Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Characterisation of retrotransposon insertion polymorphisms in whole genome sequencing data from individuals with amyotrophic lateral sclerosis

Abigail L. Savage [a], Alfredo Iacoangeli [b,c], Gerald G. Schumann [d], Alejandro Rubio-Roldan [e], Jose L. Garcia-Perez [e,f], Ahmad Al Khleifat [b], Sulev Koks [g,h], Vivien J. Bubb [a], Ammar Al-Chalabi [b,i], John P. Quinn [a,*]

[a] *Department of Pharmacology and Therapeutics, Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L69 3BX, UK*
[b] *Maurice Wohl Clinical Neuroscience Institute, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London SE5 9RT, UK*
[c] *Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London SE5 8AF, UK*
[d] *Division of Medical Biotechnology, Paul-Ehrlich-Institut, Langen 63225, Germany*
[e] *Department of Genomic Medicine and Department of Oncology, GENYO, Centre for Genomics & Oncology, PTS Granada, 18007, Spain*
[f] *MRC-HGU Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK*
[g] *Perron Institute for Neurological and Translational Science, Perth, Western Australia 6009, Australia*
[h] *Centre for Molecular Medicine and Innovative Therapeutics, Murdoch University, Perth, Western Australia 6150, Australia*
[i] *Department of Neurology, King's College Hospital, London SE5 9RS, UK*

ARTICLE INFO

Edited by: X. Carette

Keywords:
Amyotrophic lateral sclerosis
Retrotransposon
Polymorphism
Whole genome sequencing
LINE-1
Alu
SVA

ABSTRACT

The genetics of an individual is a crucial factor in understanding the risk of developing the neurodegenerative disease amyotrophic lateral sclerosis (ALS). There is still a large proportion of the heritability of ALS, particularly in sporadic cases, to be understood. Among others, active transposable elements drive inter-individual variability, and in humans long interspersed element 1 (LINE1, L1), *Alu* and SINE-VNTR-Alu (SVA) retrotransposons are a source of polymorphic insertions in the population. We undertook a pilot study to characterise the landscape of non-reference retrotransposon insertion polymorphisms (non-ref RIPs) in 15 control and 15 ALS individuals' whole genomes from Project MinE, an international project to identify potential genetic causes of ALS. The combination of two bioinformatics tools (mobile element locator tool (MELT) and TEBreak) identified on average 1250 *Alu*, 232 L1 and 77 SVA non-ref RIPs per genome across the 30 analysed. Further PCR validation of individual polymorphic retrotransposon insertions showed a similar level of accuracy for MELT and TEBreak. Our preliminary study did not identify a specific RIP or a significant difference in the total number of non-ref RIPs in ALS compared to control genomes. The use of multiple bioinformatic tools improved the accuracy of non-ref RIP detection and our study highlights the potential importance of studying these elements further in ALS.

1. Introduction

The rapidly progressive neurodegenerative disease amyotrophic lateral sclerosis (ALS) is clinically and genetically heterogeneous and characterised by the loss of motor neurons in the brain and spinal cord. The processes resulting in this neurodegeneration are not fully understood, however several mechanisms that are thought to be involved include protein aggregation, oxidative stress, mitochondrial dysfunction, axonal transport, inflammation and RNA processing and toxicity (van Es et al., 2017). An individual's risk of developing ALS is a complex interaction of genetic and environmental factors with the proportion of attributable risk from the genetic and environmental

Abbreviations: ALS, amyotrophic lateral sclerosis; HERV, human endogenous retroviruses; LINE1/L1, long interspersed element 1; LD, linkage disequilibrium; LTR, Long Terminal Repeat; MELT, mobile element locator tool; RIPs, retrotransposon insertion polymorphisms; SINE, short interspersed elements; SNPs, single nucleotide polymorphisms; SVA, SINE-VNTR-Alu; TEs, Transposable Elements; VNTR, variable number tandem repeat; WGS, whole genome sequencing.

[*] Corresponding author.
E-mail address: jquinn@liv.ac.uk (J.P. Quinn).

https://doi.org/10.1016/j.gene.2022.146799
Received 24 March 2022; Received in revised form 15 July 2022; Accepted 5 August 2022
Available online 10 August 2022

components variable between individuals (Al-Chalabi and Hardiman, 2013). There have been over 30 different genes linked to familial ALS and the genetic component has been identified in ~ 70 % of cases; the most commonly mutated genes in Europeans are *C9orf72*; *SOD1*; *TARDBP* and *FUS* (Mathis et al., 2019; Renton et al., 2014). Traditionally ALS has been separated into two groups, the approximately 10 % of individuals who have a family history of the disease with the remaining 90 % classified as sporadic. However, there is significant overlap between the two as many of the genes involved in familial ALS are also involved in the sporadic form (Al-Chalabi and Lewis, 2011; Brown and Al-Chalabi, 2017). In addition, genetics is an important factor in an individual's risk of developing sporadic ALS as the heritability is estimated at 60 %, and first-degree relatives of sporadic ALS patients have an 8-fold increased risk of developing the disease (Hanby et al., 2011; Al-Chalabi et al., 2010).

