE rich sites are often characterized by an open chromatin state, otherwise they would be less sensitive to L1 endonuclease activity. Gasior et al. (2006) demonstrated that the L1 encoded endonuclease frequently induces DSBs and such events would be more frequent at sites that are composed of open chromatin and where reintegration is favored. This suggests that L1 activity may induce DSB hotspots at repetitive TE regions but similar TE remnants at DSB sites could also be a consequence of DNA repair unrelated to L1 activity, which we will discuss in more detail in the following section.

Since there is a relation between DSB sites and retrotransposons, I wanted to test whether certain *de novo* SVA insertion (chapter 3.1), from the same individual and brain region, could be attributed to a genomic region with an unusually high DSB density. 2.53 – 4.08% of all *de novo* SVA insertions of the five tested brain regions have a fold change value greater than 2 when comparing the observed with the expected DSB counts in the flanking region of *de novo* SVAs (Figure 27). This suggests that the majority of *de novo* SVA insertions are attributable to the standard L1 mediated reintegration events. However, the 28 – 226 *de novo* SVA insertions with an unusual high DSB density in the flanking region could be related to a DSB repair pathway. The major repair pathways in human cells are NHEJ and HR but recent studies suggest that TE-templated DNA repair pathways can contribute to the maintenance of genome integrity, although not to the same extent as the major pathways. Srikanta et al. (2009) proposed an Alu element integration method distinct from the target site-primed reverse transcription (TPRT) via LINE-1, suggesting that such integrations are the cause of DNA double-strand break repair mechanisms. In 2015, Ono and colleagues provided further evidence for a TE-mediated DNA repair pathway. Using CRISPR/Cas induced DSBs, they demonstrated the existence of RT-product-mediated DSB repair (RMDR), in which a pre-existing cDNA is annealed to the DNA ends of a DSB like a ‘bridge’ or an RNA is annealed to one DSB end and cDNA is synthesized by RT to repair the DNA. In both cases, microhomologies are present at the DSB site and retrotransposon cDNAs or RNAs can serve as templates. The *de novo* SVA integrations of the present work with unusually high DSB counts (fold changes of 2 – 8 in observed compared to expected) in the flank sequence could be related to a RMDR pathway and used as template to reconnect the DNA ends. The DSB associated *de novo* SVAs are predominantly located in regions with repetitive retrotransposon like LINEs, SINEs and LTRs, which are excellent targets for annealing via microhomology and thus support RMDR repair. Moreover, the majority of *de novo* SVAs with high DSB density are unique to the respective brain region (Supplementary Figure 2, Supplement A), suggesting that they originated in the lineage of a certain brain region in late development or adolescence. The detected DSBs also reflect the current state of the individual, meaning they are present in the mature brain, and thus correlate with the lineage specific *de novo* SVA positions. In summary, the major pathways of DSB repair are HR and NHEJ but a small fraction of DSB sites may be repaired by RMDR utilizing SVA templates, thus representing a precursor to SVA mosaicism.

In the next part, we will focus on satellite regions as another genomic feature enriched in DSBs. All validated brain regions contain DSB hotspots in satellite regions, with centromeric alpha-satellites accounting for the majority (Figure 15 & Table 10). The analysis of shared DSB hotspots revealed that common breakage sites within the human brain are also located in satellite regions, especially in centromeric alpha-satellites (Figure 24). Several studies demonstrated that centromeric satellite regions are frequently affected by DSBs. Sources of this centromeric instability include collisions of replication and transcription forks, the formation of R-loop, mutagen exposure and secondary structures (Black & Giunta, 2018). Concordantly, the DSB induction in centromeric regions is often associated with proliferation but DSBs can also be induced in centromeric regions by topoisomerase IIB during quiescence (Saayman et al., 2023).

4.2.5 Recurrent breaking clusters and associated genes

Topoisomerase II is frequently mentioned in the context of DSB induction and can be associated with transcription sites and active genes. Recently, researchers investigated whether genes frequently affected by DSB clusters can be detected. To this end, Wei et al. conducted experiments on NPCs derived from the frontal brain of mice and were able to pin point so called recurrent DNA break clusters (RDC) in which 27 genes are located (Wei et al., 2016). In another study, they were able to confirm previous findings and identified several similar RDC affected genes (Wei et al., 2018). The characterized genes are in most cases associated with neuronal functions as well as described in many neurological conditions including mental disorders but also tumor, thus the authors suggested that the fragile genes respectively the DSBs may have an impact on the healthy brain. I evaluated whether the same fragile genes could be identified in a human setting and whether new fragile genes might be found. In total, 40 genes validated as fragile gene in mouse RDCs are also associated with DSB hotspots in the human brain (Figure 21). They can be associated with neurological and mental diseases like Alzheimer´s disease, schizophrenia, bipolar disorder, autism spectrum disorder, cerebellar ataxia, but also cancer. The identified genes are of particular interest because they are associated with neuronal functions and diseases. Therefore, studying the effects of DSBs on RDC genes in the mature human brain may provide insight into the susceptibility and possible association with mental disorders as well as age related neurodegenerative diseases. During differentiation of multipotent neural progenitor cells into neurons and glial cells, changes in chromatin state, DNA methylation and histone modifications are detectable, which ultimately leads to altered gene expression (Gurok et al., 2004; Yoon et al., 2018). DSBs of progenitor and mature cells may be localized differently depending on the access of chromatin and active genes, which explains the difference between human brain RDC genes and mouse NPCs.

Therefore, the next step included the identification of novel fragile genes that recur in all brain regions, similar to the mouse RDCs, to describe the state of the adult human brain and the effects of DSBs on health and disease. Several *de novo* RDCs, which are shared in all brain regions, can be identified and two genes are well described in the available literature (Figure 22 & Table 12). DLG2 and ULK4 are categorized as DSB-rich genes in the present work and are involved in neural function as well as disease. DLG2 encodes a postsynaptic scaffolding protein, which interacts with NMDA receptors, potassium channels, and regulates synaptic stability as well as potentially the synaptic plasticity. DLG2 is classified as a psychiatric risk gene and has been repeatedly documented in schizophrenia patients in association with *de novo* loss-of-function mutations (Fromer et al., 2014; Kirov et al., 2012). Other studies have reported that DLG2 deficiency is involved in excitatory synaptic deficits in the striatum and impaired synaptic integration and plasticity in the hippocampus (Griesius et al., 2022; Yoo et al., 2022). Based on multiple studies, the deficiency or loss of function of DLG2 has significant effects on the brain and shows a strong association with mental disorder, suggesting that the identified recurrent DSB hotspot in DLG2 could contribute to the disease state when DNA repair is aberrant and mutations accumulate. Another gene enriched in DSBs is ULK4, a gene that encodes a kinase family protein and was indicated to be important during neurodevelopment but has also been linked to psychiatric disorders (S. Luo et al., 2022). ULK4 deletions, e.g. intragenic fragment deletions, and SNPs have been detected in patients with schizophrenia, bipolar disorder and depression. Since ULK4 mutations could also be introduced by recurrent breaks, the DSB hotspots may be associated with mental disorders and pose another interesting research target.

Most DSB hotspots are unique to the respective brain region under scrutiny. Logically, it is possible to identify multiple genes enriched in DSBs but found exclusively in a particular brain area (Figure 23). Again, many of these fragile genes may be associated with neurological diseases such as Alzheimer’s disease, Parkinson's disease, bipolar disorder, schizophrenia or glioblastoma, as indicated by Metascape predictions. The region specific breaking genes are another indicator of somatic mosaicism in the human brain because they are potential hotspots for mutations. The question arises as to how certain genes can be affected differently across multiple brain regions. ATAC-seq of 14 distinct brain regions revealed brain region–specific chromatin accessibility, e.g. in neocortex, primary visual cortex, hippocampus, thalamus and striatum (Fullard et al., 2018). Moreover, they were able to predict region specific expression of protein-coding genes as a consequence of the different open chromatin regions, which was also demonstrated in the transcriptome analysis of different brain regions by Kang et al. (2011). Consequently, the differential expression may lead to different DSB hotspots as a result of processes related to transcription, such as TOP2 induced breaking sites or stalling of polymerases. In addition, differences in metabolic activity or oxidative stress can cause different ROS mediated breaking sites and it has also been demonstrated that open chromatin regions are more susceptible to radiation (Falk et al., 2008), thus distinct open chromatin regions may shape the different DSB hotspots.

4.2.6 Predictors of DSB hotspots

Mourad et al. (2018) showed that open and active chromatin and associated epigenetic landscapes can be predictors of DSBs in human. They analyzed ENCODE datasets and assessed the colocalization of DSBs and epigenetic marks as well as DNA-binding proteins, thus were able to accurately predict DSBs, especially with marks like DNase I hypersensitive sites, CTCF, p63, H3K4me1, H3K4me2, H3K4me3 and H3K27ac. With the predictors shown, we can evaluate whether the DSB data of the present experiments colocalize with similar epigenetic marks and are therefore also predictive, significant positions. The prefrontal cortex was analyzed because this region is well described in the literature and many epigenetic datasets like ATAC-seq, CTCF, DNase-seq, fragile sites, H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K36me3 and R-loop are available. ATAC-seq and DNase-seq are used to identify chromatin accessibility respectively open chromatin of the genome by transposase or endonuclease activity (Tsompana & Buck, 2014). CTCF is a zinc finger protein that influences gene expression by recruitment of other transcription factors or chromosomal interactions (Kim et al., 2015). CTCF and cohesin can colocalize on chromosomes to form loops to regulate the chromatin structure and act as activator or repressor of gene expression. Common fragile sites, found in numerous human samples and publications, are sensitive to replication stress and are frequently rearranged in tumors (Kumar et al., 2019). Histone H3 variants can be associated with chromatin dynamics, e.g. histone modifications like H3K4me3 and H3K27ac contribute to genome activation, and H3K9me3 can be associated with chromodomain proteins that form inactive heterochromatin (Martire & Banaszynski, 2020).

Consistent with Mourad's data, ATAC-seq, DNase-seq, CTCF and marks of H3K4me3 and H3K27ac are good predictors of DSB hotspots in the prefrontal cortex (Figure 26). This suggests that accessible chromatin regions as well as active transcription marks - indicated by H3K27ac and H3K4me3 - tend to be important regions of DSB induction and show a significant association of DSBs with gene expression. DSB hotspots associated with open and active chromatin tend to locate at promoter regions of genes in prefrontal cortex. Several studies demonstrate the generation of DSBs by TOP2 in promoter regions of active genes (Haffner et al., 2010; Ju et al., 2006) and also suggest an association with Pol II promoter-proximal pausing sites, which are often enriched in DSBs and thus affect transcription (Singh et al., 2020).

As for R-loops in the prefrontal cortex, R-loops as DNA-RNA hybrids can be formed and accumulated at transcriptional termination regions and thus induce nicks as well as DSBs if not properly removed (Ui et al., 2020). On the other hand, recent research investigates their role in DSB formation and hypothesizes that R-loops are part of the damage response to maintain the DSB affected genomic regions, especially active genes (Bader & Bushell, 2020). The formation of R-loops in damage response, which is distinct from standard co-transcriptionally formed R-loops, is still under debate. The R-loops in prefrontal cortex – only 42 R-loops of 17607 are located at DSB hotspots - do not present a quantitatively high enrichment at DSB hotspots and therefore cannot provide further support for DSB related R-loop formation in the human brain.

