Characterisation of the Function of a SINE-VNTR-*Alu* Retrotransposon to Modulate Isoform Expression at the *MAPT* Locus

Alexander Fröhlich (✉ Alexander.Froehlich@liverpool.ac.uk)

University of Liverpool <https://orcid.org/0000-0001-9083-7485>

Abigail L. Pfaff

Murdoch University

Vivien J. Bubb

University of Liverpool Institute of Translational Medicine

Sulev Koks

Murdoch University

John P. Quinn

University of Liverpool Institute of Translational Medicine

Research

Keywords: Transposable elements, MAPT, tau, Parkinson disease, ALS, FTD, motor neuron disease, SVA

Posted Date: November 16th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1065189/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Frontiers in Molecular Neuroscience on March 9th, 2022. See the published version at <https://doi.org/10.3389/fnmol.2022.815695>.

Abstract

Background:

SINE-VNTR-*Alu* retrotransposons represent one class of transposable elements which contribute to the regulation and evolution of the primate genome and have the potential to be involved in genetic instability and disease progression. However, these polymorphic elements have not been extensively analysed when addressing the missing heritability of neurodegenerative diseases, including Parkinson's disease (PD) and amyotrophic lateral sclerosis (ALS). SVA_67, a retrotransposon insertion polymorphism, is located in a 1.8 Mb region of high linkage disequilibrium, called the *MAPT* locus, which is known to contribute to increased risk of developing PD, frontotemporal dementia and other tauopathies. This study was conducted to investigate the role of SVA_67 in directing differential gene expression at this locus by characterising the impact of SVA_67 allele dosage on isoform expression of several genes in the *MAPT* locus.

Methods:

In this study, transcriptomic data from both the Parkinson's Progression Markers Initiative and New York Genome Center Consortium Target ALS cohort were used. The Parkinson's data was from gene expression in the blood and the ALS data from a variety of central nervous system (CNS) regions. The association of three different genotypes (*AA*, *PA*, *PP*) based on the presence (*P*) or absence (*A*) of SVA_67 was analysed. The unpaired Wilcoxon test was used to compare two independent groups of samples and to demonstrate statistical significance.

Results:

Using two different datasets for analysis, we could demonstrate that SVA_67 presence or absence was significantly associated with 1) changes in expression of multiple genes at this locus and 2) both differential isoform and tissue specific expression of these genes.

Conclusions:

This study highlights the importance of addressing SVA polymorphism in disease genetics to gain insight into a better understanding of the role of these regulatory domains to a variety of neurodegenerative diseases. Therefore, this study underlines an additional type of variation to be considered at the *MAPT* locus and an added layer of complexity when analysing the missing heritability of neurodegenerative diseases.

Background:

Neurodegenerative diseases, including Parkinson's disease (PD) and amyotrophic lateral sclerosis (ALS), are complex disorders involving interaction of genetic and environmental factors. Genome-wide association studies (GWAS) and targeted single nucleotide polymorphism (SNP) studies have allowed

the identification of many loci and genetic mutations and polymorphisms associated with these diseases [1–4]. However, further analyses are needed to characterise and identify the source of the current missing heritability of these diseases.

Non-coding repetitive DNA is often an overlooked source of genetic variation. Repetitive DNA can be found in both static and mobile forms, whereby transposable elements (TEs) belonging to the latter class are capable of mobilising throughout the genome [5]. Originally dismissed as “junk” DNA, TEs are now known to contribute to the regulation and evolution of the genome as well as to be involved in genetic instability and disease progression [6]. Based on their transposition strategy and intermediates formed, TEs are split into two families named DNA transposons and retrotransposons [7]. Retrotransposon sequences can propagate via a ‘copy-and-paste’ mechanism leading to a new copy at a new genomic locus of the host genome [8]. SINE-VNTR-*Alu* (SVA) elements, approximately 0.7-4 kb in length, are one member of the non-LTR (long terminal repeat) retrotransposon family with 2700-3000 copies present in the reference human genome. They consist of a 5’ CT-rich hexamer domain (CT element), an antisense *Alu*-like region, a central GC-rich variable number tandem repeat (VNTR, each repeat typically 30-50 bp in length), a SINE-R domain and a 3’ poly A tail (Figure 1a) [9, 10]. SVAs, classified A–F in order of evolutionary age based on their SINE region [11], are important contributors to genetic diversity by a variety of mechanisms which include acting as transcriptional regulatory domains and thus modulating gene expression profiles, for instance, by transcription factor (TF) binding or altering patterns of methylation [12]. Ongoing mobilisation of SVAs has led to insertions being polymorphic for their presence or absence and are thus named retrotransposon insertion polymorphisms (RIPs) [13, 14]. It should be noted that the evolutionarily youngest classes of SVAs (D-F1) may be of special interest for human physiology, health and evolution since they have introduced hominid-specific insertions which can exert novel regulatory potential to specific genomic loci [15].

To date, there are at least 13 disease-causing SVA insertions [13, 16] including the insertion causing X-linked dystonia parkinsonism (XDP), where an SVA-F element was found in intron 32 of the TATA-box binding protein associated factor 1 (*TAF1*) gene [17–20]. Interestingly, this SVA insertion was not only found to be associated with reduced expression of *TAF1*, alternative splicing and intron retention, but also the length of its CT hexamer repeat inversely correlated with age of onset of the disease [21].

Furthermore, polymorphic SVA insertions have been shown to regulate gene expression or isoform expression via intron retention in a population specific manner [19, 22]. This highlights their ability to have an impact at many levels on genetic processing and contribute to phenotypic differences within a variety of diseases including neurodegenerative disorders such as PD or ALS. We recently characterised SVA RIPs utilising whole genome sequencing, transcriptomic and clinical data in the Parkinson’s Progression Markers Initiative (PPMI) cohort, which was designed to help understand PD aetiology, identify progression markers, and enhance development of novel therapeutics [23, 24]. Eighty-one reference genome SVAs polymorphic for their presence/absence were identified, seven of which were linked with PD progression and with differential gene expression using whole blood RNA sequencing data [23]. One of these RIPs, SVA_67, is located 12 kb upstream of the KAT8 regulatory NSL complex subunit 1 (*KANSL1*) gene which is part of the microtubule-associated protein tau (*MAPT*) locus (Figure 1b). The

structurally complex *MAPT* locus contains a 900-kb inversion and is characterised by two predominant haplotypes (*H1* and *H2*) and the presence of SVA_67 is part of *H1* (direct inversion) while *H2* (inverted) is specified by its absence [25, 26]. This locus contains several genes including *MAPT* in which mutations can cause, or polymorphisms are correlated with the neurodegenerative diseases frontotemporal dementia (FTD) with parkinsonism and progressive supranuclear palsy (PSP) [27, 28]. In relation to ALS, it has been shown that 12.5% of patients with behavioural-variant FTD develop ALS, and mild features of motor neuron involvement have been reported in about 40% of patients with FTD [29, 30].