Research into the genetics of ALS has identified an increasing number of variants involved in disease susceptibility, but the majority of the heritability remains to be elucidated. We hypothesise that part of this unidentified heritability is due to classes of genetic variation that have not yet been studied extensively in large ALS and control cohorts. One such type of variation is driven by Transposable Elements (TEs) (Kazazian and Moran, 2017), mobile pieces of DNA that can drive genetic variability among individuals. Of the different classes of TEs, retrotransposons are typically active in mammals. Retrotransposons can be subdivided in Long Terminal Repeat (LTR)-containing (including human endogenous retroviruses or HERVs) and non-LTR retrotransposons, and can create new insertions within genomes through a 'copy-&-paste' mechanism. Members of the non-LTR retrotransposon group in the human genome include the families long interspersed element class 1 (LINE-1 or L1), and short interspersed elements (SINEs), majority of which are *Alus,* and the composite element SINE-VNTR-Alu (SVA). As these elements continue to mobilize in humans, their activity generates presence/absence insertion polymorphisms in the human population. Recent estimates have established that the germline retrotransposition rates in humans are 1/40 births for *Alu*, and 1/63 for both L1 and SVA elements (Feusier et al., 2019). The 1000 Genomes Project data revealed that the average European genome harbours 919 *Alu*, 123 L1 and 53 SVA non-reference retrotransposon insertions (Genomes Project C et al., 2015), however this is likely to be an underestimation due to the low coverage of the sequencing data. This highlights the contribution of non-reference retrotransposon insertion polymorphisms (non-ref RIPs) to the genetic variation of the human genome, as the ongoing retrotransposition of these elements results in insertions that are unique to specific individuals or enriched in specific ancestral groups. To date, 124 cases of genetic diseases have been demonstrated to be caused by retrotransposon insertions, including cystic fibrosis (*Alu*), heamophilia A (L1), X-linked dystonia-parkinsonism (SVA) and hereditary cancers (Hancks and Kazazian, 2016). However, these cases represent so-called Mendelian diseases caused by very deleterious mutations that are expressed with high penetrance. Retrotransposon insertions have also been suggested as candidate causal variants at an increasing range of disease loci due to their strong linkage disequilibrium (LD) with trait associated single nucleotide polymorphisms (SNPs) (Payer et al., 2017). It was shown earlier that TE insertion polymorphisms also exert regulatory effects on the human genome (Wang et al., 2017). Specifically, polymorphic TE insertions were shown to contribute to both inter-individual and population-specific differences in gene expression and to facilitate the re-wiring of transcriptional networks. A subset of genomic retrotransposons insertions can be expressed from their own promoter and several of these retrotransposon families have been reported to be transcriptionally dysregulated in ALS (Douville et al., 2011; Li et al., 2012, 2015r; Pereira et al., 2018; Prudencio et al., 2017; Savage et al., 2019; Tam et al., 2019a,b). Of particular interest was the study by Tam et al., 2019b who identified three subtypes of ALS based on their frontal and motor cortices transcriptomic profile, one of which was characterised by retrotransposon overexpression and TAR DNA-binding protein 43 dysfunction and was observed in 20 % of ALS patients (Tam et al., 2019b).

TE-derived sequences have contributed a wide variety of gene regulatory elements to the human genome (Chuong et al., 2017; Elbarbary et al., 2016; Faulkner et al., 2009; Rebollo et al., 2012), including promoters (Conley et al., 2008; Marino-Ramirez et al., 2005), enhancers (Chuong et al., 2016; Chuong et al., 2013; Notwell et al., 2015), transcription terminators (Conley and Jordan, 2012) and several classes of small RNAs (Kapusta et al., 2013; Piriyapongsa et al., 2007; Weber, 2006). Human TEs can also influence gene regulation by modulating various aspects of chromatin structure throughout the genome (Jacques et al., 2013; Lander et al., 2001; Schmidt et al., 2012; Sundaram et al., 2014; Wang et al., 2015). We have recently demonstrated that decreased methylation can be observed over specific L1 elements in the CNS which would be consistent with the model for these domains having an influence on genomic regulation in the brain (Savage et al., 2020).

Recent findings have revealed potential mechanistic links between polymorphic TE-induced gene regulatory changes and the endophenotypes that underlie human health and disease (Wang et al., 2017). For example, the disruption of the *B4GALT1* enhancer by an SVA insertion is associated with down-regulation of the gene in B-cells. *B4GALT1* encodes a glycosyltransferase that functions in the glycosylation of the Immunoglobulin G (IgG) antibody in such a way as to convert its activity from pro- to anti-inflammatory (Lauc et al., 2013). Down-regulation of this gene in individuals with the enhancer SVA insertion should thereby serve to keep the IgG molecule in a pro-inflammatory state. Consistent with this idea, the *B4GALT1* enhancer SVA insertion is linked to a genomic region implicated by genome wide association studies (GWAS) in both inflammatory conditions and autoimmune diseases such as systemic lupus erythematosus and Crohn's disease (Wang et al., 2017; Lauc et al., 2013). Changes in mRNA and protein expression of retrotransposons, both of the non-LTR class and human endogenous retroviruses, have been linked to ALS although this potential role in disease development needs to be fully understood (Pereira et al., 2018; Savage et al., 2020; Mayer et al., 2018). Retrotransposon activity has also been linked to several other neurological conditions including multiple sclerosis, Rett Syndrome, Aicardi-Goutieres syndrome, Autism-spectrum disorder and Alzheimer's disease (Faulkner and Billon, 2018; Saleh et al., 2019; Tam et al., 2019b). The mechanisms implicated in this pathogenicity are thought to occur through an inflammatory response, and it has been proposed that the presence of RNAs and DNAs from these elements could induce neurotoxicity (Volkman and Stetson, 2014).