In the previous section, we discussed the association of DSB hotspots with already described predictors of DSBs, including epigenetic marks. Another approach that can be discussed focuses on the identification of enriched sequence motifs in DSB hotspots that may also predict sites of recurrent DSBs in the human genome. In total, 9 enriched sequence motifs are detectable in all studied brain regions and each motif can be described by the association of a specific DNA-binding protein (Table 13). The identified motifs respectively the DNA-binding proteins ESR1, ZNF460, TEAD1 and STAT5B are frequently found at DSB locations or associated with DSB mechanisms. ESR1 was described as a predictor of DSBs in the work of Mourad et al. (2018) and X. Zhang et al. (2023) were able to reduce DNA damage in chondrocytes by establishing a Knock-in of ESR1. Chen et al. demonstrated a link between ZNF460 motifs and DSBs at chromatin structures like loop anchors, as well as the association of RNA Pol II, DSBs and ZNF460 motifs, suggesting a role for ZNF460 in genome stability (H. Chen et al., 2023 PREPRINT). Calses et al. (2023) showed in their recent work that TEAD proteins interact with damage response pathway proteins, co-localize with DNA damage–induced nuclear foci and affect the cellular repair of DSBs, suggesting that TEADs are important for genome stability and DNA damage responses. The involvement of STAT5B in DNA damage response or related mechanisms is not recorded but the related motif is also enriched at dimethyl sulfoxide (DMSO)-specific genomic DSBs in DMSO treated cells (Kodali et al., 2022). In summary, the common sequence motifs or ChIP-seq data of DNA-binding protein such as ESR1, ZNF460, TEAD1 and STAT5B may provide additional prediction of

DSB locations in the human brain. Moreover, the DNA-binding proteins act as transcription factors and provide further evidence that DSB-rich positions are associated with active genes or transcription.

4.2.7 Conclusion of the DSB analysis

In conclusion, the implemented BLISS based DSB detection is able to sensitively identify double-strand break positions in bulk tissue of the different human brain regions. The potential background of DSBs – as a possible consequence of fragmentation or shearing - was sufficiently eliminated by obtaining the DSB hotspots of the human genome. The DSB hotspots are characterized by a significant enrichment of DSBs in contrast to the overall genomic distribution of DSBs. When comparing the different DSB hotspots in all tested brain regions, multiple genomic regions with shared occurrence of DSBs are detectable but also brain region specific breakpoints. The differences in quantity and localization of DSBs could be the result of different exposure to oxidative stress, transcriptional patterns and state of open chromatin in the brain regions under scrutiny. As a result, the DSBs may differentially affect the genomic integrity through spontaneous aberrant repair of damaged DNA or introduction of mutations by mechanisms of NHEJ, thus a sort of mosaicism can be present. Moreover, retrotransposon classes such as the proposed SVAs could be introduced at new genomic loci by RMDR to ligate the break ends. Changing the perspective, such retrotransposon could also be the cause of certain DSB hotspots, solely because the open chromatin and TE enriched regions are a preferential source for *de novo* integration controlled by the LINE-1 machinery. The assumption of DSB accumulation in open chromatin and active gene regions is also supported by already characterized DSB predictors, including DNase-seq, H3K4me3 and H3K27ac. In addition, motifs of DNA-binding proteins can indicate DSB hotspots and provide another predictor of DSB susceptible regions in the human brain. The DNA-binding proteins should be of special interest because their relation to the DNA damage response is well supported by literature. Since the various DSB hotspots pose a precursor for somatic mosaicism, different genes can be affected by accumulation of DSBs. Previously described genes of recurrent break clusters are detectable in human, too. Moreover *de novo* fragile genes shared in all brain regions are characterizable but also region specific genes. Based on their association with disease pathways, including glioblastoma, neurodegenerative diseases and mental disorders, they represent an interesting target for studying the effects of DSB prone genes in health and disease of the human brain.

5. Conclusion of the present thesis

Currently, several studies have provided evidence that the intraindividual mutational landscape is more extensive than previously thought. Genetic differences or somatic mutational events can occur during various developmental stages, including adolescence. Consequently, somatic mutations can be observed at multiple levels, extending beyond organ dissimilarity, meaning they can even be detected between genetically distinct cells within a single organ. The present thesis aimed to identify somatic differences in various brain regions through the analysis of retrotransposition and DNA double-strand breaks.

By implementing the RDA coupled with deep sequencing, multiple somatic *de novo* SVA and LINE-1 integrations were detected across all human brain regions analyzed. The RDA offers conclusive evidence of active retrotransposition of LINE-1 and SVA contributing to somatic mosaicism in the human brain. As a result, RDA-NGS detected *de novo* SVAs can be traced back to lineages of telencephalon and metencephalon. The *de novo* integrations of LINE-1 and SVA show a preference for genomic regions enriched in GC and transposable elements as well as neural-specific genes. Unlike LINE-1 integrations, which are frequently associated with 5´-truncations that are not captured by RDA, the SVA-RDA provides more informative sites. Therefore, when introducing TEs and studying their presence/absence as stable clade markers to explain somatic mosaicism in the human brain, the SVA-RDA should be the preferred method. Moreover, the SVA-RDA provides new opportunities to explain intra- and inter-individual variations and to reconstruct the phylogeny of cell lineages and is therefore also applicable in tumor research.

The BLISS-based detection of DSB hotspots revealed multiple genomic regions with shared occurrence of DSBs but also brain region specific breakage sites. The differences in quantity and localization of DSBs are suggested to be the result of differential exposure to oxidative stress, transcriptional patterns and chromatin state. Therefore, DSBs may differentially affect the genomic integrity through spontaneous aberrant repair of damaged DNA or the introduction of mutation by mechanisms of NHEJ, leading to conditions for somatic mosaicism. The present experiments provided further evidence that marks like DNase I sensitive sites, H3K4me3, H3K27ac and several DNA-binding motifs are good predictors of DSBs in the human brain. Literature based recurrent break cluster genes are detectable in human, and in addition to *de novo* fragile genes found across all brain regions, there are also region specific ones that can be characterized.

The presented experiments provided insights into the relation of DSBs and retrotransposons. First, *de novo* integrations frequently occur in TE-rich regions where DSBs also accumulate. Second, several flanking regions of *de novo* SVAs are DSB-rich.

Therefore, retrotransposon classes such as the proposed SVAs could be introduced into new genomic positions containing retrotransposons through RMDR by ligating break ends via microhomologies. Retrotransposons may also contribute to the formation of DSB hotspots, primarily because the open chromatin and TE enriched regions offer a favorable environment for *de novo* retrotransposon integration. The detected DSBs represent the current state of the individual, i.e. they are present in the mature brain and can correlate with the lineage-specific *de novo* SVA positions, suggesting a potential precursor of SVA mosaicism. However, the extent to which retrotransposons contribute to the cause or repair of DSBs needs to be further addressed in the future.

Lastly, the described alterations in genomes of the human brain regions can potentially contribute to neurological diseases. For example, different SVA patterns respectively presence or absence of SVAs at orthologous loci and related alterations in gene expression can be associated with Parkinson’s disease. DSB prone genes are associated with pathways involved in neurological and mental diseases like Alzheimer´s disease, schizophrenia and bipolar disorder. Moreover DNA-binding proteins that are frequently found at DSB sites are associated with DSB mechanisms and are important for genome stability and DNA damage response.

In this thesis, I have explored mechanisms that affect the genomic integrity of the mature human brain and provide new targets for studying the brain's health and disease. Moreover, the methodical procedures presented here have a broader applicability in various research fields. They offer valuable tools for conducting cell lineage tracing studies and analyzing DNA damage induction or responses, particularly in the context of neurological and cancer related diseases.

References

Acuna-Hidalgo, R., Bo, T., Kwint, M. P., Van De Vorst, M., Pinelli, M., Veltman, J. A., Hoischen, A., Vissers, L. E. L. M., & Gilissen, C. (2015). Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation. *The American Journal of Human Genetics*, *97*(1), 67–74. https://doi.org/10.1016/J.AJHG.2015.05.008

Alliot, F., Godin, I., & Pessac, B. (1999). Microglia derive from progenitors, originating from the yolk sac, and which proliferate in the brain. *Developmental Brain Research*, *117*(2), 145–152. https://doi.org/10.1016/S0165-3806(99)00113-3

Alt, F. W., & Schwer, B. (2018). DNA double-strand breaks as drivers of neural genomic change, function, and disease. *DNA Repair (Amst)*, *71*, 158–163. https://doi.org/10.1016/J.DNAREP.2018.08.019

Amemiya, H. M., Kundaje, A., & Boyle, A. P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports*, *9*, 9354. https://doi.org/10.1038/s41598-019-45839-z

Aneichyk, T., Hendriks, W. T., Yadav, R., Shin, D., Gao, D., Vaine, C. A., Collins, R. L., Domingo, A., Currall, B., Stortchevoi, A., Multhaupt-Buell, T., Penney, E. B., Cruz, L., Dhakal, J., Brand, H., Hanscom, C., Antolik, C., Dy, M., Ragavendran, A., … Talkowski, M. E. (2018). Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell*, *172*(5), 897-909.e21. https://doi.org/10.1016/J.CELL.2018.02.011

Bader, A. S., & Bushell, M. (2020). DNA:RNA hybrids form at DNA double-strand breaks in transcriptionally active loci. *Cell Death & Disease*, *11*, 280. https://doi.org/10.1038/s41419-020-2464-6

Baeken, M. W., Moosmann, B., & Hajieva, P. (2020). Retrotransposon activation by distressed mitochondria in neurons. *Biochemical and Biophysical Research Communications*, *525*(3), 570–575. https://doi.org/10.1016/j.bbrc.2020.02.106

Bailey, T. L., & Grant, C. E. (2021). SEA: Simple Enrichment Analysis of motifs. *BioRxiv*, *2021.08.23.457422*. https://doi.org/10.1101/2021.08.23.457422

Baillie, J. K., Barnett, M. W., Upton, K. R., Gerhardt, D. J., Richmond, T. A., De Sapio, F., Brennan, P., Rizzu, P., Smith, S., Fell, M., Talbot, R. T., Gustincich, S., Freeman, T. C., Mattick, J. S., Hume, D. A., Heutink, P., Carninci, P., Jeddeloh, J. A., & Faulkner, G. J. (2011). Somatic retrotransposition alters the genetic landscape of the human brain. *Nature*, *479*, 534–537. https://doi.org/10.1038/nature10531

Baranello, L., Kouzine, F., Wojtowicz, D., Cui, K., Przytycka, T. M., Zhao, K., & Levens, D. (2014). DNA Break Mapping Reveals Topoisomerase II Activity Genome-Wide. *International Journal of Molecular Sciences*, *15*(7), 13111–13122. https://doi.org/10.3390/IJMS150713111

Barnada, S. M., Isopi, A., Tejada-Martinez, D., Goubert, C., Patoori, S., Pagliaroli, L., Tracewell, M., & Trizzino, M. (2022). Genomic features underlie the co-option of SVA transposons as cis-regulatory elements in human pluripotent stem cells. *PLoS Genetics*, *18*(6). https://doi.org/10.1371/JOURNAL.PGEN.1010225

Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T. G., Fan, J., Garmiri, P., da Costa Gonzales, L. J., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., … Zhang, J. (2023). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, *51*(D1), D523–D531. https://doi.org/10.1093/NAR/GKAC1052

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Mikkelsen, T. S., & Thomson, J. A. (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology*, *28*(10), 1045–1048. https://doi.org/10.1038/NBT1010-1045

Bhat, A., Ghatage, T., Bhan, S., Lahane, G. P., Dhar, A., Kumar, R., Pandita, R. K., Bhat, K. M., Ramos, K. S., & Pandita, T. K. (2022). Role of Transposable Elements in Genome Stability: Implications for Health and Disease. *International Journal of Molecular Sciences*, *23*(14), 7802. https://doi.org/10.3390/IJMS23147802