SVAs have the potential to exert regulatory influences on genes distant from the closest gene to which they are found. As non-coding RIPs may lead to interpersonal differences in expression patterns, we aimed to extend our previous findings [23] by analysing the association of the SVA_67 RIP with isoform expression of six genes located in the block at the *MAPT* locus including *MAPT*, *KANSL1*, corticotropin releasing hormone receptor 1 (*CRHR1*), leucine rich repeat containing 37A (*LRR37A*), pleckstrin homology and RUN domain containing M1 (*PLEKHM1*) and ADP ribosylation factor like GTPase 17A (*ARL17A*). Using the datasets from both the PPMI and New York Genome Center Consortium Target ALS (henceforth NYGC ALS) cohort, we could show that the SVA_67 genotype was significantly associated with differential isoform expression of all genes of interest in the *MAPT* locus. Our data only addressed the ability of the SVA to alter isoform expression and current data does not give us sufficient power to address isoform expression in disease progression. These approaches could nevertheless lead to a more precise understanding of transposable elements as contributors to a variety of neurodegenerative diseases.

Methods:

Bioinformatic analysis of the *MAPT* locus

The *MAPT* locus was examined using the UCSC Genome Browser hg38 (<https://genome.ucsc.edu/>), which contains a set of tools allowing the visualisation of a defined genomic region. The 'Repeat Masker' tool (<http://repeatmasker.org>) was used to screen and identify low complexity DNA sequences and interspersed repeats, including retrotransposons and tandem repeat DNA within the reference genome. This method, where 'Repeat Masker' annotations were overlaid with conservation data, specifically the 'vertebrate multiz alignment' & conservation of 100 vertebrate species from 'Phylogenetic Analysis with Space/Time' models (PHAST program) [31], allowed us to determine if the SVA of interest (SVA_67) was human specific or present in other primates. To analyse the potential of this non-coding region to have regulatory function, the genomic region of SVA_67 was overlaid with data from 'The Encyclopedia of DNA Elements' (ENCODE) [32]. Signals for H3K4Me1, H3K4Me3 and H3K27Ac histone marks, which are found near regulatory regions, promoters, and active regulatory elements, respectively, were analysed.

Genotyping of SVA_67 in PPMI and NYGC ALS cohort and analysis of differential isoform expression

In this study, we utilised data from the Parkinson's Progression Markers Initiative (PPMI) and New York Genome Center Consortium Target ALS (NYGC ALS) cohort. The structural variant caller Delly2 (<https://github.com/dellytools/delly>) [33], with default settings, was used to genotype SVA_67 in individuals of the PPMI and the NYGC ALS cohorts [23]. The PPMI is a longitudinal study and SVA_67 was genotyped in 608 individuals, including 179 healthy controls, 371 PD and 58 SWEDD (scans without evidence of dopaminergic deficit) subjects [34]. For the NYGC cohort, the SVA_67 genotype was available for 197 subjects, including 142 ALS, 29 ALS + other neurological disorders (OND), 4 OND, 21 healthy controls and 1 other motor neuron disease (spinal bulbar muscular atrophy). OND diagnosis can include the following diseases: dementia with lewy bodies, peripheral neuropathy, Alzheimer's disease, multiple system atrophy, frontotemporal dementia, argyrophilic grain disease, cerebral amyloid angiopathy, acute meningitis, cerebrovascular disease, chronic traumatic encephalopathy, and Parkinson's disease. RNA-seq data from different tissues of these subjects from the NYGC ALS cohort were available. The majority represented tissues from the central nervous system (CNS), which were complemented by choroid and liver tissues. Taken together, RNA-seq data from the following tissues (total 1,004) were included: cerebellum (163), 141 frontal cortex (141), medial motor cortex (120), lateral motor cortex (119), unspecified motor cortex (10), occipital cortex (75), temporal cortex (5), sensory cortex (2), lumbar spinal cord (114), cervical spinal cord (132), thoracic spinal cord (72), choroid (32) and liver (19).

In order to evaluate the effect of SVA RIP genotypes on the expression profile, differential isoform expression analysis based on the PPMI whole blood RNA-seq data and NYGC ALS RNA-seq data was performed on all subjects. In this analysis, all subjects (cases and controls) were combined. Isoform quantification of RNA-seq data was performed by using the Salmon tool (<https://salmon.readthedocs.io>). Salmon-generated quant files were imported into R using *tximport* function from the *tximport* package [35] of R. Counts were extracted with the *DESeqDataSetFromTximport* function and raw counts were normalised using the median-of-ratios method, implemented in the *DESeq2* package [36]. The *DESeq2* package in R was also used to detect statistically significant differences in the isoform expression profiles between the different genotypes of SVA_67. The association of three different genotypes (AA, PA, PP) based on the presence (P) or absence (A) of SVA_67 was analysed. The *ggplot2* package [37] in R containing the *geom_boxplot* function was used to visualise the data specifying the *stat_summary* function to *mean*. Six genes (*MAPT*, *KANSL1*, *PLEKHM1*, *CRHR1*, *LRCC37A*, and *ARL17A*) centred around this RIP at the *MAPT* locus were selected for analysis. The unpaired Wilcoxon test was used to compare two independent groups of samples and to demonstrate statistical significance.

Results:

Bioinformatic analysis of the MAPT locus

The *MAPT* locus on chromosome 17q21.31 was analysed using UCSC genome browser, specifically evaluating SVAs, ENCODE data and evolutionary DNA conservation over this region (Figure 1). Using the 'Repeat Masker' data track for analysis, four SVA retrotransposons were identified, including two SVA-C elements as well as one element of the classes D and F (Figure 1b). UCSC genome browser additionally

indicated three potential SVA-A elements which were excluded because these elements had a length of 58, 127 and 128 bp, respectively, and did not show the characteristic structure of SVA elements. We previously identified the SVA-F element, termed SVA_67, as a RIP showing that it transposed relatively recently in evolutionary age. SVA_67 (hg38, chr17:46,237,519 – 46,238,226), located 12 kb upstream of the *KANSL1* gene, represented a truncated element with a length of 707 bp. This element was anti-sense with respect to the orientation of the genes (defined by their 5' transcriptional start sites (TSS)) *LRRC37A*, *CRHR1* and *MAPT*, and showed a sense orientation relative to *KANSL1*, *ARL17A* and *PLEKHM1* (Figure 1b, c). These six genes were selected for isoform expression analyses. The distances between SVA_67 and the TSS of analysed gene isoforms (Table 1) ranged from approximately 12 kb to 800,000 kb and are summarised in Table 1. Using the 'vertebrate multiz alignment' & conservation track of 100 vertebrate species, this analysis showed that this genomic region was not conserved in chimps, gorillas, gibbons, rhesus macaques, rats, and mice, indicating that SVA_67 was human specific (Figure 1c). We overlaid the genomic region of SVA_67 with data from ENCODE to analyse if this SVA was associated with signals for H3K4Me1 (marker of regulatory domains associated with enhancers), H3K4Me3 (marker of regulatory domains associated with promoters) and H3K27Ac (marker of active regulatory domains associated with active enhancer elements) marks (Figure 1c). No major histone marks associated with active chromatin were observed in proximity to SVA_67, although this may not be surprising as SVAs are not well captured and aligned to specific genomic locations utilising short read DNA sequence due to the repetitive nature and primary sequence homology of the SVA retrotransposons.