Here, we focus on the characterisation of polymorphic non-LTR retrotransposons and their contribution to genetic variation within known ALS haploblocks as a first step towards an understanding of how these might influence gene expression at these loci. We hypothesise that non-ref RIPs could play a role in the genetic heritability of ALS, either through common variants contributing to risk or through rare or even *de novo* insertions that could generate a phenotype. Large-scale whole genome sequencing (WGS) projects, such as Project MinE (https://www.projectmine.com) (Project MinE, 2018), are paving the way to enable these elements to be studied. However, such studies are computationally intense, and specialist bioinformatics tools are required to characterise this type of genetic variation accurately (Ewing, 2015; Goerner-Potvin and Bourque, 2018). Therefore, we undertook an objective comparison to validate two such tools, mobile element locator tool (MELT) (https://melt.igs.umaryland.edu) (Gardner et al., 2017) and TEBreak (https://github.com/adamewing/tebreak) (Salvador-Palomeque et al., 2019; Schauer et al., 2018), to characterise non-ref RIPs in a pilot study of 15 ALS and 15 control whole genomes from Project MinE. MELT, a tool developed to be used on a population scale, has been used to genotype over 2500 genomes from the 1000 Genomes Project (Gardner et al., 2017; Sudmant et al., 2015) and has recently been used as part of the genome aggregation database (gnomAD) to characterise *Alu*, L1 and SVA insertions in over 10,000 genomes and this data is available on the gnomAD-SV browser (https://gnomad.broadinstitute.org/). We were able to compare the non-ref RIPs identified by each tool and characterise the genetic landscape of these variants in known ALS-associated regions. Although this pilot study is too small to enable case-control comparisons of the frequency of each insertion, we expanded the number of samples analysed using PCR amplification and LD analysis with known disease associated SNPs. We investigated whether characterisation of polymorphic TE insertions may allow us to better understand their ability to influence the outcome of this neurological disorder.

2. Results

2.1. Comparison of two bioinformatics tools, TEBreak and MELT, to identify non-reference RIPs in WGS data

The analysis of WGS data from 30 whole genomes which represent 15 control and 15 ALS individuals from the UK samples of Project MinE (Project MinE, 2018) sequenced to a depth of ~ 30x to identify polymorphic retrotransposon insertions, yielded a different number for non-ref RIPs of each non-LTR retrotransposon family per genome (*Alu*, L1 and SVA) when the two tools TEBreak and MELT were compared (Fig. 1A-C). TEBreak identified fewer *Alu* non-ref RIPs (on average 991 vs 1113), a greater number of L1 non-ref RIPs (on average 179 vs 157) and a greater number of SVA non-ref RIPs (on average 56 vs 44) per genome compared to MELT. The differences in the average number of non-ref RIPs observed between TEBreak and MELT were significantly different: *Alu*, $p = 5x10^{-24}$; L1, $p = 3.0x10^{-12}$; SVA, $p = 5x10^{-11}$. The mean number of unique non-ref RIPs per genome identified in total (data from TEBreak and MELT combined) were as follows: *Alu*, 1250; L1, 232; SVA, 77. The loci of non-ref RIPs identified by both TEBreak and MELT per genome showed the greatest concordance for *Alus* (68 %) followed by L1s (45 %) and SVAs (26 %) (Fig. 1D). In total across the 30 genomes analysed, TEBreak identified 4129 *Alu*, 802 L1 and 344 SVA unique non-ref RIPs compared to MELT, that identified a total of 4305 *Alu*, 785 L1 and 306 SVA unique non-ref RIPs. Using a bed file provided with TEBreak of 38,512 non-ref RIPs previously identified we determined how many of the insertions detected in the study had been reported previously. This bed file was compiled from insertions identified from 15 publications, including from the 1 k genomes project analysis, and the dbRIP database (Ewing, 2015; Sudmant et al., 2015). Of the *Alu* non-ref RIPs identified by TEBreak and MELT 83.1 % and 82.5 % respectively had been reported previously. For the L1 non-ref RIPs this was 76.8 % of those identified by TEBreak and 66.8 % of those by MELT and SVAs 59.3 % by TEBreak and 60.5 % by MELT.

There are multiple subfamilies of *Alu*, L1 and SVA retrotransposons that have shown different dynamics of expansion during evolution (Lander et al., 2001; Batzer and Deininger, 2002; Wang et al., 2005). Both TEBreak and MELT classify the *Alu* and L1 non-ref insertions by subfamily, however MELT does not determine the subtype of each SVA insertion whereas TEBreak does. In addition, many of the L1 insertions were classified by MELT as ambiguous, therefore here we report on the subfamily distribution using TEBreak data alone. The largest proportion (66 %) of non-ref RIP *Alus* belonged to the currently most active subfamily *AluYa5* (Figure S1A), which represents the most abundant subfamily with the highest levels of mobilization in humans (Lander et al., 2001; Gardner et al., 2017; Batzer and Deininger, 2002; Bennett et al., 2008; Konkel et al., 2015; Mills et al., 2007; Wang et al., 2006). Consistently, the L1 non-ref RIPs were classified into the subfamilies L1PA2, L1preTa and L1Ta, and 79 % of the insertions were members of the youngest and human-specific subfamily L1Ta (Figure S1B). The SVA non-ref RIPs were divided into the subtypes A-F and the majority belonged to the younger subtypes D-F with the largest proportion (55 %) classified as F (Figure S1C). For all three of the non-LTR retrotransposon families, there was no significant difference in the distribution of the subfamilies between control and ALS genomes.