Bizzotto, S., Dou, Y., Ganz, J., Doan, R. N., Kwon, M., Bohrson, C. L., Kim, S. N., Bae, T., Abyzov, A., Park, P. J., & Walsh, C. A. (2021). Landmarks of human embryonic development inscribed in somatic mutations. *Science (New York, N.Y.)*, *371*(6535), 1249–1253. https://doi.org/10.1126/SCIENCE.ABE1544

Black, E. M., & Giunta, S. (2018). Repetitive Fragile Sites: Centromere Satellite DNA as a Source of Genome Instability in Human Diseases. *Genes*, *9*(12). https://doi.org/10.3390/GENES9120615

Bodea, G. O., McKelvey, E. G. Z., & Faulkner, G. J. (2018). Retrotransposon-induced mosaicism in the neural genome. *Open Biology*, *8*(7), 180074. https://doi.org/10.1098/RSOB.180074

Bundo, M., Toyoshima, M., Okada, Y., Akamatsu, W., Ueda, J., Nemoto-Miyauchi, T., Sunaga, F., Toritsuka, M., Ikawa, D., Kakita, A., Kato, M., Kasai, K., Kishimoto, T., Nawa, H., Okano, H., Yoshikawa, T., Kato, T., & Iwamoto, K. (2014). Increased l1 retrotransposition in the neuronal genome in schizophrenia. *Neuron*, *81*(2), 306–313. https://doi.org/10.1016/J.NEURON.2013.10.053

Cai, X., Evrony, G. D., Lehmann, H. S., Elhosary, P. C., Mehta, B. K., Poduri, A., & Walsh, C. A. (2014). Single-Cell, Genome-wide Sequencing Identifies Clonal Somatic Copy-Number Variation in the Human Brain. *Cell Reports*, *8*(5), 1280–1289. https://doi.org/10.1016/J.CELREP.2014.07.043

Calses, P. C., Pham, V. C., Guarnaccia, A. D., Choi, M., Verschueren, E., Bakker, S. T., Pham, T. H., Hinkle, T., Liu, C., Chang, M. T., Kljavin, N., Bakalarski, C., Haley, B., Zou, J., Yan, C., Song, X., Lin, X., Rowntree, R., Ashworth, A., … Lill, J. R. (2023). TEAD Proteins Associate With DNA Repair Proteins to Facilitate Cellular Recovery From DNA Damage. *Molecular & Cellular Proteomics*, *22*(2), 100496. https://doi.org/10.1016/J.MCPRO.2023.100496

Campbell, I. M., Shaw, C. A., Stankiewicz, P., & Lupski, J. R. (2015). Somatic Mosaicism: Implications for Disease and Transmission Genetics. *Trends in Genetics : TIG*, *31*(7), 382–392. https://doi.org/10.1016/J.TIG.2015.03.013

Cannan, W. J., & Pederson, D. S. (2016). Mechanisms and Consequences of Double-strand DNA Break Formation in Chromatin. *Journal of Cellular Physiology*, *231*(1), 3–14. https://doi.org/10.1002/JCP.25048

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., & Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, *14*. https://doi.org/10.1186/1471-2105-14-128

Chen, H., Xu, K., Chen, B., Xie, Q., Sun, Y., Xu, X., Wang, J., Li, Y., Hu, P., Yue, S., Yu, G., Wang, J., Li, H., & Bo, X. (2023). Sensitivity of DNA double-strand break loci is coupled with spatial interaction density of chromatin. *11 July 2023, PREPRINT (Version 1) Available at Research Square*. https://doi.org/10.21203/RS.3.RS-3120397/V1

Ciccia, A., & Elledge, S. J. (2010). The DNA damage response: making it safe to play with knives. *Molecular Cell*, *40*(2), 179–204. https://doi.org/10.1016/J.MOLCEL.2010.09.019

Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, *10*, 691–703. https://doi.org/10.1038/nrg2640

Crosetto, N., Mitra, A., Silva, M. J., Bienko, M., Dojer, N., Wang, Q., Karaca, E., Chiarle, R., Skrzypczak, M., Ginalski, K., Pasero, P., Rowicka, M., & Dikic, I. (2013). Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nature Methods*, *10*, 361–365. https://doi.org/10.1038/nmeth.2408

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., & Davies, R. M. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/GIGASCIENCE/GIAB008

Denli, A. M., Narvaiza, I., Kerman, B. E., Pena, M., Benner, C., Marchetto, M. C. N., Diedrich, J. K., Aslanian, A., Ma, J., Moresco, J. J., Moore, L., Hunter, T., Saghatelian, A., & Gage, F. H. (2015). Primate-Specific ORF0 Contributes to Retrotransposon-Mediated Diversity. *Cell*, *163*(3), 583–593. https://doi.org/10.1016/J.CELL.2015.09.025

Deweese, J. E., & Osheroff, N. (2009). The DNA cleavage reaction of topoisomerase II: wolf in sheep's clothing. *Nucleic Acids Research*, *37*(3), 738–748. https://doi.org/10.1093/NAR/GKN937

Domcke, S., & Shendure, J. (2023). A reference cell tree will serve science better than a reference cell atlas. *Cell*, *186*(6), 1103–1114. https://doi.org/10.1016/j.cell.2023.02.016

Doyle, G. A., Crist, R. C., Karatas, E. T., Hammond, M. J., Ewing, A. D., Ferraro, T. N., Hahn, C. G., & Berrettini, W. H. (2017). Analysis of LINE-1 Elements in DNA from Postmortem Brains of Individuals with Schizophrenia. *Neuropsychopharmacology : Official Publication of the American College of Neuropsychopharmacology*, *42*, 2602–2611. https://doi.org/10.1038/NPP.2017.115

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B. K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., … Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/NATURE11247

Evrony, G. D., Lee, E., Mehta, B. K., Benjamini, Y., Johnson, R. M., Cai, X., Yang, L., Haseley, P., Lehmann, H. S., Park, P. J., & Walsh, C. A. (2015). Cell lineage analysis in human brain using endogenous retroelements. *Neuron*, *85*(1), 49–59. https://doi.org/10.1016/J.NEURON.2014.12.028

Falk, M., Lukášová, E., & Kozubek, S. (2008). Chromatin structure influences the sensitivity of DNA to γ-radiation. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, *1783*(12), 2398–2414. https://doi.org/10.1016/J.BBAMCR.2008.07.010

Flasch, D. A., Macia, Á., Sánchez, L., Ljungman, M., Heras, S. R., García-Pérez, J. L., Wilson, T. E., & Moran, J. V. (2019). Genome-wide de novo L1 Retrotransposition Connects Endonuclease Activity with Replication. *Cell*, *177*(4), 837-851.e28. https://doi.org/10.1016/J.CELL.2019.02.050

Fogwe, L. A., Reddy, V., & Mesfin, F. B. (2022). *Neuroanatomy, Hippocampus*. In StatPearls. StatPearls Publishing. https://pubmed.ncbi.nlm.nih.gov/29489273/ Accessed: 2023-04-20.

Fromer, M., Pocklington, A. J., Kavanagh, D. H., Williams, H. J., Dwyer, S., Gormley, P., Georgieva, L., Rees, E., Palta, P., Ruderfer, D. M., Carrera, N., Humphreys, I., Johnson, J. S., Roussos, P., Barker, D. D., Banks, E., Milanova, V., Grant, S. G., Hannon, E., … O'Donovan, M. C. (2014). De novo mutations in schizophrenia implicate synaptic networks. *Nature*, *506*, 179–184. https://doi.org/10.1038/nature12929

Fullard, J. F., Hauberg, M. E., Bendl, J., Egervari, G., Cirnaru, M. D., Reach, S. M., Motl, J., Ehrlich, M. E., Hurd, Y. L., & Roussos, P. (2018). An atlas of chromatin accessibility in the adult human brain. *Genome Research*, *28*(8), 1243–1252. https://doi.org/10.1101/gr.232488.117

Galtier, N., Gouy, M., & Gautier, C. (1996). SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Bioinformatics*, *12*(6), 543–548. https://doi.org/10.1093/BIOINFORMATICS/12.6.543

Gasior, S. L., Wakeman, T. P., Xu, B., & Deininger, P. L. (2006). The Human LINE-1 Retrotransposon Creates DNA Double-strand Breaks. *Journal of Molecular Biology*, *357*(5), 1383–1393. https://doi.org/10.1016/J.JMB.2006.01.089

Gianfrancesco, O., Geary, B., Savage, A. L., Billingsley, K. J., Bubb, V. J., & Quinn, J. P. (2019). The Role of SINE-VNTR-Alu (SVA) Retrotransposons in Shaping the Human Genome. *International Journal of Molecular Sciences*, *20*(23), 5977. https://doi.org/10.3390/IJMS20235977

Ginhoux, F., Greter, M., Leboeuf, M., Nandi, S., See, P., Gokhan, S., Mehler, M. F., Conway, S. J., Ng, L. G., Stanley, E. R., Samokhvalov, I. M., & Merad, M. (2010). Fate mapping analysis reveals that adult microglia derive from primitive macrophages. *Science*, *330*(6005), 841–845. https://doi.org/10.1126/science.1194637

Gittens, W. H., Johnson, D. J., Allison, R. M., Cooper, T. J., Thomas, H., & Neale, M. J. (2019). A nucleotide resolution map of Top2-linked DNA breaks in the yeast and human genome. *Nature Communications*, *10*, 4846. https://doi.org/10.1038/S41467-019-12802-5

Gouy, M., Guindon, S., & Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, *27*(2), 221–224. https://doi.org/10.1093/MOLBEV/MSP259

Griesius, S., O'Donnell, C., Waldron, S., Thomas, K. L., Dwyer, D. M., Wilkinson, L. S., Hall, J., Robinson, E. S. J., & Mellor, J. R. (2022). Reduced expression of the psychiatric risk gene DLG2 (PSD93) impairs hippocampal synaptic integration and plasticity. *Neuropsychopharmacology*, *47*, 1367–1378. https://doi.org/10.1038/s41386-022-01277-6

Grimwood, J., Gordon, L. A., Olsen, A., Terry, A., Schmutz, J., Lamerdin, J., Hellsten, U., Goodstein, D., Couronne, O., Tran-Gyamil, M., Aerts, A., Altherr, M., Ashworth, L., Bajorek, E., Black, S., Branscomb, E., Caenepeel, S., Carrano, A., Caoile, C., … Lucas, S. M. (2004). The DNA sequence and biology of human chromosome 19. *Nature*, *428*(6982), 529–535. https://doi.org/10.1038/NATURE02399

Grover, D., Mukerji, M., Bhatnagar, P., Kannan, K., Samir, K., & Brahmachari, S. K. (2004). Alu repeat analysis in the complete human genome: trends and variations with respect to genomic composition. *Bioinformatics (Oxford, England)*, *20*(6), 813–817. https://doi.org/10.1093/BIOINFORMATICS/BTH005

Gurok, U., Steinhoff, C., Lipkowitz, B., Ropers, H. H., Scharff, C., & Nuber, U. A. (2004). Gene Expression Changes in the Course of Neural Progenitor Cell Differentiation. *The Journal of Neuroscience*, *24*(26), 5982–6002. https://doi.org/10.1523/JNEUROSCI.0809-04.2004

Ha, H., Loh, J. W., & Xing, J. (2016). Identification of polymorphic SVA retrotransposons using a mobile element scanning method for SVA (ME-Scan-SVA). *Mobile DNA*, *7*, 15. https://doi.org/10.1186/S13100-016-0072-X

Haffner, M. C., Aryee, M. J., Toubaji, A., Esopi, D. M., Albadine, R., Gurel, B., Isaacs, W. B., Bova, G. S., Liu, W., Xu, J., Meeker, A. K., Netto, G., De Marzo, A. M., Nelson, W. G., & Yegnasubramanian, S. (2010). Androgen-induced TOP2B mediated double strand breaks and prostate cancer gene rearrangements. *Nature Genetics*, *42*(8), 668–675. https://doi.org/10.1038/NG.613