Table 1
Analysed gene isoforms in this study and their specifics.

Name	Transcript ID	bp	Protein	Biotype*	Location on chromosome 17	Distance SVA_67 to TSS (bp)
MAPT-201	ENST00000262410	6815	833aa	PC	45,894,554-46,028,334	342,965
MAPT-202	ENST00000334239	1107	352aa	PC	45,894,668-46,024,183	342,851
MAPT-203	ENST00000344290	2609	736aa	PC	45,894,560-46,024,425	342,959
MAPT-204	ENST00000351559	5639	441aa	PC	45,894,554-46,028,334	342,965
MAPT-205	ENST00000415613	2331	776aa	PC	45,962,338-46,024,171	275,181
MAPT-206	ENST00000420682	1239	412aa	PC	45,962,338-46,024,171	275,181
MAPT-207	ENST00000431008	1233	410aa	PC	45,962,338-46,024,171	275,181
MAPT-208	ENST00000446361	5345	383aa	PC	45,894,674-46,028,334	342,845
MAPT-209	ENST00000535772	1353	381aa	PC	45,894,551-46,024,225	342,968
MAPT-210	ENST00000570299	889	-	PT	45,894,576-46,018,730	342,943
MAPT-211	ENST00000571311	544	59aa	NMD	45,894,566-45,978,425	342,953
MAPT-212	ENST00000571987	2277	758aa	PC	45,962,338-46,024,171	275,181
MAPT-213	ENST00000572440	5015	-	RI	45,975,492-45,980,506	262,027
MAPT-215	ENST00000576518	6778	-	RI	45,972,783-46,024,431	264,736
MAPT-217	ENST00000680542	2988	412aa	PC	45,894,663-46,025,873	342,856
MAPT-218	ENST00000680674	1936	424aa	PC	45,894,527-46,024,655	342,992

Abbreviations: PC, protein coding; PT, processed transcript; NMD, nonsense mediated decay; RI, retained intron; TSS, transcriptional start site.

Name	Transcript ID	bp	Protein	Biotype*	Location on chromosome 17	Distance SVA_67 to TSS (bp)
KANSL1-201	ENST00000262419	5309	1105aa	PC	46,029,956-46,192,800	44,719
KANSL1-202	ENST00000432791	5574	1105aa	PC	46,029,916-46,193,429	44,090
KANSL1-203	ENST00000571698	2276	678aa	PC	46,039,870-46,193,901	43,618
KANSL1-204	ENST00000572218	9095	-	RI	46,029,916-46,046,121	191,398
KANSL1-205	ENST00000572904	4973	1105aa	PC	46,029,916-46,223,676	13,843
KANSL1-206	ENST00000573286	5437	-	RI	46,037,325-46,043,101	194,418
KANSL1-207	ENST00000573682	358	-	RI	46,032,165-46,033,512	204,007
KANSL1-208	ENST00000574590	5147	1104aa	PC	46,029,916-46,225,389	12,130
KANSL1-209	ENST00000574655	1681	-	PT	46,152,904-46,225,371	12,148
KANSL1-210	ENST00000574963	1493	-	RI	46,031,074-46,032,566	204,953
KANSL1-211	ENST00000575318	4935	1041aa	PC	46,029,917-46,225,367	12,152
KANSL1-212	ENST00000576137	815	-	RI	46,033,099-46,039,075	198,444
KANSL1-213	ENST00000576248	452	-	PT	46,119,967-46,171,223	66,296
KANSL1-214	ENST00000576739	592	9aa	PC	46,172,117-46,224,902	12,617
KANSL1-215	ENST00000576870	2848	-	RI	46,029,918-46,040,065	197,454
KANSL1-216	ENST00000577114	629	-	PT	46,050,650-46,121,338	116,181
KANSL1-217	ENST00000638269	3365	-	RI	46,092,870-46,225,361	12,158
Abbreviations: PC, protein coding; PT, processed transcript; NMD, nonsense mediated decay; RI, retained intron; TSS, transcriptional start site.						

Name	Transcript ID	bp	Protein	Biotype*	Location on chromosome 17	Distance SVA_67 to TSS (bp)
KANSL1-218	ENST00000638275	5352	1041aa	PC	46,029,917-46,193,400	44,119
KANSL1-220	ENST00000638551	899	-	PT	46,032,186-46,035,160	202,359
KANSL1-221	ENST00000638902	1971	-	RI	46,170,277-46,223,685	13,834
KANSL1-222	ENST00000639099	1724	-	PT	46,152,904-46,225,367	12,152
KANSL1-223	ENST00000639150	1815	523aa	PC	46,033,080-46,224,693	12,826
KANSL1-224	ENST00000639279	387	-	PT	46,088,428-46,094,641	142,878
KANSL1-225	ENST00000639356	1717	-	PT	46,148,125-46,225,367	12,152
KANSL1-226	ENST00000639375	3392	-	RI	46,049,399-46,225,355	12,164
KANSL1-228	ENST00000639520	2655	-	PT	46,221,186-46,225,376	12,143
KANSL1-229	ENST00000639531	2731	897aa	PC	46,032,257-46,172,183	65,336
KANSL1-231	ENST00000639853	1425	475aa	PC	46,038,636-46,171,416	66,103
KANSL1-233	ENST00000640519	1970	-	PT	46,033,045-46,035,924	201,595
KANSL1-236	ENST00000648792	4933	1061aa	PC	46,029,985-46,225,373	12,146
PLEKHM1-201	ENST00000430334	5240	1056aa	PC	45,435,900-45,490,721	746,798
PLEKHM1-202	ENST00000446609	2825	63aa	NMD	45,440,163-45,487,858	749,661
PLEKHM1-203	ENST00000579131	499	-	PT	45,437,916-45,441,059	796,460
PLEKHM1-204	ENST00000579197	4995	63aa	NMD	45,435,904-45,490,728	746,791
Abbreviations: PC, protein coding; PT, processed transcript; NMD, nonsense mediated decay; RI, retained intron; TSS, transcriptional start site.						