2.2. PCR validation of non-reference RIPs

Non-ref RIPs from each non-LTR retrotransposon family (12 *Alu*, 11

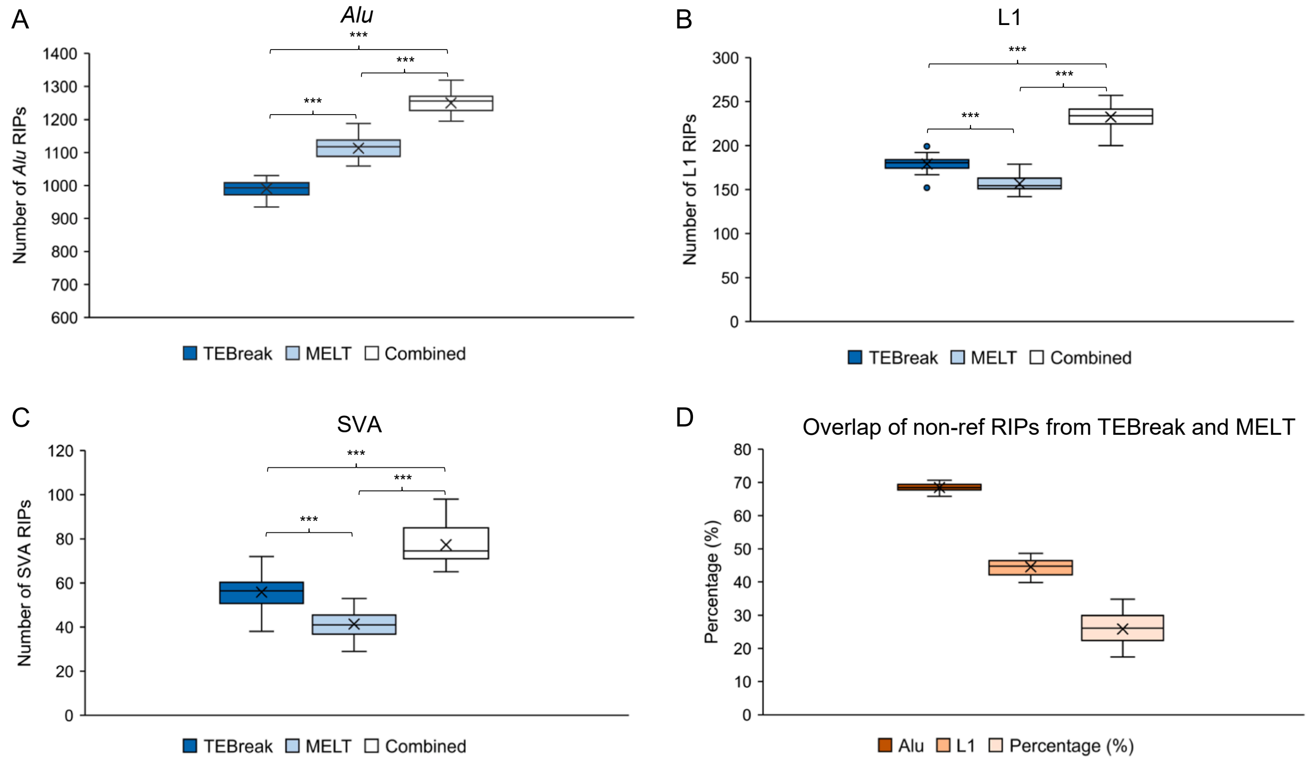

Fig. 1. The average number of non-reference RIPs identified per individual differed between the bioinformatic tool MELT and TEbreak used to analyse WGS datasets. A –C Numbers of non-ref *Alu*, L1 and SVA RIPs identified in each genome by TEBreak, MELT and the two tools in combination, respectively. D – The percentage overlap of non-ref *Alu*, L1 and SVA RIPs that were identified by both TEBreak and MELT. Two-tailed T-test ***=p < 0.001.

L1 and 6 SVA insertions) were validated by PCR in multiple DNA samples (a minimum of 9 for each insertion and a total of 335 PCRs overall) to include individuals with different genotypes (homozygous present, heterozygous and homozygous absent) called by TEBreak and MELT (Fig. 2). Of the 29 insertions, chosen for PCR validation 93 % were called by both tools in at least one genome. Five of the retrotransposon insertions were chosen for validation based on their presence in introns of ALS associated genes, two were retrotransposition competent L1s and the remainder were chosen at random. There were no false positives identified in this validation, namely insertions called present by the bioinformatics tools but absent in the PCR assay. However, there were multiple instances of false negatives where the retrotransposon insertion was present in the PCR validation assay but not identified by TEBreak or MELT. We did not further validate by sequencing analyses of the obtained PCR products. The false negative and false positive rates, accuracy, sensitivity and specificity for each software tool were calculated (Table 1) and the raw numbers are reported in Table S1. Of the three retrotransposon families, the *Alu* non-ref RIPs were the most accurately called by TEBreak and MELT with the lowest false negative rate and highest accuracy and sensitivity (Table 1). SVAs were the most poorly called non-ref RIPs by both tools, most likely because they harbour long tracts of tandem repeats consisting of 35–50 repeat copies that can accumulate to ~ 2300 bp, making them difficult to characterise from short read sequencing data (Wang et al., 2005; Chaisson et al., 2015). In addition, individual SVA elements can be polymorphic in length due to the differing numbers of repeat copies their variable number tandem repeat (VNTR) regions contain and arise after their de novo insertion into a particular locus; this adds a further layer of complexity when analysing non-ref SVA RIPs. This is exemplified by the SVA insertion validated in Fig. 2C, where two different length alleles of the SVA insertion can be detected (Fig. 2C, compare lanes 8 and 9). The specific changes were not further determined by sequence analysis of the obtained PCR product. The false negative rate, accuracy and sensitivity were very similar between the two tools for calling *Alus* and SVAs. Notably, combining the calls made by both tools increased the accuracy and sensitivity and reduced the false negative rate in all three retrotransposon families, with perfect agreement between the combined calls for the polymorphic *Alus* and their PCR validation. In some instances the false negatives of a particular insertion was due to the lack of detection by the caller in any of the genomes analysed, for example an *Alu* in the intron of the *TRNAU1AP* gene was not detected by MELT, whereas other false negatives were due to the lack of detection in a particular genome but was identified in other individuals by the caller, for example an *Alu* in the *UNC13A* gene. Further inspection of the insertion site of the *Alu*