Hancks, D. C., Ewing, A. D., Chen, J. E., Tokunaga, K., & Kazazian, H. H. (2009). Exon-trapping mediated by the human retrotransposon SVA. *Genome Research*, *19*(11), 1983–1991. https://doi.org/10.1101/GR.093153.109

Hancks, D. C., Goodier, J. L., Mandal, P. K., Cheung, L. E., & Kazazian, H. H. (2011). Retrotransposition of marked SVA elements by human L1s in cultured cells. *Human Molecular Genetics*, *20*(17), 3386–3400. https://doi.org/10.1093/HMG/DDR245

Harris, L. W., Lockstone, H. E., Khaitovich, P., Weickert, C. S., Webster, M. J., & Bahn, S. (2009). Gene expression in the prefrontal cortex during adolescence: implications for the onset of schizophrenia. *BMC Medical Genomics*, *2*, 28. https://doi.org/10.1186/1755-8794-2-28

Hathaway, W. R., & Newton, B. W. (2022). *Neuroanatomy, Prefrontal Cortex*. In StatPearls. StatPearls Publishing. https://pubmed.ncbi.nlm.nih.gov/29763094/ Accessed: 2023-04-20.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell*, *38*(4), 576–589. https://doi.org/10.1016/J.MOLCEL.2010.05.004

Hitz, B. C., Lee, J.-W., Jolanki, O., Kagda, M. S., Graham, K., Sud, P., Gabdank, I., Strattan, J. S., Sloan, C. A., Dreszer, T., Rowe, L. D., Podduturi, N. R., Malladi, V. S., Chan, E. T., Davidson, J. M., Ho, M., Miyasato, S., Simison, M., Tanaka, F., … Cherry, J. M. (2023). The ENCODE Uniform Analysis Pipelines. *BioRxiv*, bioRxiv 2023.04.04.535623. https://doi.org/10.1101/2023.04.04.535623

Jacob-Hirsch, J., Eyal, E., Knisbacher, B. A., Roth, J., Cesarkas, K., Dor, C., Farage-Barhom, S., Kunik, V., Simon, A. J., Gal, M., Yalon, M., Moshitch-Moshkovitz, S., Tearle, R., Constantini, S., Levanon, E. Y., Amariglio, N., & Rechavi, G. (2018). Whole-genome sequencing reveals principles of brain retrotransposition in neurodevelopmental disorders. *Cell Research*, *28*, 187–203. https://doi.org/10.1038/cr.2018.8

Jakovcevski, I., Filipovic, R., Mo, Z., Rakic, S., & Zecevic, N. (2009). Oligodendrocyte Development and the Onset of Myelination in the Human Fetal Brain. *Frontiers in Neuroanatomy*, *3*, 5. https://doi.org/10.3389/NEURO.05.005.2009

Jimsheleishvili, S., & Dididze, M. (2022). *Neuroanatomy, Cerebellum*. In StatPearls. StatPearls Publishing. https://pubmed.ncbi.nlm.nih.gov/30844194/ Accessed: 2023-04-20.

Ju, B. G., Lunyak, V. V., Perissi, V., Garcia-Bassets, I., Rose, D. W., Glass, C. K., & Rosenfeld, M. G. (2006). A topoisomerase IIβ-mediated dsDNA break required for regulated transcription. *Science*, *312*(5781), 1798–1802. https://doi.org/10.1126/science.1127196

Kagda, M. S., Lam, B., Litton, C., Small, C., Sloan, C. A., Spragins, E., Tanaka, F., Whaling, I., Gabdank, I., Youngworth, I., Strattan, J. S., Hilton, J., Jou, J., Au, J., Lee, J.-W., Andreeva, K., Graham, K., Lin, K., Simison, M., … Hitz, B. C. (2023). Data navigation on the ENCODE portal. *ArXiv Genomics (q-Bio.GN)*, *version, v2*. https://doi.org/10.48550/arXiv.2305.00006

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A. M. M., Pletikos, M., Meyer, K. A., Sedmak, G., Guennel, T., Shin, Y., Johnson, M. B., Krsnik, Ž., Mayer, S., Fertuzinhos, S., Umlauf, S., Lisgo, S. N., Vortmeyer, A., … Šestan, N. (2011). Spatiotemporal transcriptome of the human brain. *Nature*, *478*(7370), 483–489. https://doi.org/10.1038/NATURE10523

Karolchik, D., Hinricks, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., & Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, *32*(Database issue), D493–D496. https://doi.org/10.1093/NAR/GKH103

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Research*, *12*(4), 656–664. https://doi.org/10.1101/GR.229202

Khan, A., & Mathelier, A. (2017). Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. *BMC Bioinformatics*, *18*(1), 287. https://doi.org/10.1186/S12859-017-1708-7

Kim, S., Yu, N. K., & Kaang, B. K. (2015). CTCF as a multifunctional protein in genome regulation and gene expression. *Experimental & Molecular Medicine*, *47*, e166. https://doi.org/10.1038/emm.2015.33

Kirov, G., Pocklington, A. J., Holmans, P., Ivanov, D., Ikeda, M., Ruderfer, D., Moran, J., Chambert, K., Toncheva, D., Georgieva, L., Grozeva, D., Fjodorova, M., Wollerton, R., Rees, E., Nikolov, I., Van De Lagemaat, L. N., Bayés, A., Fernandez, E., Olason, P. I., … Owen, M. J. (2012). De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Molecular Psychiatry*, *17*, 142–153. https://doi.org/10.1038/mp.2011.154

Kodali, S., Meyer-Nava, S., Landry, S., Chakraborty, A., Rivera-Mulia, J. C., & Feng, W. (2022). Epigenomic signatures associated with spontaneous and replication stress-induced DNA double strand breaks. *Frontiers in Genetics*, *13*, 907547. https://doi.org/10.3389/fgene.2022.907547

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., & Maayan, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, *44*(W1), W90–W97. https://doi.org/10.1093/NAR/GKW377

Kumar, R., Nagpal, G., Kumar, V., Usmani, S. S., Agrawal, P., & Raghava, G. P. S. (2019). HumCFS: a database of fragile sites in human chromosomes. *BMC Genomics*, *19*(Suppl 9). https://doi.org/10.1186/S12864-018-5330-5

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., McEwan, P., … International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*, 860–921. https://doi.org/10.1038/35057062

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/NMETH.1923

Lazarini, F., Gabellec, M. M., Moigneu, C., Chaumont, F. De, Olivo-Marin, J. C., & Lledo, P. M. (2014). Adult Neurogenesis Restores Dopaminergic Neuronal Loss in the Olfactory Bulb. *The Journal of Neuroscience*, *34*(43), 14430–14442. https://doi.org/10.1523/JNEUROSCI.5366-13.2014

Leduc, F., Faucher, D., Nkoma, G., Grégoire, M. C., Arguin, M., Wellinger, R. J., & Boissonneault, G. (2011). Genome-wide mapping of DNA strand breaks. *PloS One*, *6*(2), e17353. https://doi.org/10.1371/JOURNAL.PONE.0017353

Lensing, S. V., Marsico, G., Hänsel-Hertsch, R., Lam, E. Y., Tannahill, D., & Balasubramanian, S. (2016). DSBCapture: in situ capture and sequencing of DNA breaks. *Nature Methods*, *13*, 855–857. https://doi.org/10.1038/nmeth.3960

Lisitsyn, N., Lisitsyn, N., & Wigler, M. (1993). Cloning the differences between two complex genomes. *Science (New York, N.Y.)*, *259*(5097), 946–951. https://doi.org/10.1126/SCIENCE.8438152

Liu, D. D., He, J. Q., Sinha, R., Eastman, A. E., Toland, A. M., Morri, M., Neff, N. F., Vogel, H., Uchida, N., & Weissman, I. L. (2023). Purification and characterization of human neural stem and progenitor cells. *Cell*, *186*(6), 1179-1194.e15. https://doi.org/10.1016/J.CELL.2023.02.017

Lodato, M. A., Woodworth, M. B., Lee, S., Evrony, G. D., Mehta, B. K., Karger, A., Lee, S., Chittenden, T. W., D’Gama, A. M., Cai, X., Luquette, L. J., Lee, E., Park, P. J.,

& Walsh, C. A. (2015). Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science (New York, N.Y.)*, *350*(6256), 94–98. https://doi.org/10.1126/SCIENCE.AAB1785

Luo, S., Zheng, N., & Lang, B. (2022). ULK4 in Neurodevelopmental and Neuropsychiatric Disorders. *Frontiers in Cell and Developmental Biology*, *10*, 873706. https://doi.org/10.3389/FCELL.2022.873706

Luo, Y., Hitz, B. C., Gabdank, I., Hilton, J. A., Kagda, M. S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., Baymuradov, U. K., Graham, K., Litton, C., Miyasato, S. R., Strattan, J. S., Jolanki, O., Lee, J. W., Tanaka, F. Y., Adenekan, P., … Cherry, J. M. (2020). New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Research*, *48*(D1), D882–D889. https://doi.org/10.1093/NAR/GKZ1062

Martire, S., & Banaszynski, L. A. (2020). The roles of histone variants in fine-tuning chromatin organization and function. *Nature Reviews Molecular Cell Biology*, *21*, 522–541. https://doi.org/10.1038/s41580-020-0262-8

Mayer, R., Brero, A., von Hase, J., Schroeder, T., Cremer, T., & Dietzel, S. (2005). Common themes and cell type specific variations of higher order chromatin arrangements in the mouse. *BMC Cell Biology*, *6*, 44. https://doi.org/10.1186/1471-2121-6-44

McClintock, B. (1956). Controlling Elements and the Gene. *Cold Spring Harbor Symposia on Quantitative Biology*, *21*, 197–216. https://doi.org/10.1101/SQB.1956.021.01.017

McConnell, M. J., Lindberg, M. R., Brennand, K. J., Piper, J. C., Voet, T., Cowing-Zitron, C., Shumilina, S., Lasken, R. S., Vermeesch, J. R., Hall, I. M., & Gage, F. H. (2013). Mosaic Copy Number Variation in Human Neurons. *Science (New York, N.Y.)*, *342*(6158), 632–637. https://doi.org/10.1126/SCIENCE.1243472

Menini, A. (2010). *The Neurobiology of Olfaction*. CRC Press/Taylor & Francis. https://pubmed.ncbi.nlm.nih.gov/21882432/ Accessed: 2023-04-20.