Name	Transcript ID	bp	Protein	Biotype*	Location on chromosome 17	Distance SVA_67 to TSS (bp)
PLEKHM1-205	ENST00000580205	2385	-	PT	45,452,938-45,460,432	777,087
PLEKHM1-206	ENST00000580404	2628	-	PT	45,435,903-45,446,273	791,246
PLEKHM1-207	ENST00000581448	3292	335aa	NMD	45,437,471-45,490,729	746,790
PLEKHM1-208	ENST00000581932	1163	-	RI	45,477,074-45,482,525	754,994
PLEKHM1-209	ENST00000582035	1834	-	RI	45,452,938-45,454,771	782,748
PLEKHM1-211	ENST00000583150	479	-	PT	45,439,615-45,448,112	789,407
PLEKHM1-212	ENST00000584420	603	140aa	PC	45,475,211-45,490,738	746,781
PLEKHM1-213	ENST00000585506	332	-	RI	45,439,514-45,440,373	797,146
PLEKHM1-214	ENST00000586084	553	99aa	NMD	45,475,696-45,490,741	746,778
PLEKHM1-215	ENST00000586562	550	-	PT	45,475,597-45,490,749	746,770
PLEKHM1-216	ENST00000589780	587	152aa	PC	45,475,366-45,490,741	746,778
PLEKHM1-217	ENST00000590991	327	45aa	NMD	45,439,546-45,445,507	792,012
PLEKHM1-218	ENST00000591580	650	113aa	PC	45,437,549-45,445,533	791,986
CRHR1-201	ENST00000293493	2621	430aa	PC	45,784,280-45,835,827	453,239
CRHR1-202	ENST00000314537	2537	415aa	PC	45,784,320-45,835,828	453,199
CRHR1-203	ENST00000339069	2490	314aa	PC	45,784,280-45,835,827	453,239
CRHR1-204	ENST00000347197	2462	145aa	NMD	45,784,307-45,835,825	453,212
Abbreviations: PC, protein coding; PT, processed transcript; NMD, nonsense mediated decay; RI, retained intron; TSS, transcriptional start site.						

Name	Transcript ID	bp	Protein	Biotype*	Location on chromosome 17	Distance SVA_67 to TSS (bp)
CRHR1-205	ENST00000352855	1146	375aa	PC	45,784,527-45,834,764	452,992
CRHR1-206	ENST00000398285	2399	444aa	PC	45,784,545-45,835,828	452,974
CRHR1-207	ENST00000535778	1272	52aa	NMD	45,833,138-45,835,644	404,381
CRHR1-208	ENST00000577353	1206	401aa	PC	45,784,545-45,834,764	452,974
CRHR1-209	ENST00000580876	223	75aa	PC	45,833,138-45,834,645	404,381
CRHR1-210	ENST00000580955	479	93aa	NMD	45,807,074-45,830,125	430,445
CRHR1-211	ENST00000581479	606	-	RI	45,830,400-45,833,202	407,119
CRHR1-212	ENST00000582766	828	-	RI	45,784,569-45,830,940	452,950
CRHR1-213	ENST00000583888	777	152aa	NMD	45,816,558-45,834,841	420,961
CRHR1-214	ENST00000619154	2370	154aa	PC	45,784,280-45,835,827	453,239
ARL17A-202	ENST00000336125	5242	177aa	PC	46,552,756-46,579,691	341,465
ARL17A-203	ENST00000445552	798	125aa	PC	46,516,702-46,579,670	342,151
ARL17A-206	ENST00000622488	524	88aa	PC	46,528,685-46,579,673	342,154
LRRC37A-201	ENST00000320254	5177	1700aa	PC	46,295,131-46,337,794	56,905
LRRC37A-202	ENST00000393465	4997	1635aa	PC	46,295,099-46,337,777	57,580
LRRC37A-203	ENST00000496930	3860	738aa	PC	46,292,733-46,337,792	55,214
Abbreviations: PC, protein coding; PT, processed transcript; NMD, nonsense mediated decay; RI, retained intron; TSS, transcriptional start site.						

Genotype of SVA₆₇ is significantly associated with differential isoform expression of several genes in the MAPT locus

We assessed whether the SVA₆₇ genotype correlated with isoform expression of six genes (*MAPT*, *KANSL1*, *PLEKHM1*, *CRHR1*, *LRCC37A*, *ARL17A*) centred around this RIP at the *MAPT* locus (Figure 1b). Transcriptomic data of the PPMI and NYGC ALS cohort was utilised to determine if the SVA RIP had an influence on isoform expression. The association of three different genotypes (*AA*, *PA*, *PP*) based on the presence (*P*) or absence (*A*) of SVA₆₇ was analysed. Initial analysis of one isoform of each gene of interest using the NYGC ALS cohort demonstrated that SVA₆₇ allele dosage is significantly associated with differential expression of the specific isoforms MAPT-208, KANSL1-214, LRRC37A-202, CRHR1-205, PLEKHM1-212 and ARL17A-203 (Figure 2). Three of these isoforms (MAPT-208, PLEKHM1-212 and ARL17A-203) had significantly increased expression with *PP* genotype compared to both *PA* and *AA* genotype (Figure 2a, d-f). This was in contrast with the levels of expression of the isoforms KANSL1-214, LRRC37A-202 and CRHR1-205 where the opposite pattern was observed (Figure 2b, c). When extending the analysis for all detectable isoforms of the genes of interest using the NYGC ALS cohort, 14/16 *MAPT*, 3/3 *LRRC37A*, 3/3 *ARL17A*, 7/12 *CRHR1*, 12/17 *PLEKHM1* and 23/28 *KANSL1* isoforms were significantly associated with at least one SVA₆₇ genotype (Additional File 1: Fig. S1, 3-5, 7, 9, 10). When utilising the PPMI cohort, 2/4 *MAPT*, 3/3 *LRRC37A*, 2/2 *ARL17A*, 5/6 *CRHR1*, 7/9 *PLEKHM1* and 7/7 *KANSL1* isoforms were significantly associated with at least one SVA₆₇ genotype (Additional File 1: Fig. S2-4, 6, 8, 11). More isoforms of these genes were detected in the NYGC ALS RNAseq than blood derived from the PPMI cohort. The majority of the corresponding RNAseq data derived from tissues of the CNS (brain and spinal cord) with just a small portion from choroid and liver. Further characterisation and validation will be required to address the significance of that data; however, it could be as simple as the heterogeneity of the CNS compared to blood. Analysed isoforms of each gene and additional information including isoform ID and biotype are summarised in Table 1.

Presence of SVA₆₇ correlates with gene expression in an isoform-specific manner

We next assessed if SVA₆₇ had the same observed influence on each isoform of a gene regarding positive or negative impact on expression. A major consideration when analysing this data was that the expression could be from different cell types in the brain and therefore such positive or negative findings could be generated by two different cell populations. Nevertheless, utilising the NYGC ALS cohort, there was at least one isoform of each analysed gene present that demonstrated higher and lower expression, respectively, in individuals with *PP* genotypes compared to both *PA* and *AA* genotypes. For example, when addressing *MAPT*, the isoforms MAPT-202 and MAPT-212 showed significantly reduced expression with *PP* genotype compared to both *PA* and *AA* (Figure 3a), while two other isoforms of this same gene (MAPT-208 and MAPT-217) showed the opposite effect (Figure 3b). All four displayed isoforms represented protein coding isoforms encoding distinct proteins with different lengths and there was no association with a specific TSS although several are defined for the gene (Figure 3c, Table 1). This

isoform-specific effect of SVA₆₇ on expression was also visible when using the PPMI cohort, where subjects with *PP* genotypes showed significantly increased and reduced, respectively, expression of a specific isoform compared to *AA*, *PA* or both *PA* and *AA* genotypes. For example, *KANSL1* isoforms KANSL1-201 and KANSL1-211 showed significantly increased expression with *PP* genotype compared to both *PA* and *AA*, while isoforms KANSL1-209 and KANSL1-214 were associated with reduced expression when the subjects had both alleles of SVA₆₇ present (Additional File 1: Fig. S10). No statistical significance regarding positive and negative effect was obtained for *MAPT* isoforms in the PPMI cohort. Here, subjects with *PP* genotype correlated significantly with increased expression of isoform MAPT-208 (*PP* vs *PA* and *AA*) and MAPT-215 (*PP* vs *AA*) (Additional File 1: Fig. S2). The association of SVA₆₇ genotype did not reach statistical significance for the isoform MAPT-211, however a trend was visible which indicated a reduced expression with *PP* and *PA* genotypes compared to subjects having no allele of SVA₆₇ present.