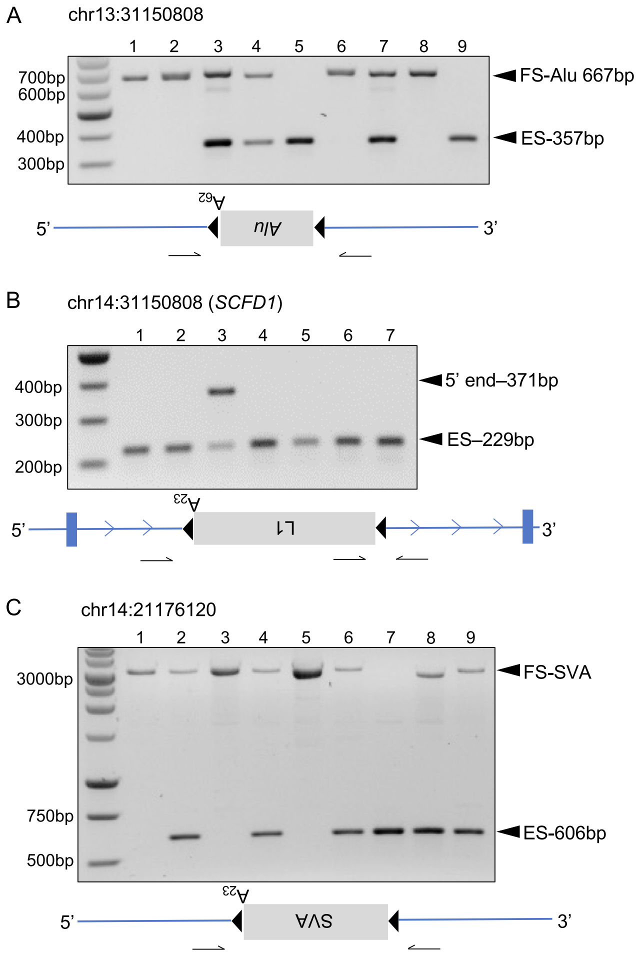

Fig. 2. PCR validation of non-reference RIPs identified in WGS datasets using TEBreak and MELT. (a)- Representative gel image of 9 different ALS and control individuals (lanes 1–9) for the PCR validation of a specific intergenic *Alu* insertion into chr13 using one primer pair to amplify the empty site (357 bp) and the filled site (667 bp). 147 *Alu* validation PCRs were carried out for 12 different *Alu* non-ref RIPs in a minimum of 9 ALS and control individuals each whose genomes had been analysed using TEBreak and MELT. (b)– Representative gel image of 7 different individuals (lanes 1–7) for the multiplex PCR validation of an intronic full length L1 insertion in the *SCFD1* gene at position 31,150,808 of chromosome 14. Primers flanking the insertion site were used to generate an empty site product (229 bp) and a primer in the 5′end end of the L1 sequence to generate a PCR product (371 bp) if L1 was present. 117 L1 non-ref RIP validation PCRs were carried out for 11 different L1 RIPs in a minimum of 9 ALS and control individuals each whose genomes had been analysed using TEBreak and MELT. (c)– Representative gel image of 9 different individuals (lanes 1–9) for the PCR validation of a SVA insertion into chr14 using one primer pair to amplify the empty and filled sites. The empty site PCR product was 606 bp; the filled site was approximately 3000 bp, however TEBreak or MELT could not definitively predict the length of the filled site product due to the large size and possible variability in length due to the VNTR domain. 71 SVA non-ref RIP validation PCRs were carried out for 6 different SVA RIPs in a minimum of 9 ALS and control individuals each whose genomes had been analysed using TEBreak and MELT. ES – empty site, FS – filled site.

Table 1
The false negative and false positive rates, accuracy, sensitivity and specificity for each family of non-ref RIPs are reported with 95% confidence intervals.

		TEBreak	MELT	Combined
Alu	False positive rate	0 (0, 0.05)	0 (0, 0.05)	0 (0, 0.05)
	False negative rate	0.04 (0.01, 0.11)	0.08 (0.03, 0.17)	0 (0, 0.05)
	Accuracy	0.98 (0.94, 1.00)	0.96 (0.91, 0.98)	1.00 (0.95, 1.00)
	Sensitivity	0.96 (0.89, 0.99)	0.92 (0.83, 0.97)	1.00 (0.95, 1.00)
	Specificity	1.00 (0.95, 1.00)	1.00 (0.95, 1.00)	1.00 (0.95, 1.00)
L1	False positive rate	0 (0, 0.06)	0 (0, 0.06)	0 (0, 0.06)
	False negative rate	0.21 (0.11, 0.35)	0.10 (0.03, 0.21)	0.06 (0.01–0.16)
	Accuracy	0.91 (0.84, 0.95)	0.96 (0.90, 0.99)	0.97 (0.93, 0.99)
	Sensitivity	0.79 (0.65, 0.89)	0.90 (0.79, 0.97)	0.94 (0.84, 0.99)
	Specificity	1.00 (0.94, 1.00)	1.00 (0.94, 1.00)	1.00 (0.94, 1.00)
SVA	False positive rate	0 (0, 0.08)	0 (0, 0.08)	0 (0, 0.08)
	False negative rate	0.35 (0.17, 0.56)	0.38 (0.20, 0.59)	0.19 (0.07, 0.39)
	Accuracy	0.87 (0.77, 0.94)	0.86 (0.76, 0.93)	0.93 (0.84, 0.98)
	Sensitivity	0.65 (0.44, 0.83)	0.62 (0.41, 0.80)	0.81 (0.61, 0.93)
	Specificity	1.00 (0.92, 1.00)	1.00 (0.92, 1.00)	1.00 (0.92, 1.00)

into the intron of the *TRNAU1AP* gene in the gnomAD-SV browser (https://gnomad.broadinstitute.org/) identified an insertion of 388 bp at the site reported by TEBreak detected by the structural variant callers DELLY and Manta. MELT has also been used to call variants in the genomes analysed in the gnomadSV browser and the lack of detection in over 10,000 genomes suggests it is not particular features of the genomes analysed in our study but difficulties in MELT detecting this insertion at this particular locus.