Merkle, F. T., Fuentealba, L. C., Sanders, T. A., Magno, L., Kessaris, N., & Alvarez-Buylla, A. (2014). Adult neural stem cells in distinct microdomains generate previously unknown interneuron types. *Nature Neuroscience*, *17*, 207–214. https://doi.org/10.1038/NN.3610

Möhner, J., Scheuren, M., Woronzow, V., Schumann, S., & Zischler, H. (2023). RDA coupled with deep sequencing detects somatic SVA-retrotranspositions and mosaicism in the human brain. *Frontiers in Cell and Developmental Biology*, *11*(1201258). https://doi.org/10.3389/FCELL.2023.1201258

Morimoto, S., Tsuda, M., Bunch, H., Sasanuma, H., Austin, C., & Takeda, S. (2019). Type II DNA Topoisomerases Cause Spontaneous Double-Strand Breaks in Genomic DNA. *Genes*, *10*(11), 868. https://doi.org/10.3390/GENES10110868

Mourad, R., Ginalski, K., Legube, G., & Cuvier, O. (2018). Predicting double-strand DNA breaks using epigenome marks or DNA at kilobase resolution. *Genome Biology*, *19*(1), 34. https://doi.org/10.1186/S13059-018-1411-7

Ogawa, H., Horitani, K., Izumiya, Y., & Sano, S. (2022). Somatic Mosaicism in Biology and Disease. *Annual Review of Physiology*, *84*, 113–133. https://doi.org/10.1146/ANNUREV-PHYSIOL-061121-040048

Ono, R., Ishii, M., Fujihara, Y., Kitazawa, M., Usami, T., Kaneko-Ishino, T., Kanno, J., Ikawa, M., & Ishino, F. (2015). Double strand break repair by capture of retrotransposon sequences and reverse-transcribed spliced mRNA sequences in mouse zygotes. *Scientific Reports*, *5*, 12281. https://doi.org/10.1038/srep12281

Pace, J. K., & Feschotte, C. (2007). The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome Research*, *17*, 422–432. https://doi.org/10.1101/GR.5826307

Penzkofer, T., Jäger, M., Figlerowicz, M., Badge, R., Mundlos, S., Robinson, P. N., & Zemojtel, T. (2017). L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. *Nucleic Acids Research*, *45*(D1), D68–D73. https://doi.org/10.1093/NAR/GKW925

Pfaff, A. L., Bubb, V. J., Quinn, J. P., & Koks, S. (2021). Reference SVA insertion polymorphisms are associated with Parkinson’s Disease progression and differential gene expression. *Npj Parkinson’s Disease*, *7*, 44. https://doi.org/10.1038/s41531-021-00189-4

Piovesan, A., Pelleri, M. C., Antonaros, F., Strippoli, P., Caracausi, M., & Vitale, L. (2019). On the length, weight and GC content of the human genome. *BMC Research Notes*, *12*(1), 106. https://doi.org/10.1186/s13104-019-4137-z

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, *26*(6), 841–842. https://doi.org/10.1093/BIOINFORMATICS/BTQ033

Raiz, J., Damert, A., Chira, S., Held, U., Klawitter, S., Hamdorf, M., Löwer, J., Strätling, W. H., Löwer, R., & Schumann, G. G. (2012). The non-autonomous retrotransposon SVA is trans -mobilized by the human LINE-1 protein machinery. *Nucleic Acids Research*, *40*(4), 1666–1683. https://doi.org/10.1093/NAR/GKR863

Ramirez-Amaya, V., Marrone, D. F., Gage, F. H., Worley, P. F., & Barnes, C. A. (2006). Integration of New Neurons into Functional Neural Networks. *The Journal of Neuroscience*, *26*(47), 12237–12241. https://doi.org/10.1523/JNEUROSCI.2195-06.2006

Rehman, A., & Al Khalili, Y. (2022). *Neuroanatomy, Occipital Lobe*. In StatPearls. StatPearls Publishing. https://pubmed.ncbi.nlm.nih.gov/31335040/ Accessed: 2023-04-20.

Richardson, S. R., Morell, S., & Faulkner, G. J. (2014). L1 Retrotransposons and Somatic Mosaicism in the Brain. *Annual Review of Genetics*, *48*, 1–27. https://doi.org/10.1146/ANNUREV-GENET-120213-092412

Saayman, X., Graham, E., Nathan, W. J., Nussenzweig, A., & Esashi, F. (2023). Centromeres as universal hotspots of DNA breakage, driving RAD51-mediated recombination during quiescence. *Molecular Cell*, *83*(4), 523-538.e7. https://doi.org/10.1016/J.MOLCEL.2023.01.004

Salim, S. (2017). Oxidative Stress and the Central Nervous System. *The Journal of Pharmacology and Experimental Therapeutics*, *360*(1), 201–205. https://doi.org/10.1124/JPET.116.237503

Savage, A. L., Bubb, V. J., Breen, G., & Quinn, J. P. (2013). Characterisation of the potential function of SVA retrotransposons to modulate gene expression patterns. *BMC Evolutionary Biology*, *13*, 101. https://doi.org/10.1186/1471-2148-13-101

Scheuren, M., Möhner, J., & Zischler, H. (2023). R-loop landscape in mature human sperm: Regulatory and evolutionary implications. *Frontiers in Genetics*, *14*, 613. https://doi.org/10.3389/FGENE.2023.1069871/BIBTEX

Scully, R., Panday, A., Elango, R., & Willis, N. A. (2019). DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nature Reviews Molecular Cell Biology*, *20*, 698–714. https://doi.org/10.1038/s41580-019-0152-0

Sharma, V., Collins, L. B., Chen, T. H., Herr, N., Takeda, S., Sun, W., Swenberg, J. A., & Nakamura, J. (2016). Oxidative stress at low levels can induce clustered DNA lesions leading to NHEJ mediated mutations. *Oncotarget*, *7*(18), 25377–25390. https://doi.org/10.18632/ONCOTARGET.8298

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE*, *11*(10), e0163962. https://doi.org/10.1371/JOURNAL.PONE.0163962

Singh, S., Szlachta, K., Manukyan, A., Raimer, H. M., Dinda, M., Bekiranov, S., & Wang, Y. H. (2020). Pausing sites of RNA polymerase II on actively transcribed genes are enriched in DNA double-stranded breaks. *Journal of Biological Chemistry*, *295*(12), 3990–4000. https://doi.org/10.1074/JBC.RA119.011665

Skene, N. G., & Grant, S. G. N. (2016). Identification of vulnerable cell types in major brain disorders using single cell transcriptomes and expression weighted cell type enrichment. *Frontiers in Neuroscience*, *10*, 16. https://doi.org/10.3389/fnins.2016.00016

Smit, A., & Hubley, R. (2008). *RepeatModeler Open-1.0. 2008-2015 <http://www.repeatmasker.org>.*

Smit, A., Hubley, R., & Green, P. (2013). *RepeatMasker Open-4.0. 2013-2015 <http://www.repeatmasker.org>.*

Spalding, K. L., Bergmann, O., Alkass, K., Bernard, S., Salehpour, M., Huttner, H. B., Boström, E., Westerlund, I., Vial, C., Buchholz, B. A., Possnert, G., Mash, D. C., Druid, H., & Frisén, J. (2013). Dynamics of hippocampal neurogenesis in adult humans. *Cell*, *153*(6), 1219–1227. https://doi.org/10.1016/J.CELL.2013.05.002

Srikanta, D., Sen, S. K., Huang, C. T., Conlin, E. M., Rhodes, R. M., & Batzer, M. A. (2009). An alternative pathway for Alu retrotransposition suggests a role in DNA double-strand break repair. *Genomics*, *93*(3), 205–212. https://doi.org/10.1016/J.YGENO.2008.09.016

Stecher, G., Tamura, K., & Kumar, S. (2020). Molecular Evolutionary Genetics Analysis (MEGA) for macOS. *Molecular Biology and Evolution*, *37*(4), 1237–1239. https://doi.org/10.1093/MOLBEV/MSZ312

Stefanatos, R., & Sanz, A. (2018). The role of mitochondrial ROS in the aging brain. *FEBS Letters*, *592*(5), 743–758. https://doi.org/10.1002/1873-3468.12902

Stiles, J., & Jernigan, T. L. (2010). The Basics of Brain Development. *Neuropsychology Review*, *20*, 327–348. https://doi.org/10.1007/S11065-010-9148-4

Suarez, N. A., Macia, A., & Muotri, A. R. (2018). LINE-1 Retrotransposons in Healthy and Diseased Human Brain. *Developmental Neurobiology*, *78*(5), 434–455. https://doi.org/10.1002/DNEU.22567

Suberbielle, E., Sanchez, P. E., Kravitz, A. V., Wang, X., Ho, K., Eilertson, K., Devidze, N., Kreitzer, A. C., & Mucke, L. (2013). Physiological Brain Activity Causes DNA Double Strand Breaks in Neurons — Exacerbation by Amyloid-β. *Nature Neuroscience*, *16*(5), 613–621. https://doi.org/10.1038/NN.3356

Tamura, K., Stecher, G., & Kumar, S. (2021). MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution*, *38*(7), 3022–3027. https://doi.org/10.1093/MOLBEV/MSAB120

Tang, W., Mun, S., Joshi, A., Han, K., & Liang, P. (2018). Mobile elements contribute to the uniqueness of human genome with 15,000 human-specific insertions and 14 Mbp sequence increase. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, *25*(5), 521–533. https://doi.org/10.1093/DNARES/DSY022

Tchurikov, N. A., Alembekov, I. R., Klushevskaya, E. S., Kretova, A. N., Keremet, A. M., Sidorova, A. E., Meilakh, P. B., Chechetkin, V. R., Kravatskaya, G. I., & Kravatsky, Y. V. (2022). Genes Possessing the most Frequent DNA DSBs Are Highly Associated with Development and Cancers, and Essentially Overlap with the rDNA-Contacting Genes. *International Journal of Molecular Sciences*, *23*(13), 7201. https://doi.org/10.3390/ijms23137201

Tsompana, M., & Buck, M. J. (2014). Chromatin accessibility: a window into the genome. *Epigenetics & Chromatin*, *7*, 33. https://doi.org/10.1186/1756-8935-7-33

Turinetto, V., & Giachino, C. (2015). Multiple facets of histone variant H2AX: a DNA double-strand-break marker with several biological functions. *Nucleic Acids Research*, *43*(5), 2489–2498. https://doi.org/10.1093/NAR/GKV061

Ui, A., Chiba, N., & Yasui, A. (2020). Relationship among DNA double-strand break (DSB), DSB repair, and transcription prevents genome instability and cancer. *Cancer Science*, *111*(5), 1443–1451. https://doi.org/10.1111/CAS.14404

Upton, K. R., Gerhardt, D. J., Jesuadian, J. S., Richardson, S. R., Sánchez-Luque, F. J., Bodea, G. O., Ewing, A. D., Salvador-Palomeque, C., Van Der Knaap, M. S., Brennan, P. M., Vanderver, A., & Faulkner, G. J. (2015). Ubiquitous L1 Mosaicism in Hippocampal Neurons. *Cell*, *161*(2), 228–239. https://doi.org/10.1016/J.CELL.2015.03.026

Varga, T., & Aplan, P. D. (2005). Chromosomal aberrations induced by double strand DNA breaks. *DNA Repair*, *4*(9), 1038–1046. https://doi.org/10.1016/J.DNAREP.2005.05.004

Venkateshappa, C., Harish, G., Mahadevan, A., Srinivas Bharath, M. M., & Shankar, S. K. (2012). Elevated oxidative stress and decreased antioxidant function in the human hippocampus and frontal cortex with increasing age: Implications for neurodegeneration in Alzheimer's disease. *Neurochemical Research*, *37*(8), 1601–1614. https://doi.org/10.1007/s11064-012-0755-8

Versteeg, R., van Schaik, B. D. C., van Batenburg, M. F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H. J., & van Kampen, A. H. C. (2003). The Human Transcriptome Map Reveals Extremes in Gene Density, Intron Length, GC Content, and Repeat Pattern for Domains of Highly and Weakly Expressed Genes. *Genome Research*, *13*(9), 1998–2004. https://doi.org/10.1101/GR.1649303

Vilenchik, M. M., & Knudson, A. G. (2003). Endogenous DNA double-strand breaks: Production, fidelity of repair, and induction of cancer. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(22), 12871–12876. https://doi.org/10.1073/PNAS.2135498100

Wang, H., Xing, J., Grover, D., Hedges Kyudong Han, D. J., Walker, J. A., & Batzer, M. A. (2005). SVA Elements: A Hominid-specific Retroposon Family. *Journal of Molecular Biology*, *354*(4), 994–1007. https://doi.org/10.1016/J.JMB.2005.09.085