A tissue specific influence of SVA₆₇ allele dosage on isoform expression

We analysed the effect of SVA₆₇ on specific isoforms in different tissues by comparing the influence of this SVA RIP in two different datasets (PPMI and NYGC ALS cohort). The six isoforms LRRC37A-201, ARL17A-203, CRHR1-203, PLEKHM1-202, KANSL1-208 and KANSL1-209 were differentially influenced by SVA₆₇ allele dosage demonstrating opposing effects on expression when comparing both datasets (Additional File 1: Fig. S3, S4, 7-10). Using the PPMI cohort, isoform PLEKHM1-202 showed significantly increased expression with *PP* genotype compared to both *PA* and *AA* genotype, while the opposite effect is visible when utilising the NYGC ALS cohort. We do not believe this is disease specific as we have both cases and controls in our datasets to give us power to address transcriptional changes, however isoform association with disease can be considered in the future as the number of cases and controls increases in such data sets. A tissue-specific effect was also detectable for the two *KANSL1* isoforms KANSL1-208 and KANSL1-209. When using the PPMI dataset for analysis, both isoforms were associated with reduced expression when having two alleles of SVA₆₇ present, while the same genotype led to a decreased expression when using the NYGC ALS cohort for analysis. Regarding LRRC37A-201, this isoform showed an increased expression with one copy of SVA₆₇ present compared to *AA* in the PPMI cohort. Again, the opposite effect was visible when utilising the NYGC ALS dataset. ARL17A-203 and CRHR1-203 isoforms showed a significantly reduced expression with *PP* genotype compared to both *PA* and *AA* genotype (PPMI dataset) (Figure 4a, c), while the opposite effect was visible when utilising the NYGC ALS cohort (Figure 4b, d).

Discussion:

In this study we analysed the impact of SVA₆₇ allele dosage on isoform expression of six genes (*MAPT*, *KANSL1*, *CRHR1*, *PLEKHM1*, *ARL17A* and *LRRC37A*) at the *MAPT* locus (Figure 1) in order to gain insight into the role of SVAs to modify the transcriptome. Using two different datasets (PPMI and NYGC ALS cohort) for analysis, we could demonstrate that SVA₆₇ was significantly associated with 1) changes in

expression of multiple genes at this locus and 2) both differential isoform and tissue specific expression of these genes (Figure 2-4). The data is consistent with SVAs having functional consequences for gene expression over long distances directly by activator or repressor mechanisms or their ability to modulate genome structure by looping mechanisms affecting 3D structures in such as transcriptional hubs [38].

SVAs have previously been demonstrated to be involved in genetic instability and disease progression via genome regulation mechanisms [6, 10, 13]. It has been demonstrated that SVAs can modulate gene expression profiles by various mechanisms, for instance, by transcription factor binding, altering patterns of methylation and interaction with distant promoters through 3D chromatin structure by recruiting CCCTC-binding factor (CTCF), a master regulator of 3D chromatin structure [12, 39]. Indeed, TEs have shown to have multiple binding sites for CTCF which has a well-established role in chromatin looping and topologically associated domain (TAD) formation [40–43]. This could be one model to allow SVA_67 to act at many genes to enhance or reduce gene isoform expression or alter chromatin structure via looping. This is consistent with the obtained results showing that SVA_67 was capable of affecting isoform expression at both the most proximal (e.g., *KANSL1*) and most distant (e.g., *PLEKHM1*) TSS relative to its location at the *MAPT* locus (Figure 2, Additional File 1: Fig. S7-10).

We have previously characterised the function of SVAs to modulate gene expression profiles by *in vitro* and *in vivo* models including an SVA upstream of the FUS RNA binding protein (*FUS*) gene in which genetic mutations have been linked to many diseases including ALS [44] and an SVA upstream of the Parkinsonism associated deglycase DJ-1 gene, also termed *PARK7*, a gene associated with PD [45]. These demonstrated the action of SVAs to have classical regulatory domains when analysed in reporter gene constructs and indeed that the SVAs were a composite regulatory domain containing multiple functional domains. In our recent study, we could identify 81 reference SVAs polymorphic for their presence/absence, seven of which, including SVA_67, were associated with the progression of PD using the PPMI cohort [23]. We have extended that previous work by addressing differential gene expression on an isoform-based level using the same cohort. We observed that SVA_67 had an isoform-specific correlation with gene expression showing potentially the capacity of SVA_67 to differentially modulate the transcriptome (Figure 3). No study to date has shown this regulatory influence of an SVA on an isoform-specific level, although intron retention has been proposed for the action of an SVA within an intron of the *TAF1* gene [19]. However, intron retention is not operating at the *MAPT* locus as the SVA is intergenic in nature (Figure 1). To date, studies have focused predominantly on differential gene expression analyses, however, these approaches are limited as they do not account for isoform diversity [9, 10, 19, 46]. Many of the genes at this locus have been associated with CNS functional parameters [47–50] and encode more than one isoform which are generated, for instance, through mechanisms such as alternative splicing or alternative usage of transcription start sites [51]. It is important to differentiate between isoforms, as some of these can represent protein-coding isoforms with different functions and/or subcellular localisations, while others do not lead to a protein product (Table 1). Dick *et al.* have reported the first genome-wide study of differential transcript usage (DTU), the measure of the relative contribution of one isoform to overall gene expression, in PD [52]. The authors demonstrated that PD subjects showed a decrease in the relative usage of a thioesterase superfamily member 5 (*THEM5*)

transcript (involved in mitochondrial fatty acid metabolism), while the concomitant increase of a shorter isoform, which more likely localises to the extracellular space than to the mitochondria, may therefore not recapitulate the function of the full-length protein [52, 53]. Indeed, these observed changes in the relative expression of specific gene isoforms may affect the ratio of the resulting protein isoforms, which in turn could affect cellular signalling pathway or metabolism through variation of, for instance, function and subcellular localisation [52]. In our study SVA_67 allele dosage was significantly associated, for instance, with *PLEKHM1* isoform expression (Figure 2, Additional File 1: Fig. S7, 8). McEwan *et al.* demonstrated in their study that *PLEKHM1* regulates clearance of protein aggregates, which are implicated in the pathogenesis of multiple diseases like ALS or PD, in an autophagy- and LC3-interacting region-dependent manner and therefore showed a potential mechanism that links *PLEKHM1* to an increased risk of PD [47]. Our data suggest that SVA_67 could influence *PLEKHM1* function by modulating isoform expression which could result in a gene specific DTU. This may have an impact on the potential role of *PLEKHM1* in maintaining appropriate cellular functions or even cell survival. Future work could address the role of *PLEKHM1* and other genes of interest in this study in neurodegenerative diseases by analysing the influence of isoform switches and DTU. More importantly it demonstrates the potential for multigenic regulatory effects of a single variant over a large region of the genome giving greater insight into the complex mechanisms that have to be factored in to understand how a variant is affecting not only genomic regulation but disease progression.