2.3. Distribution of non-reference RIPs across ALS and control genomes

There was no significant difference in the overall number of *Alu*, L1 and SVA non-ref RIPs in the genomes of individuals with ALS compared to controls using data combined from both tools (Fig. 3). Of the total number of unique non-ref RIPs (5101 *Alu*, 1086 L1 and 480 SVAs) identified across the 15 control and 15 ALS genomes, 45 %, 37 % and 44 % of *Alu*, L1 and SVA insertions, respectively, occurred into genes. The majority of those genic insertions were in introns, however 49 *Alu*, 3 L1 and 2 SVA insertions occurred into non-coding exons, and 3 *Alu* and 2 L1s in coding exons. The largest proportion of non-ref RIPs located in non-coding exons inserted in the 3′UTR of genes (72.2 %) followed by the exons of ncRNAs (22.2 %) and the remainder were in the 5′UTR of genes (5.6 %). The non-ref RIPs identified in coding exons occurred in both controls and ALS genomes or had been reported previously in the literature with an updated list available at Github (*https://github.com/adamewing/tebreak*) to complement the following manuscript (Ewing, 2015).

Pathway analysis was performed to identify the function of the genes containing non-ref RIPs and identified a 1.23-fold enrichment for those genes expressed in brain tissue (Bonferroni corrected p = 1.2x10^{-18}) (Fig. 4A). There was also an enrichment for genes encoding proteins located in cellular components related to synapses such as the postsynaptic membrane (2.5-fold Bonferroni corrected p = 8.1x10^{-7}) and postsynaptic density (2.6-fold Bonferroni corrected P = 1.9x10^{-6}) (Fig. 4B). This bias towards neuronal pathways reflects hereditary or early developmental somatic variation but does not include adult brain somatic insertions. Notably, TEBreak and MELT analysis of the 30 genomes identified 10 *Alu* and 1 L1 non-ref RIPs within the introns of 5/32 ALS-associated genes analysed (Table 2), and these were found to be common insertions occurring in multiple individuals. Eight of these non-ref RIPs were located in *ERBB4*, although in 4 different introns, a large gene 1.16 Mb in length; however the number of RIPs per kb in *ERBB4* was one of the lowest out of these five ALS-associated genes (Table 2). The majority of the insertions into the *ERBB4* gene were in introns towards the 5′ end of the gene (introns 1 and 2) where the largest introns are located. The insertions into *ERBB4* were characterised by the hallmarks of L1 medicated target primed reverse transcription, such as target site duplications 10–24 bp in length.

In an initial experiment to address these RIPs in a larger N number the allele frequencies for 2 *Alu* and 1 L1 RIPs located in the ALS-associated genes (*SPAST*, *FIG4* and *ERBB4*) were determined through PCR amplification in ALS and controls from the MNDA UK DNA bank cohort (Table 3). However, there was no significant association of these particular variants with ALS identified.

In 11 haploblocks containing SNPs associated with ALS through GWAS, there were 47 *Alu*, 7 L1 and 6 SVA non-ref RIPs identified, and the majority of these insertions occurred in multiple individuals (Table 4). Of those non-ref RIPs in ALS-associated haploblocks, 5 *Alu*, 1 L1 and 1 SVA insertions were located in introns of the reported genes linked to the GWAS SNPs (*UNC13A*, *SCFD1*, *SARM1* and *DPP6*). An L1 and *Alu* non-ref RIP had inserted into the same intron that contained an ALS-associated SNP in the genes *SCFD1* (rs10139154) and *UNC13A* (rs12608932) respectively (Table 4). However, the GWAS SNPs rs10139154 and rs12608932 did not tag either of these two non-ref RIPs, (the intronic full length L1Ta in *SCFD1* and the intronic *AluYb9* in *UNC13A*), when analysed for linkage disequilibrium (LD) between the variants. On analysis of *UNC13A* in 196 individuals from the MNDA UK DNA bank for which the SNP and *Alu* genotypes were available (a subset of those in genotyped Table 3), the presence of the *Alu* and the non-risk allele of the SNP were inherited together (D’=1). Although the presence of the *Alu* insertion was predictive of the non-risk allele of rs12608932, it was not true for the reverse. This can be attributed to the lower frequency of the present allele of the *Alu* insertion compared to the frequency of the non-risk allele of rs12608932 (r^2 = 0.06). In 192 individuals from the MNDA UK DNA bank (a subset of those genotyped in Table 3) analysed, the L1 non-ref RIP in *SCFD1* was not in LD with rs10139154 (r^2 = 0.03, D’=0.6). Association analysis of the *Alu* insertion in *UNC13A* and the *L1* insertion in SCFD1 in the MNDA UK DNA bank samples did not identify an association with disease (Table 3). While we could not find strong non-ref RIP candidates linked to ALS, our data does not rule out that such transposable element insertions could play a role in ALS progression or severity by modifying genomic parameters such as RNA splicing or gene expression in individuals harbouring a known genetic risk.

3. Discussion

Candidate gene and large-scale sequencing studies have identified many genes and genetic variants that are involved in the development of or an increased susceptibility to ALS (Mejzini et al., 2019). The majority of these studies have addressed single nucleotide variants, as they are the easiest and most cost effective to study on a large scale. However, initiatives such as Project MinE (https://www.projectmine.com/) are enabling the analysis of whole genome sequencing data to address additional types of genetic variation, such as structural variants, tandem repeats and the presence or absence of retrotransposon insertions, to aid the discovery of novel variants contributing to disease susceptibility.