Wang, Y., Bae, T., Thorpe, J., Sherman, M. A., Jones, A. G., Cho, S., Daily, K., Dou, Y., Ganz, J., Galor, A., Lobon, I., Pattni, R., Rosenbluh, C., Tomasi, S., Tomasini, L., Yang, X., Zhou, B., Akbarian, S., Ball, L. L., … Abyzov, A. (2021). Comprehensive identification of somatic nucleotide variants in human brain tissue. *Genome Biology*, *22*, 92. https://doi.org/10.1186/s13059-021-02285-3

Wei, P. C., Chang, A. N., Kao, J., Du, Z., Meyers, R. M., Alt, F. W., & Schwer, B. (2016). Long Neural Genes Harbor Recurrent DNA Break Clusters in Neural Stem/Progenitor Cells. *Cell*, *164*(4), 644–655. https://doi.org/10.1016/J.CELL.2015.12.039

Wei, P. C., Lee, C. S., Du, Z., Schwer, B., Zhang, Y., Kao, J., Zurita, J., & Alt, F. W. (2018). Three classes of recurrent DNA break clusters in brain progenitors identified by 3D proximity-based break joining assay. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(8), 1919–1924. https://doi.org/10.1073/pnas.1719907115

Wright, C. F., Prigmore, E., Rajan, D., Handsaker, J., McRae, J., Kaplanis, J., Fitzgerald, T. W., FitzPatrick, D. R., Firth, H. V., & Hurles, M. E. (2019). Clinically-relevant postzygotic mosaicism in parents and children with developmental disorders in trio exome sequencing data. *Nature Communications*, *10*, 2985. https://doi.org/10.1038/S41467-019-11059-2

Xie, Z., Bailey, A., Kuleshov, M. V., Clarke, D. J. B., Evangelista, J. E., Jenkins, S. L., Lachmann, A., Wojciechowicz, M. L., Kropiwnicki, E., Jagodnik, K. M., Jeon, M., & Ma'ayan, A. (2021). Gene Set Knowledge Discovery with Enrichr. *Current Protocols*, *1*(3), e90. https://doi.org/10.1002/CPZ1.90

Yan, W. X., Mirzazadeh, R., Garnerone, S., Scott, D., Schneider, M. W., Kallas, T., Custodio, J., Wernersson, E., Li, Y., Gao, L., Federova, Y., Zetsche, B., Zhang, F., Bienko, M., & Crosetto, N. (2017). BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nature Communications*, *8*, 15058. https://doi.org/10.1038/ncomms15058

Yoo, T., Joshi, S., Prajapati, S., Cho, Y. S., Kim, J., Park, P. H., Bae, Y. C., Kim, E., & Kim, S. Y. (2022). A Deficiency of the Psychiatric Risk Gene DLG2/PSD-93 Causes Excitatory Synaptic Deficits in the Dorsolateral Striatum. *Frontiers in Molecular Neuroscience*, *15*, 938590. https://doi.org/10.3389/fnmol.2022.938590

Yoon, K. J., Vissers, C., Ming, G. li, & Song, H. (2018). Epigenetics and epitranscriptomics in temporal patterning of cortical neural progenitor competence. *The Journal of Cell Biology*, *217*(6), 1901–1914. https://doi.org/10.1083/JCB.201802117

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics (Oxford, England)*, *30*(5), 614–620. https://doi.org/10.1093/BIOINFORMATICS/BTT593

Zhang, X., Xiang, S., Zhang, Y., Liu, S., Lei, G., Hines, S., Wang, N., & Lin, H. (2023). In vitro study to identify ligand-independent function of estrogen receptor-α in suppressing DNA damage-induced chondrocyte senescence. *The FASEB Journal*, *37*(2), e22746. https://doi.org/10.1096/FJ.202201228R

Zhang, X., Zhang, R., & Yu, J. (2020). New Understanding of the Relevant Role of LINE-1 Retrotransposition in Human Disease and Immune Modulation. *Frontiers in Cell and Developmental Biology*, *8*, 657. https://doi.org/10.3389/fcell.2020.00657

Zhao, B., Wu, Q., Ye, A. Y., Guo, J., Zheng, X., Yang, X., Yan, L., Liu, Q. R., Hyde, T. M., Wei, L., & Huang, A. Y. (2019). Somatic LINE-1 retrotransposition in cortical neurons and non-brain tissues of Rett patients and healthy individuals. *PLoS Genetics*, *15*(4), e1008043. https://doi.org/10.1371/JOURNAL.PGEN.1008043

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., Benner, C., & Chanda, S. K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications*, *10*(1), 1523. https://doi.org/10.1038/S41467-019-09234-6

Supplement A

I. Reference list of RDC related genes of mouse NPCs (Wei PC et al., 2018; PMID: 29432181):

Npas3 Lsamp Nrxn1 Ptn Dgki Nfia Ctnna2 Sdk1 Grid2 Csmd1 Pard3b Prkg1 Maml2 Csmd3 Tcf4 Lrp1b Cdh13 Grik2 Nrxn3 Gpc6 Ctnnd2 Rbfox1 Cadm2 Park2 Pacrg Qk Agpat4 Map3k4 Slc22a3 Igf2r Sema6d Nlgn1 Auts2 Opcml Ntm Pcdh9 Ptk2 Ago2 Trappc9 Sox5 Sox6 Pik3c2a Wwox Prim2 Dst Nkain2 Anks1b Apaf1 Gphn Mctp1 2210408I21Rik Fam172a Arhgap26 Fgf1 Dcc Lrrc4c Ptprt Zhx3 Chd6 Astn2 Inpp4b Grip1 Fgf12 Fhod3 AW554918 Celf4 Macrod2 Naaladl2 Tnik Magi2 Ccser1 Nav2 Rbms3 Bai3 Erbb4 Rev3l Nav3 Nxn Abr Wnk2 Cenpp Phf2 Fam120a Fhit Kcnma1 Nrg3 Fgf14 Slc1a3 Ptprm Dlgap1 Lrba Dclk2 Vav3 Dpp6 Lphn3 Creb5 Stim1 Clpb Nup98 Dock1 Pid1 Dner Trip12 Agap1 Nckap5 Srgap2 Utrn Hdac9 Dgkb Mdga2 Mark3 Adarb2 Dip2c Chrm3 Ptprg Vcl Adk Kat6b Zmiz1 Wdr70 2410089E03Rik Nipbl Oxr1 Trappc9 Zbtb20 Pde10a Prkce Cdh4 Dclk1 Nbea Dpyd Ptprd Csmd2 Slc4a4 Exoc4 Chchd6 Cntn4 Gabrg3 Gabra5 Gabrb3 Large1 Nr3c2 Arhgap10 Kirrel3 Rora Tcf12 Dmd Il1rapl1 Pcdh11x

II. List of RDC genes shared in mouse and human:

SOX5 CSMD1 CHRM3 DIP2C NRXN1 CADM2 CENPP AUTS2 DGKI MDGA2 SOX6 DGKB CSMD3 NAALADL2 PARD3B SEMA6D PTK2 DST NR3C2 DMD RBFOX1 PACRG FGF14 EXOC4 NAV2 MAP3K4 VAV3 PID1 ASTN2 PTPRG GRID2 GRIP1 NBEA DOCK1 CHD6 NPAS3 PTPRD LRRC4C CTNND2 CCSER1

III. List of *de novo* RDC genes:

ACYP2 AGBL4 AKAP13 AKAP6 ANKRD26P1 ARHGEF18 ART1 C12orf40 CDC27 COPG2 CPA6 CRTC3 DCK DLG2 FAM157A FAM230F FBXW11 FSCN1 HFM1 LDOC1 LINC00486 LMO7 LOC101930421 LOC401478 LRCH3 MIR2110 MIR6788 NAGPA-AS1 NLRC4 ODF2 PALM2-AKAP2 PGAM1P5 PIK3CB REEP2 RMDN1 RPL29P2 SCAT8 SFMBT2 SNORD168 SNX16 SPHKAP SULF1 TNK2-AS1 ULK4 VCX3B WDR27 ZNF652 ZNF732

IV. List of region-specific RDC genes

Olfactory bulb:

ANO2 POLR3K COPE EXOC4 PSMB5

Calcarine sulcus:

RIT2 XPR1 POLR3B EFNA2 MUSK KCND3 SLC28A3 FAF1 TNRC18 ZMYND11 ADAM23 SEL1L3 SLC18A1 KDM4C SCLT1 GRIA4 DRD5 CHRNA10 NPSR1 NDEL1 TUBA1A SLC24A2 KLHL15 PHF14

Cerebellum:

ZNF521 SRSF5 HMCN1 GPM6A COX7B2 CLASP2 JHY TRPV1 PDE11A SYT1 UBE2H ERICH1 SSH2 WASHC2A IST1 STAU1 HIST3H2BB VAV3 SYNJ2 PTPRG GRID2 FSTL4 PIP4K2A TOP2B GFM2 BAG4 NXPE3 SATB1 RETREG1 PTPRZ1 FLRT3 CEP112 NTNG1 PPY GNAQ ABCA13 YWHAZ TPGS1 CDH18 SEPHS2 GABRR2 MAB21L2 FRMD5 STAG2 THSD7B OGFRL1 DMXL2 PRAG1

Hippocampus:

MRPL3 PPP1R14C NBPF12 TMEM135 POU2F3 KLHL14 KATNAL1 EIF2S3B ATXN7L1 APBA1 VIT LRRN1 SOD2 SLC5A11 SLC19A1 IP6K2 RFC1 SPG11 THAP8 SAMD12 DOCK1 TGIF1 HCN1 PARD3 NPAS3 ATG10 MTERF4 DGUOK LRRC4C LOC100132202 GRIN2B PTPRD HPX GAP43 MTMR12 NLN SNRPD1 GSG1L GRIP1 BCAT1 ADGRB3 ARHGAP24 NBPF8 ITGB1 CHD6 PHF3

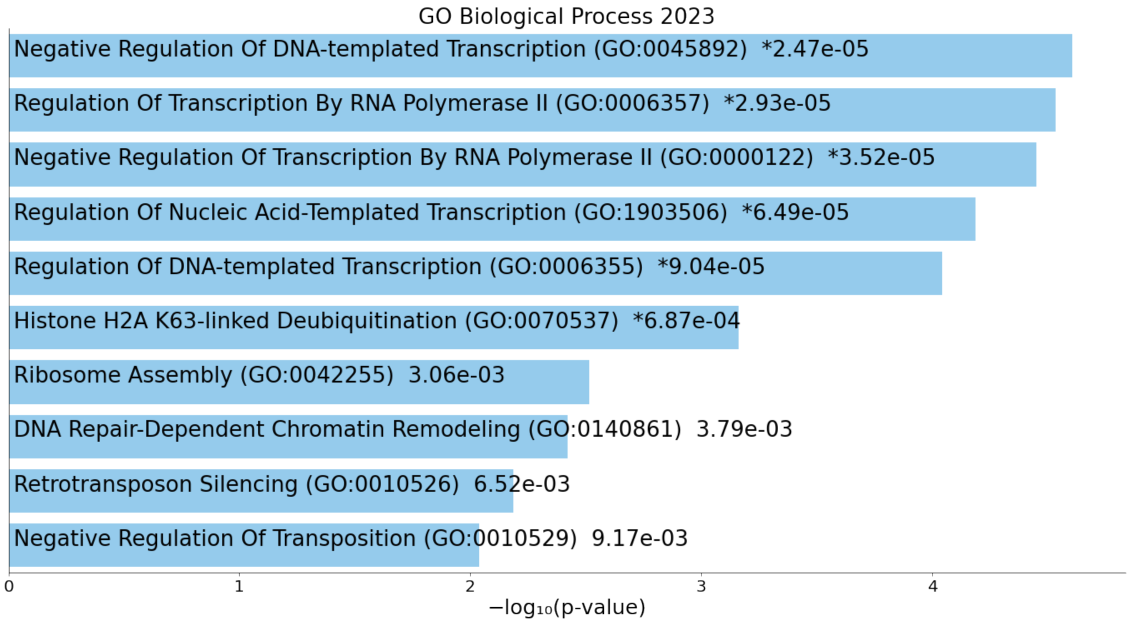
Prefrontal cortex:

PSIP1 LYPD1 GSK3B NDUFA12 PTRH2 SERPINB5 UBE4A SEMA6D ENPP1 DENND6A VPS13D EFCAB11 HECW2 OXNAD1 KYAT3 CPEB3 IQGAP2 MAP9 RUFY3 FAM120C ASPM ANKRD49 TNKS NUF2 TRPM7 INTS4 MTFR1 GEMIN2 CCT8 CCNH VLDLR NDC1 RFX7 OARD1 ELP4 COX17 STRN AIMP1 MAP3K13 RNF168 SPTLC2 COX7C NUBP2 TTBK2 FAM126B VPS4B TCTN3 PCNP DCTN4 CCDC90B MRPL32 ATP5MC1 PABPN1 GTPBP8 CPEB4 SELENOK EIF2A FGD4 NDUFA5 MTPN PTPN4 CNST LMBRD1 TPRKB TXNDC16 KIFAP3 C12orf60 SMNDC1 MRPS28 GLO1 BECN1 PSMA5 RABGGTB CSDE1 CYCS R3HDM1 HSPA13 NGDN XRCC5 ERP44 TENM2 TRIM23 RNF20 TNRC6A WHAMM GLDC KCNN2 TCERG1 UGGT1 ABHD2 EHBP1 ZFYVE1 APLP2 IPMK NXPH1 DDOST ALG10 PIP4P1 SIRT6 THG1L TRMO CHD1 POP4 DNAJC16 SNRPC ORMDL2 RBM7 NOA1 TRMT1L AZIN1 NETO2 PPP2R5B ECD TAF11 GTPBP1 DIAPH3 PGAP2 SEL1L NOP58 HEXIM2 ATAD1 TUBE1 TIMM9 EIF2S1 ZBTB3 SCFD2 MRPL45 MAP2 RNF34 DNAJC1 MRPL34 NDUFS4 APTX METTL14 PRPF38B PPM1B FRYL LIMCH1 RPH3A SFXN5 RBFOX1 BTRC CAMTA1 MRPS18A SMURF1 TRAP1 ARAP1 NDUFB5 RPS6KB2 NTRK2 DDX5 ACAD9 PDE12 MRPS21 MRPL54 SATB2 FAM72A KIF20A RBBP5 SLIT3 PRMT1 KIF5C CCT5 EIF3G ACTR6 PSMB3 CCZ1B HIF1A FGF14 MAPK13 ALS2 APEX1 CSTF2T PCNA PNKD TP53INP1 TESMIN CLIC4 NRG2 L3MBTL2 MAST2 FKBP15 ZNF283 SPICE1 CNTN6 ZNF138 OCIAD1 ZNF689 CIC ZGRF1 ESYT2 ZNF684 TRIM4 GLIDR ZNF527 ZNF286A ZNF548 FN1 CHTOP SUPT6H ZNF774 CBX6 CCNB2

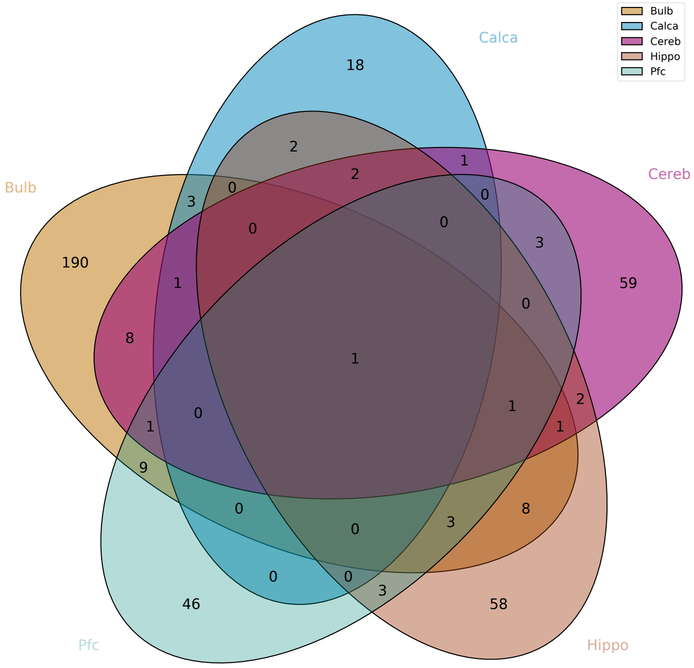
Supplementary Table 1: Shared Motif information

RANK	ALT_ID	CONSENSUS	TP	TP%	FP	FP%	ENR_RATIO	SCORE_THR	PVALUE	EVALUE	QVALUE
Pfc											
1	ZNF460	GCCTCMGCCTCCCRAG	673	29.49	51	2.23	13	9.9	9.14E-140	1.89E-136	1.13E-136
2	ZNF135	CCTCGACCTCCYRR	723	31.68	196	8.59	3.68	6.7	6.99E-72	1.45E-68	4.32E-69
6	ESR1	ARGGTCACSRTGACCTK	222	9.73	27	1.18	7.96	1.9	1.33E-39	2.76E-36	2.74E-37
24	MAFK	NGMTGACTCAGCMNH	517	22.66	222	9.73	2.32	8.5	2.81E-28	5.82E-25	1.45E-26
55	ZNF85	DHDGAGATTACWKCAK	755	33.09	440	19.28	1.71	5.5	3.23E-20	6.69E-17	7.25E-19
295	PROX1	YAAGACGYCTTA	547	23.97	390	17.09	1.4	2.4	1.63E-07	3.37E-04	6.82E-07
313	Zfx	SSSGCCBVGGCCTS	1296	56.79	1060	46.45	Zfx	4	6.32E-07	1.31E-03	2.48E-06
438	TEAD1	NNACATTCCAGSN	1221	53.51	1046	45.84	1.17	4.8	1.28E-04	2.65E-01	3.61E-04
476	Stat5b	NNNTTCCCAGAANNN	17	0.74	2	0.09	6	15	3.64E-04	7.54E-01	9.45E-04
Cereb											
1	ZNF460	GCCTCMGCCTCCCRAG	139	13.19	3	0.28	35	14	8.92E-38	1.85E-34	1.18E-34
2	ZNF135	CCTCGACCTCCYRR	158	14.99	28	2.66	5.48	9.3	1.74E-23	3.61E-20	1.15E-20
13	ESR1	ARGGTCACSRTGACCTK	61	5.79	10	0.95	5.64	3.8	2.36E-10	4.88E-07	2.40E-08
24	MAFK	NGMTGACTCAGCMNH	89	8.44	32	3.04	2.73	11	1.09E-07	2.26E-04	6.02E-06
18	ZNF85	DHDGAGATTACWKCAK	505	47.91	343	32.54	1.47	3.9	1.55E-08	3.20E-05	1.13E-06
54	PROX1	YAAGACGYCTTA	113	10.72	59	5.6	1.9	4.6	2.37E-05	4.90E-02	5.79E-04
5	Zfx	SSSGCCBVGGCCTS	483	45.83	287	27.23	1.68	3.9	8.53E-13	1.77E-09	2.25E-10
127	TEAD1	NNACATTCCAGSN	616	58.44	507	48.1	1.21	4.7	6.52E-04	1.35E+00	6.76E-03
50	Stat5b	NNNTTCCCAGAANNN	137	13	75	7.12	1.82	9.6	1.26E-05	2.61E-02	3.33E-04
Bulb											
1	ZNF460	GCCTCMGCCTCCCRAG	81	21.54	9	2.39	8.2	7.9	7.50E-16	1.55E-12	1.50E-12
2	ZNF135	CCTCGACCTCCYRR	47	12.5	2	0.53	16	14	2.40E-12	4.97E-09	2.40E-09
14	ESR1	ARGGTCACSRTGACCTK	27	7.18	3	0.8	7	3.7	4.45E-06	9.20E-03	6.35E-04
55	MAFK	NGMTGACTCAGCMNH	56	14.89	30	7.98	1.84	9.1	3.55E-03	7.34E+00	1.29E-01

5	ZNF85	DHDGAGATTACWKCAK	54	14.36	9	2.39	5.5	9.8	3.37E-09	6.98E-06	1.35E-06
37	PROX1	YAAGACGYCTTA	23	6.12	6	1.6	3.43	7.7	1.20E-03	2.49E+00	6.49E-02
15	Zfx	SSSGCCBVGGCCTS	100	26.6	46	12.23	2.15	7.2	5.21E-06	1.08E-02	6.94E-04
30	TEAD1	NNACATTCCAGSN	173	46.01	116	30.85	1.49	6.5	5.44E-04	1.13E+00	3.52E-02
12	Stat5b	NNNTTCCCAGAANNN	58	15.43	18	4.79	3.11	11	2.58E-06	5.33E-03	4.29E-04
Hippo											
1	ZNF460	GCCTCMGCCTCCCRAG	256	16.08	8	0.5	28.6	12	2.88E-65	5.96E-62	3.92E-62
2	ZNF135	CCTCGACCTCCYRR	232	14.57	12	0.75	17.9	10	3.92E-54	8.12E-51	2.67E-51
11	ESR1	ARGGTCACSRTGACCTK	109	6.85	15	0.94	6.87	0.4	5.15E-19	1.07E-15	6.38E-17
58	MAFK	NGMTGACTCAGCMNH	168	10.55	81	5.09	2.06	10	2.17E-08	4.50E-05	5.09E-07
12	ZNF85	DHDGAGATTACWKCAK	228	14.32	80	5.03	2.83	8.8	8.26E-18	1.71E-14	9.37E-16
50	PROX1	YAAGACGYCTTA	58	3.64	11	0.69	4.92	9.3	4.11E-09	8.52E-06	1.12E-07
9	Zfx	SSSGCCBVGGCCTS	745	46.8	428	26.88	1.74	2.9	1.47E-20	3.04E-17	2.09E-18
53	TEAD1	NNACATTCCAGSN	672	42.21	475	29.84	1.41	6.1	4.73E-09	9.80E-06	1.21E-07
10	Stat5b	NNNTTCCCAGAANNN	235	14.76	75	4.71	3.11	10	1.54E-20	3.18E-17	2.09E-18
Calca											
4	ZNF460	GCCTCMGCCTCCCRAG	150	16.69	47	5.23	3.15	5.2	4.84E-14	1.00E-10	1.92E-11
12	ZNF135	CCTCGACCTCCYRR	149	16.57	59	6.56	2.5	5.4	1.80E-10	3.73E-07	2.38E-08
15	ESR1	ARGGTCACSRTGACCTK	35	3.89	1	0.11	18	5	5.37E-10	1.11E-06	5.69E-08
62	MAFK	NGMTGACTCAGCMNH	54	6.01	19	2.11	2.75	11	2.53E-05	5.23E-02	6.47E-04
20	ZNF85	DHDGAGATTACWKCAK	82	9.12	23	2.56	3.46	10	2.97E-09	6.14E-06	2.36E-07
104	PROX1	YAAGACGYCTTA	179	19.91	120	13.35	1.49	2	3.83E-04	7.92E-01	5.84E-03
13	Zfx	SSSGCCBVGGCCTS	207	23.03	100	11.12	2.06	6.1	4.92E-10	1.02E-06	5.68E-08
9	TEAD1	NNACATTCCAGSN	305	33.93	162	18.02	1.88	7.2	1.75E-11	3.63E-08	3.10E-09
3	Stat5b	NNNTTCCCAGAANNN	185	20.58	62	6.9	2.95	9.2	1.06E-15	2.19E-12	5.60E-13



Supplementary Figure 1: Enrichr GO pathways of 83 prefrontal cortex genes (genes with DSB hotspot at promoter region that are exclusive in prefrontal cortex on Chr. 19). Enrichr is available on https://maayanlab.cloud/Enrichr/; accessed on 11 July 2023 (E. Y. Chen et al., 2013; Kuleshov et al., 2016; Xie et al., 2021).