It is known that TE derived sequences can contribute to the regulation of the human genome and that domains with regulatory potential can function in a tissue-specific manner. We previously utilised Genotype-Tissue Expression (GTEx) database [54] to characterise the potential of SVA_67 to act in other tissues using its proxy SNP (rs55653937) [23]. This approach led, based on the findings that the tagging SNP for SVA_67 was identified as eQTL (expression quantitative trait loci) for over 30 genes including the six genes of interest in this study, to the assumption that SVA_67 could also influence gene expression in other tissues in addition to whole blood. We validated that by using transcriptomic data derived from different tissues (NYGC ALS vs PPMI) and demonstrated that SVA_67 correlated with a tissue-specific effect on isoform expression of six genes in the *MAPT* locus (Figure 4). Although we do not wish to draw conclusions regarding specific expression and disease progression, we do want to highlight the plethora of transcriptomic changes associated with one SVA insertion. We are not able to demonstrate that causation of differential gene expression is directed by the SVA which would await functional validation by such as CRISPR.

Conclusions:

Our study presented here demonstrated that SVA insertions have the potential to influence expression of multiple genes over large distances and that regulation can be isoform specific. SVAs could be influencing the disease course of PD and ALS through modulation of isoform expression and usage which ultimately could affect protein levels and biological processes. Therefore, this study highlighted an additional type of variation to be considered at the *MAPT* locus and an added layer of complexity when analysing the missing heritability of neurodegenerative diseases.

Abbreviations

ALS

Amyotrophic lateral sclerosis

ARL17A

ADP ribosylation factor like GTPase 17A

CRHR1

Corticotropin releasing hormone receptor 1

CTCF

CCCTC-binding factor

DTU

Differential transcript usage

ENCODE

The Encyclopedia of DNA Elements

eQTL

Expression quantitative trait loci

FTD

Frontotemporal dementia

FUS

FUS RNA binding protein

GTEx

Genotype-Tissue Expression

KANSL1

KAT8 regulatory NSL complex subunit 1

LRRC37A

Leucine rich repeat containing 37A

MAPT

Microtubule-associated protein tau

NYGC ALS

New York Genome Center Consortium Target ALS

PD

Parkinson's disease

PLEKHM1

Pleckstrin homology and RUN domain containing M1

PPMI

Parkinson's Progression Marker Initiative

PSP

Progressive supranuclear palsy

RIP

Retrotransposon insertion polymorphism

SVA
SINE-VNTR-*Alu*
TAD
Topologically associated domain
TAF1
TATA-box binding protein associated factor 1
TE
Transposable element
THEM5
Thioesterase superfamily member 5
TSS
Transcriptional start site
VNTR
Variable number tandem repeat
XDP
X-linked dystonia parkinsonism

Declarations

Ethics approval and consent to participate:

Not applicable.

Consent for publication:

Not applicable.

Availability of data and materials:

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests:

The authors declare that they have no competing interests.

Funding:

A.F. & J.P.Q. are funded by the Andrzej Wlodarski Memorial Research Fund. A.F. is recipient of Andrzej Wlodarski Memorial Research PhD scholarship. A.L.P. and S.K. are funded by MSWA and Perron Institute for Neurological and Translational Science. This study was funded by the Motor Neurone Disease Association (Grant Number: 41/523).

Author's contributions: Author's contributions: Study concept and design: A.F., A.L.P., V.J.B., S.K., J.P.Q.; Analysis and interpretation of data: A.F., A.L.P., V.J.B., S.K., J.P.Q.; Drafting of the manuscript: A.F.; Critical revision of the manuscript for important intellectual content: A.F., A.L.P., V.J.B., S.K., J.P.Q.; all authors reviewed and approved the final manuscript.

Acknowledgements:

Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org. PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including Abbvie, Allergan, Amathus therapeutics, Avid Radiopharmaceuticals, Biogen Idec, Biolegend, Briston-Myers Squibb, Celgene, Denali, GE Healthcare, Genentech, GlaxoSmithKline, janssen neuroscience, Lilly, Lundbeck, Merck, Meso Scale Discovery, Pfizer, Piramal, Prevail Therapeutics, Roche, Sanofi Genzyme, Servier, Takeda, Teva, UCB, Verily and Voyager Therapeutics. We acknowledge the Target ALS Human Postmortem Tissue Core, New York Genome Center for Genomics of Neurodegenerative Disease, Amyotrophic Lateral Sclerosis Association and TOW Foundation for RNA-seq data of human ALS patients. This work was supported by resources provided by the Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia.

Author's Information:

Department of Pharmacology and Therapeutics, Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool, United Kingdom

Alexander Fröhlich, Vivien J. Bubb & John P. Quinn

Centre for Molecular Medicine and Innovative Therapeutics, Murdoch University, Perth, WA Australia

Abigail L. Pfaff & Sulev Koks

Perron Institute for Neurological and Translational Science, Perth, WA Australia

Abigail L. Pfaff & Sulev Koks

References

1. Taylor JP, Brown RH, Cleveland DW. Decoding ALS: from genes to mechanism. *Nature*. 2016;539 7628:197–206; doi: 10.1038/nature20413.
2. van Rheenen W, Shatunov A, Dekker AM, McLaughlin RL, Diekstra FP, Pulit SL, et al. Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat Genet*. 2016;48 9:1043–8; doi: 10.1038/ng.3622.