We utilised two bioinformatics tools, TEBreak and MELT, to carry out

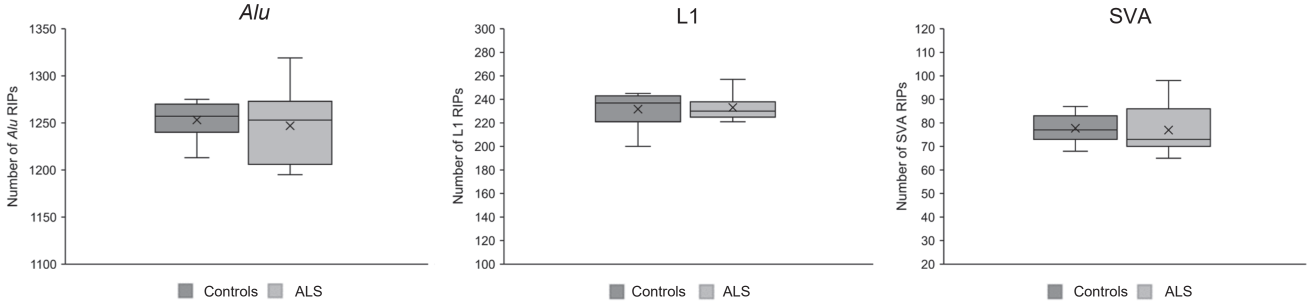

Fig. 3. Identification of comparable numbers of non-reference RIPs in control and ALS genomes. The number of *Alu*, L1 and SVA non-ref RIPs identified in the genome of each individual using TEBreak and MELT were combined and no significant difference in the number in those with ALS compared to the unaffected control was found (two-tailed T-test). In total, the number of different non-ref RIPs found across the 30 individuals were as follows: 5101 Alu, 1086 L1 and 480 SVA insertions (controls n = 15, ALS n = 15).

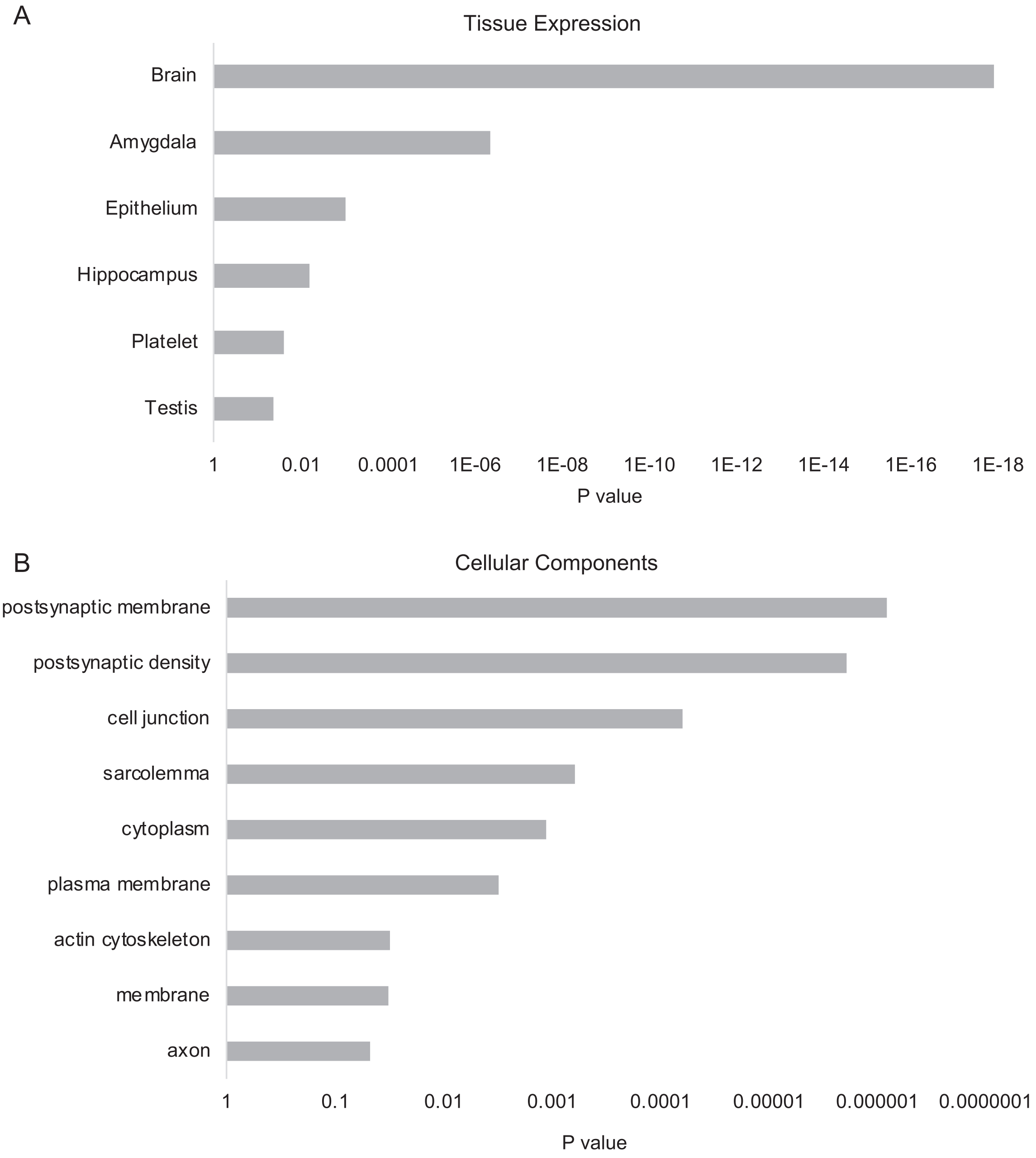

Fig. 4. Genes containing a non-reference RIP were enriched for expression in the brain and neuronal structures. (a) - Six tissues were significantly enriched for the expression of genes containing a non-ref RIP (Bonferroni adjusted p-value > 0.05). (b) – The proteins encoded by genes containing a non-ref RIP were significantly enriched in particular cellular components, most frequently those relating to neurons (Bonferroni adjusted p-value > 0.05).

Table 2
Numbers of different non-ref RIPs located in introns of ALS-linked genes and identified using TEBreak and MELT to analyse WGS data.