Supplementary Figure 2: Venn diagram of DSB enriched *de novo* SVA flanks for each brain region analyzed.

49 Spadina Ave. Suite 200
Toronto ON M5V 2J1 Canada
www.biorender.com

Confirmation of Publication and Licensing Rights

August 11th, 2023
Science Suite Inc.

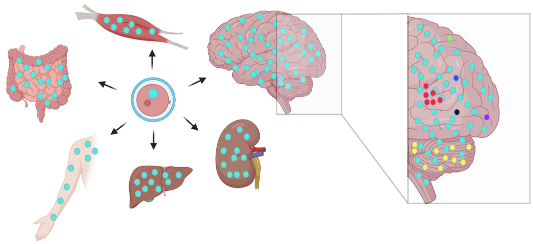
Subscription:	*Student Plan*
Agreement number:	*PH25PTVS4Q*
Journal name:	*Gutenberg Open Science - UB JGU Mainz*

To whom this may concern,

This document is to confirm that Jonas Möhner has been granted a license to use the BioRender content, including icons, templates and other original artwork, appearing in the attached completed graphic pursuant to BioRender's Academic License Terms. This license permits BioRender content to be sublicensed for use in journal publications.

All rights and ownership of BioRender content are reserved by BioRender. All completed graphics must be accompanied by the following citation: "Created with BioRender.com".

BioRender content included in the completed graphic is not licensed for any commercial uses beyond publication in a journal. For any commercial use of this figure, users may, if allowed, recreate it in BioRender under an Industry BioRender Plan.



For any questions regarding this document, or other questions about publishing with BioRender refer to our BioRender Publication Guide, or contact BioRender Support at support@biorender.com.

49 Spadina Ave. Suite 200
Toronto ON M5V 2J1 Canada
www.biorender.com

Confirmation of Publication and Licensing Rights

August 11th, 2023
Science Suite Inc.

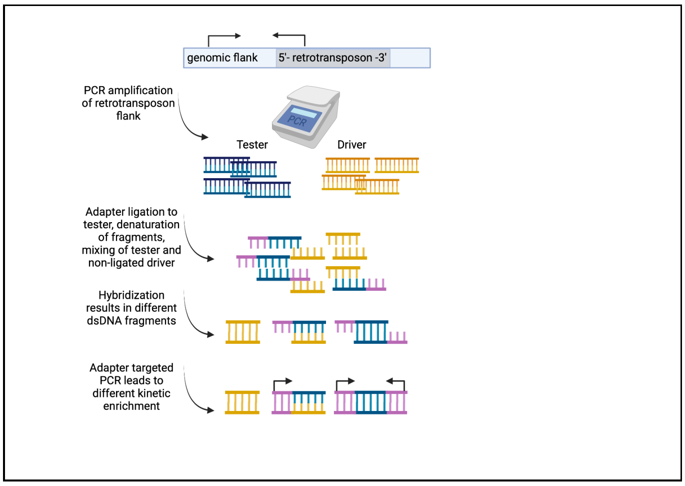
Subscription:	*Student Plan*
Agreement number:	*NN25PU1YGN*
Journal name:	*Gutenberg Open Science - UB JGU Mainz*

To whom this may concern,

This document is to confirm that Jonas Möhner has been granted a license to use the BioRender content, including icons, templates and other original artwork, appearing in the attached completed graphic pursuant to BioRender's Academic License Terms. This license permits BioRender content to be sublicensed for use in journal publications.

All rights and ownership of BioRender content are reserved by BioRender. All completed graphics must be accompanied by the following citation: “Created with BioRender.com”.

BioRender content included in the completed graphic is not licensed for any commercial uses beyond publication in a journal. For any commercial use of this figure, users may, if allowed, recreate it in BioRender under an Industry BioRender Plan.


genomic flank
5'- retrotransposon -3'
PCR amplification of retrotransposon flank
PCR
Tester
Driver
Adapter ligation to tester, denaturation of fragments, mixing of tester and non-ligated driver
Hybridization results in different dsDNA fragments
Adapter targeted PCR leads to different kinetic enrichment

For any questions regarding this document, or other questions about publishing with BioRender refer to our BioRender Publication Guide, or contact BioRender Support at support@biorender.com.

49 Spadina Ave. Suite 200
Toronto ON M5V 2J1 Canada
www.biorender.com

Confirmation of Publication and Licensing Rights

August 11th, 2023
Science Suite Inc.

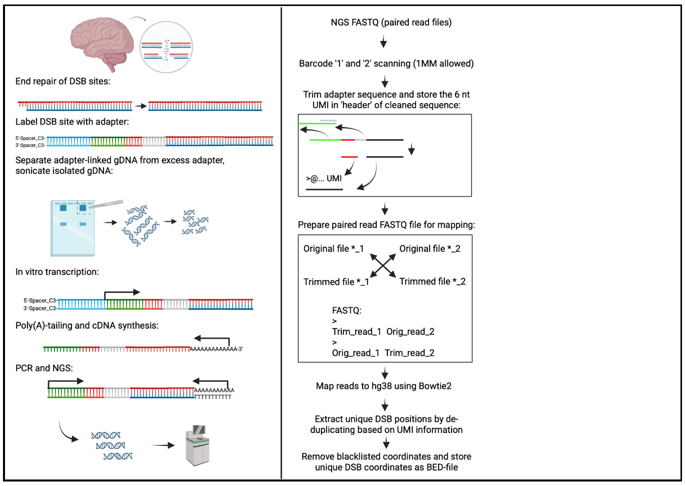
Subscription: *Student Plan*
Agreement number: *BB25PU5IN0*
Journal name: *Gutenberg Open Science - UB JGU Mainz*

To whom this may concern,

This document is to confirm that Jonas Möhner has been granted a license to use the BioRender content, including icons, templates and other original artwork, appearing in the attached completed graphic pursuant to BioRender's Academic License Terms. This license permits BioRender content to be sublicensed for use in journal publications.

All rights and ownership of BioRender content are reserved by BioRender. All completed graphics must be accompanied by the following citation: "Created with BioRender.com".

BioRender content included in the completed graphic is not licensed for any commercial uses beyond publication in a journal. For any commercial use of this figure, users may, if allowed, recreate it in BioRender under an Industry BioRender Plan.


End repair of DSB sites:
Label DSB site with adapter:
5'-Spacer_C3-
3'-Spacer_C3-
Separate adapter-linked gDNA from excess adapter, sonicate isolated gDNA:
In vitro transcription:
5'-Spacer_C3-
3'-Spacer_C3-
Poly(A)-tailing and cDNA synthesis:
AAAAAAAAAAAAA-3'
PCR and NGS:
AAAAAAAAAAA
TTTTTTTTTT
NGS FASTQ (paired read files)
Barcode '1' and '2' scanning (1MM allowed)
Trim adapter sequence and store the 6 nt UMI in 'header' of cleaned sequence:
>@... UMI
Prepare paired read FASTQ file for mapping:
Original file *_1
Original file *_2
Trimmed file *_1
Trimmed file *_2
FASTQ:
>
Trim_read_1 Orig_read_2
>
Orig_read_1 Trim_read_2
Map reads to hg38 using Bowtie2
Extract unique DSB positions by de-duplicating based on UMI information
Remove blacklisted coordinates and store unique DSB coordinates as BED-file

For any questions regarding this document, or other questions about publishing with BioRender refer to our BioRender Publication Guide, or contact BioRender Support at support@biorender.com.

49 Spadina Ave. Suite 200
Toronto ON M5V 2J1 Canada
www.biorender.com

Confirmation of Publication and Licensing Rights

August 11th, 2023
Science Suite Inc.

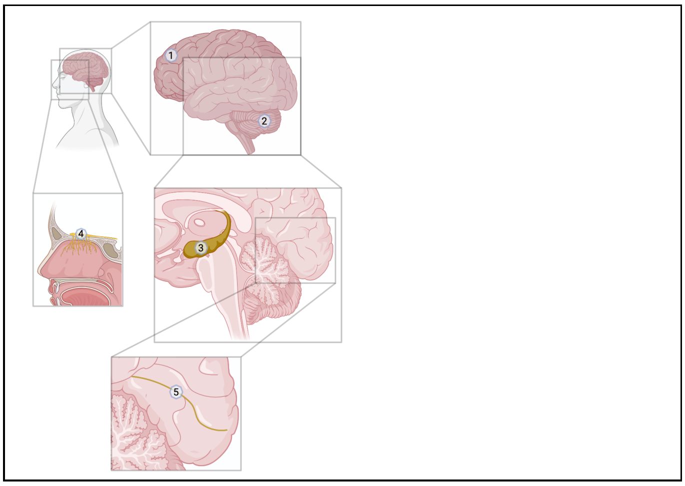
Subscription:	*Student Plan*
Agreement number:	*SM25PTWZDX*
Journal name:	*Gutenberg Open Science - UB JGU Mainz*

To whom this may concern,

This document is to confirm that Jonas Möhner has been granted a license to use the BioRender content, including icons, templates and other original artwork, appearing in the attached completed graphic pursuant to BioRender's Academic License Terms. This license permits BioRender content to be sublicensed for use in journal publications.

All rights and ownership of BioRender content are reserved by BioRender. All completed graphics must be accompanied by the following citation: “Created with BioRender.com”.

BioRender content included in the completed graphic is not licensed for any commercial uses beyond publication in a journal. For any commercial use of this figure, users may, if allowed, recreate it in BioRender under an Industry BioRender Plan.



For any questions regarding this document, or other questions about publishing with BioRender refer to our BioRender Publication Guide, or contact BioRender Support at support@biorender.com.

Author contributions

RDA coupled deep sequencing detects somatic SVA-retrotranspositions and mosaicism in the human brain (chapter 3.1):

The manuscript was written by Jonas Möhner and Prof. Dr. Hans Zischler. The Concept of *de novo* SVA enrichment via RDA was created by Hans Zischler and Jonas Möhner established the experimental procedures. Valentina Woronzow provided further knowledge during experimental validations. Sven Schumann isolated the human tissue samples. The bioinformatic pipeline was created by Jonas Möhner and analysis of NGS data was also carried out by Jonas Möhner. First interpretation of bioinformatic results were interpretated by Jonas Möhner. Prof. Dr. Hans Zischler, Maurice Scheuren and Sven Schumann provided further knowledge for interpretation of data. The graphics were generated by Jonas Möhner and validated by Maurice Scheuren.

Ethics statement

Human tissue samples were obtained as part of the body donation program of the Institute of Anatomy, University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany. The people donated their body voluntarily for medical education and research and the present study was conducted within the parameters of the written permission we received from the body donor during lifetime. This human research on post-mortem tissue was reviewed and approved by the ethics committee of Landesärztekammer Rheinland-Pfalz, Mainz, Germany (24/05/2022; Ref.# 2022-16488).

Danksagung

Eidesstattliche Versicherung

Hiermit versichere ich, Jonas Möhner, dass ich die vorgelegte Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Ich habe oder hatte die jetzt als Dissertation vorgelegte Arbeit nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht. Auch habe oder hatte ich die vorgelegte Dissertation oder Teile der Arbeit nicht als Dissertation bei einer anderen Fakultät oder einem anderen Fachbereich eingereicht.

Ort, Datum

(Jonas Möhner)

Lebenslauf