3. Nalls MA, Blauwendraat C, Vallerga CL, Heilbron K, Bandres-Ciga S, Chang D, et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* 2019;18 12:1091–102; doi: 10.1016/S1474-4422(19)30320-5.
4. Blauwendraat C, Nalls MA, Singleton AB. The genetic architecture of Parkinson's disease. *Lancet Neurol.* 2020;19 2:170–8; doi: 10.1016/S1474-4422(19)30287-X.
5. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. *Genome Biol.* 2018;19 1:199; doi: 10.1186/s13059-018-1577-z.
6. Ayarpadikannan S, Kim HS. The impact of transposable elements in genome evolution and genetic instability and their implications in various diseases. *Genomics Inform.* 2014;12 3:98–104; doi: 10.5808/GI.2014.12.3.98.
7. Savage AL, Schumann GG, Breen G, Bubb VJ, Al-Chalabi A, Quinn JP. Retrotransposons in the development and progression of amyotrophic lateral sclerosis. *J Neurol Neurosurg Psychiatry.* 2019;90 3:284–93; doi: 10.1136/jnnp-2018-319210.
8. Elbarbary RA, Lucas BA, Maquat LE. Retrotransposons as regulators of gene expression. *Science.* 2016;351 6274:aac7247; doi: 10.1126/science.aac7247.
9. Gianfrancesco O, Bubb VJ, Quinn JP. SVA retrotransposons as potential modulators of neuropeptide gene expression. *Neuropeptides.* 2017;64:3–7; doi: 10.1016/j.npep.2016.09.006.
10. Hancks DC, Kazazian HH. SVA retrotransposons: Evolution and genetic instability. *Semin Cancer Biol.* 2010;20 4:234–45; doi: 10.1016/j.semcancer.2010.04.001.
11. Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, et al. SVA elements: a hominid-specific retroposon family. *J Mol Biol.* 2005;354 4:994–1007; doi: 10.1016/j.jmb.2005.09.085.
12. Hancks DC, Goodier JL, Mandal PK, Cheung LE, Kazazian HH. Retrotransposition of marked SVA elements by human L1s in cultured cells. *Hum Mol Genet.* 2011;20 17:3386–400; doi: 10.1093/hmg/ddr245.
13. Hancks DC, Kazazian HH. Roles for retrotransposon insertions in human disease. *Mob DNA.* 2016;7:9; doi: 10.1186/s13100-016-0065-9.
14. Gianfrancesco O, Geary B, Savage AL, Billingsley KJ, Bubb VJ, Quinn JP. The Role of SINE-VNTR-Alu (SVA) Retrotransposons in Shaping the Human Genome. *Int J Mol Sci.* 2019;20 23; doi: 10.3390/ijms20235977.
15. Vasieva O, Cetiner S, Savage A, Schumann GG, Bubb VJ, Quinn JP. Potential impact of primate-specific SVA retrotransposons during the evolution of human cognitive function. *Trends in Evolutionary Biology.* 2017;6.
16. Bychkov I, Baydakova G, Filatova A, Migiaev O, Marakhonov A, Pechatnikova N, et al. Complex Transposon Insertion as a Novel Cause of Pompe Disease. *Int J Mol Sci.* 2021;22 19; doi: 10.3390/ijms221910887.
17. Delvallée C, Nicaise S, Antin M, Leuvrey AS, Nourisson E, Leitch CC, et al. A BBS1 SVA F retrotransposon insertion is a frequent cause of Bardet-Biedl syndrome. *Clin Genet.* 2021;99 2:318–

- 24; doi: 10.1111/cge.13878.
18. Yamamoto G, Miyabe I, Tanaka K, Kakuta M, Watanabe M, Kawakami S, et al. SVA retrotransposon insertion in exon of MMR genes results in aberrant RNA splicing and causes Lynch syndrome. *Eur J Hum Genet.* 2021;29 4:680–6; doi: 10.1038/s41431-020-00779-5.
 19. Makino S, Kaji R, Ando S, Tomizawa M, Yasuno K, Goto S, et al. Reduced neuron-specific expression of the TAF1 gene is associated with X-linked dystonia-parkinsonism. *Am J Hum Genet.* 2007;80 3:393–406; doi: 10.1086/512129.
 20. Aneichyk T, Hendriks WT, Yadav R, Shin D, Gao D, Vaine CA, et al. Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell.* 2018;172 5:897-909.e21; doi: 10.1016/j.cell.2018.02.011.
 21. Bragg DC, Mangkalaphiban K, Vaine CA, Kulkarni NJ, Shin D, Yadav R, et al. Disease onset in X-linked dystonia-parkinsonism correlates with expansion of a hexameric repeat within an SVA retrotransposon in. *Proc Natl Acad Sci U S A.* 2017;114 51:E11020-E8; doi: 10.1073/pnas.1712526114.
 22. Wang L, Rishishwar L, Mariño-Ramírez L, Jordan IK. Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements. *Nucleic Acids Res.* 2017;45 5:2318–28; doi: 10.1093/nar/gkw1286.
 23. Pfaff AL, Bubb VJ, Quinn JP, Koks S. Reference SVA insertion polymorphisms are associated with Parkinson's Disease progression and differential gene expression. *NPJ Parkinsons Dis.* 2021;7 1:44; doi: 10.1038/s41531-021-00189-4.
 24. Initiative PPM. The Parkinson Progression Marker Initiative (PPMI). *Prog Neurobiol.* 2011;95 4:629–35; doi: 10.1016/j.pneurobio.2011.09.005.
 25. Zody MC, Jiang Z, Fung HC, Antonacci F, Hillier LW, Cardone MF, et al. Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat Genet.* 2008;40 9:1076–83; doi: 10.1038/ng.193.
 26. Wider C, Vilariño-Güell C, Jasinska-Myga B, Heckman MG, Soto-Ortolaza AI, Cobb SA, et al. Association of the MAPT locus with Parkinson's disease. *Eur J Neurol.* 2010;17 3:483–6; doi: 10.1111/j.1468-1331.2009.02847.x.
 27. Im SY, Kim YE, Kim YJ. Genetics of Progressive Supranuclear Palsy. *J Mov Disord.* 2015;8 3:122–9; doi: 10.14802/jmd.15033.
 28. Strang KH, Golde TE, Giasson BI. MAPT mutations, tauopathy, and mechanisms of neurodegeneration. *Lab Invest.* 2019;99 7:912–28; doi: 10.1038/s41374-019-0197-x.
 29. Bang J, Spina S, Miller BL. Frontotemporal dementia. *Lancet.* 2015;386 10004:1672–82; doi: 10.1016/S0140-6736(15)00461-4.
 30. van Es MA, Hardiman O, Chio A, Al-Chalabi A, Pasterkamp RJ, Veldink JH, et al. Amyotrophic lateral sclerosis. *Lancet.* 2017;390 10107:2084–98; doi: 10.1016/S0140-6736(17)31287-4.
 31. Hubisz MJ, Pollard KS, Siepel A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief Bioinform.* 2011;12 1:41–51; doi: 10.1093/bib/bbq072.

32. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489 7414:57–74; doi: 10.1038/nature11247.
33. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 2012;28 18:i333-i9; doi: 10.1093/bioinformatics/bts378.
34. Marek K, Chowdhury S, Siderowf A, Lasch S, Coffey CS, Caspell-Garcia C, et al. The Parkinson's progression markers initiative (PPMI) - establishing a PD biomarker cohort. *Ann Clin Transl Neurol*. 2018;5 12:1460–77; doi: 10.1002/acn3.644.
35. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res*. 2015;4:1521; doi: 10.12688/f1000research.7563.2.
36. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15 12:550; doi: 10.1186/s13059-014-0550-8.
37. Wickham H, Sievert C, SpringerLink. ggplot2: Elegant Graphics for Data Analysis. 2016.
38. Ferrari R, Grandi N, Tramontano E, Dieci G. Retrotransposons as Drivers of Mammalian Brain Evolution. *Life (Basel)*. 2021;11 5; doi: 10.3390/life11050376.
39. Wang DC, Wang W, Zhang L, Wang X. A tour of 3D genome with a focus on CTCF. *Semin Cell Dev Biol*. 2019;90:4–11; doi: 10.1016/j.semcdb.2018.07.020.
40. Kentepozidou E, Aitken SJ, Feig C, Stefflova K, Ibarra-Soria X, Odom DT, et al. Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. *Genome Biol*. 2020;21 1:5; doi: 10.1186/s13059-019-1894-x.
41. Pugacheva EM, Kubo N, Loukinov D, Tajmul M, Kang S, Kovalchuk AL, et al. CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention. *Proc Natl Acad Sci U S A*. 2020;117 4:2020–31; doi: 10.1073/pnas.1911708117.
42. Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res*. 2008;18 11:1752–62; doi: 10.1101/gr.080663.108.
43. Schwalie PC, Ward MC, Cain CE, Faure AJ, Gilad Y, Odom DT, et al. Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes. *Genome Biol*. 2013;14 12:R148; doi: 10.1186/gb-2013-14-12-r148.
44. Savage AL, Wilm TP, Khursheed K, Shatunov A, Morrison KE, Shaw PJ, et al. An evaluation of a SVA retrotransposon in the FUS promoter as a transcriptional regulator and its association to ALS. *PLoS One*. 2014;9 6:e90833; doi: 10.1371/journal.pone.0090833.
45. Savage AL, Bubb VJ, Breen G, Quinn JP. Characterisation of the potential function of SVA retrotransposons to modulate gene expression patterns. *BMC Evol Biol*. 2013;13:101; doi: 10.1186/1471-2148-13-101.
46. Hall A, Moore AK, Hernandez DG, Billingsley KJ, Bubb VJ, Quinn JP, et al. A SINE-VNTR-. *Int J Mol Sci*. 2020;21 22; doi: 10.3390/ijms21228486.

47. McEwan DG, Popovic D, Gubas A, Terawaki S, Suzuki H, Stadel D, et al. PLEKHM1 regulates autophagosome-lysosome fusion through HOPS complex and LC3/GABARAP proteins. *Mol Cell*. 2015;57 1:39–54; doi: 10.1016/j.molcel.2014.11.006.
48. Caillet-Boudin ML, Buée L, Sergeant N, Lefebvre B. Regulation of human MAPT gene expression. *Mol Neurodegener*. 2015;10:28; doi: 10.1186/s13024-015-0025-8.
49. Moreno-Igoa M, Hernández-Charro B, Bengoa-Alonso A, Pérez-Juana-del-Casal A, Romero-Ibarra C, Nieva-Echebarria B, et al. KANSL1 gene disruption associated with the full clinical spectrum of 17q21.31 microdeletion syndrome. *BMC Med Genet*. 2015;16:68; doi: 10.1186/s12881-015-0211-0.
50. de la Tremblaye PB, Benoit SM, Schock S, Plamondon H. CRHR1 exacerbates the glial inflammatory response and alters BDNF/TrkB/pCREB signaling in a rat model of global cerebral ischemia: implications for neuroprotection and cognitive recovery. *Prog Neuropsychopharmacol Biol Psychiatry*. 2017;79 Pt B:234–48; doi: 10.1016/j.pnpbp.2017.06.021.
51. Elkon R, Ugalde AP, Agami R. Alternative cleavage and polyadenylation: extent, regulation and function. *Nat Rev Genet*. 2013;14 7:496–506; doi: 10.1038/nrg3482.
52. Dick F, Nido GS, Alves GW, Tysnes OB, Nilsen GH, Dölle C, et al. Differential transcript usage in the Parkinson's disease brain. *PLoS Genet*. 2020;16 11:e1009182; doi: 10.1371/journal.pgen.1009182.
53. Zhuravleva E, Gut H, Hynx D, Marcellin D, Bleck CK, Genoud C, et al. Acyl coenzyme A thioesterase Them5/Acot15 is involved in cardiolipin remodeling and fatty liver development. *Mol Cell Biol*. 2012;32 14:2685–97; doi: 10.1128/MCB.00312-12.
54. Consortium G. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45 6:580–5; doi: 10.1038/ng.2653.

Figures

Figure 1

Structure of a full-length SVA and schematic representation of the location of SVA₆₇. (a) A typical SVA element (full-length) contains five domains. Considered 5' to 3', SVAs are composed of a (CCCTCT)_n hexamer repeat, an inverted Alu-like region, a GC-rich variable number tandem repeat (VNTR), a SINE-R domain and a poly A-tail. SVA elements are flanked by target site duplications (TSDs). (b) Schematic modified from UCSC genome browser hg38 of the MAPT locus on chromosome 17. This locus contains multiple genes and four SVAs including the SVA₆₇ RIP (red bar). Six genes (MAPT, KANSL1, CRHR1, LRR37A, PLEKHM1 and ARL17A) were selected for analysis. SVA₆₇ (chr17:46,237,519 – 46,238,226) is anti-sense with respect to the orientation (defined as their 5' TSS) of the LRR37A, CRHR1 and MAPT gene, and shows a sense orientation relative to the KANSL1, ARL17A and PLEKHM1 gene. (c) ENCODE data from UCSC showing the levels of enrichment of histone marks in proximity of SVA₆₇. Signals for H3K27Ac (marker of active regulatory domains associated with active enhancer elements), H3K4Me1 (marker of regulatory domains associated with enhancers) and H3K4Me3 (marker of a regulatory domains associated with promoters) are indicated. The region has been overlaid with a conservation track of several primates and rodents which shows that SVA₆₇ is human specific.

Figure 2

The expression of several gene isoforms at the MAPT locus is associated with SVA_67 genotype. Association of SVA_67 genotype with expression of isoform MAPT-208 (a), KANSL1-214 (b), LRRC37A-202 (c), CRHR1-205 (d), PLEKHM1-212 (e) and ARL17A-203 (f) was analysed by the presence or absence of the SVA RIP (AA, PA, PP) using the NYGC ALS cohort. Wilcoxon test was applied to demonstrate

statistical significance indicated as asterisks. *P <= 0.05, **P <= 0.01, ***P <= 0.001, ****P <= 0.0001, ns > 0.05.

Figure 3

SVA_67 influences expression in an isoform-specific manner. Using the NYGC ALS cohort, isoforms MAPT-202 and MAPT-212 show significantly reduced expression with PP genotype (a), while the opposite effect is observed for isoforms MAPT-217 and MAPT-208 (b). Wilcoxon test was applied to demonstrate

statistical significance indicated as asterisks. *P <= 0.05, **P <= 0.01, ***P <= 0.001, ****P <= 0.0001, ns > 0.05. (c) Overview of the MAPT gene and all its isoforms as shown on Ensembl. Boxes represent exons, connecting lines introns. Filled boxes show coding sequence, and empty boxes UTRs (untranslated regions). Analysed isoforms of interest in (a) and (b) are highlighted as asterisks (red/black).

Figure 4

Tissue specific influence of SVA_67 genotype. Using the PPMI cohort, isoforms ARL17A-203 and CRHR1-203 show significantly reduced expression with PP genotype (a, c), while the opposite effect is observed when using the NYGC ALS cohort (b, d). Wilcoxon test was applied to demonstrate statistical significance indicated as asterisks. *P <= 0.05, **P <= 0.01, ***P <= 0.001, ****P <= 0.0001, ns > 0.05.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AdditionalFile1.pdf](#)