Gene coordinates (hg19)			Gene	Number of RIPs			No. of RIPs per kb
				Alu	L1	SVA	
chr2	32,288,680	32,382,706	*SPAST*	1	-	-	0.011
chr2	202,564,986	202,645,895	*ALS2*	1	–	–	0.012
chr2	212,240,442	213,403,352	*ERBB4*	7	1	–	0.007
chr6	110,012,424	110,146,634	*FIG4*	1	–	–	0.007
chr12	111,890,018	112,037,480	*ATXN2*	1	–	–	0.007

a small pilot study to analyse polymorphic non-reference non-LTR retrotransposon insertions in ALS and control genomes from Project MinE. Our analysis on average identified 1250 *Alu*, 232 L1 and 77 SVA non-ref RIPs per genome, which is higher than the number identified in the 1000 Genomes Project data set (for Europeans – 919 *Alu*, 123 L1 and 53 SVA insertions) using MELT alone. This may be attributed to the higher depth of coverage of the Project MinE sequencing and the use of two tools to complete the analysis. An additional factor to consider is read length, as the Project MinE UK samples are 150 bp in length whereas the genomes from Phase 3 of the 1000 genome are a range of read length > 70bps. Combining the calls made by both tools generated the highest level of accuracy when compared to the PCR validation (Fig. 2) and RIP detection was greatly improved when using two callers, however the use of two tools for genotyping a large number of genomes can be impractical. The VCF (variant caller format) output from MELT enables the data to be

Table 3

The insertion allele frequencies (IAF) of non-ref RIPs in ALS linked genes or haploblocks reveal that there is no significant difference in the frequency between ALS cases and controls (chi-squared test). Five non-ref RIPs from table 1 or 3 were chosen to be PCR validated in a larger number of controls and ALS samples from the MNDA UK DNA bank cohort.

					IAF		Number of samples	
Coordinates (hg19)			Family	Gene	Controls	ALS	Controls	ALS
chr2	32,365,841	32,365,856	*Alu*	*SPAST*	0.07	0.03	50	54
chr6	110,102,976	110,102,981	*Alu*	*FIG4*	0.66	0.67	48	53
chr19	17,752,493	17,752,493	*Alu*	*UNC13A*	0.09	0.08	143	142
chr2	213,114,162	213,114,173	L1	*ERBB4*	0.31	0.31	48	54
chr14	31,150,808	31,150,825	L1	*SCFD1*	0.12	0.12	142	137

Table 4

Numbers of different non-ref RIPs that were located in haploblocks containing SNPs associated with ALS risk and identified using TEBreak and MELT to analyse WGS data. The list of SNPs was taken from the GWAS catalog and filtered to retain those SNPs that were identified in European populations and reached genome wide significance (p-value$<5x10^{-8}$).

					Number of RIPs			
Haploblock (hg19)			GWAS SNP	Reported Gene	*Alu*	L1	SVA	No. of RIPs per kb
chr3	38,356,116	40,221,298	rs616147	*MOBP*	2	1	-	0.002
chr5	148,662,624	150,561,298	rs10463311	*TNIP1*	3	–	2	0.003
chr7	153,674,019	154,964,730	rs10260404	*DPP6*	4	–	1	0.004
chr9	26,111,757	28,224,283	rs2814707, rs3849942, rs3849943, rs10122902	*C9orf72*	3	–	–	0.001
chr12	57,548,860	59,308,666	rs113247976,	*KIF5A*	6	–	1	0.004
chr12	64,032,461	65,559,695	rs74654358	*TBK1*	2	–	–	0.001
chr14	29,972,145	32,383,265	rs10139154	*SCFD1*	4	3	–	0.003
chr17	21,290,357	27,334,244	rs35714695, rs34517613	*SARM1*	12	2	2	0.003
chr19	16,374,416	18,409,862	rs12608932	*UNC13A*	6	–	–	0.003
chr21	32,668,642	34,376,999	rs13048019	*SOD1*	4	1	–	0.003
chr21	44,506,268	46,177,105	rs75087725	*C21orf2*	1	–	–	0.0006

integrated with other types of genetic variation and to be used in downstream analyses more easily than the text output from TEBreak. MELT has been used both in the 1000 Genomes Project (Gardner et al., 2017; Sudmant et al., 2015) and the genome aggregation database (gnomAD) to characterise *Alu*, L1 and SVA insertions and this data is available on the gnomAD-SV browser (https://gnomad.broadinstitute.org/). Therefore, at present MELT is the most suitable tool of the two for use in large-scale analyses such as for the Project MinE dataset if a single tool was to be used.

Although our study is comparatively small, it has highlighted the need to validate targets identified by the bioinformatics analyses due to the inconsistencies found between the two tools and the false negatives identified when the insertions were PCR validated (Fig. 2). We considered PCR an appropriate tool to validate the RIPs identified with our bioinformatic pipeline. Additional variation within the VNTR regions, both central and flanking CT domain, of SVA elements is not captured by bioinformatics tools addressing presence or absence, such as TEBreak and MELT, and is difficult to characterise in short read sequencing data. The importance of changes in the length of an SVA insertion is highlighted by the disease causing SVA insertion in *TAF1* that causes X-linked dystonia parkinsonism. In this case the length of the 5′ CCCTCT repeat within the SVA has been shown to modify the age on onset of the disease (Bragg et al., 2017). Therefore, further analyses of polymorphism intrinsic within an SVA insertion is required to determine if it is also an important parameter for disease risk.

De novo retrotransposition events, depending on the site of their insertion, could affect gene function through multiple mechanisms such as loss of function mutations, exonisation and changes in mRNA expression and splicing (Savage et al., 2019; Chuong et al., 2017; Payer and Burns, 2019). By using the expression quantitative trait loci (eQTL) analytical paradigm, hundreds of associations between human TE insertion variants and gene expression levels have been discovered (Wang and Jordan, 2018). These include population-specific gene regulatory effects as well as coordinated changes to gene regulatory networks. Additionally, analyses of linkage disequilibrium patterns with previously characterized GWAS trait variants have uncovered TE insertion polymorphisms that are likely causal variants for a variety of common complex diseases. For instance, applying the eQTL approach uncovered that polymorphic TE loci are associated with differences in expression between European and African population groups, and a single *Alu* locus was shown to be indirectly associated with the expression of numerous genes via the regulation of the B cell-specific transcription factor *PAX5* (Wang et al., 2017). A recent study identified 211 and 176 polymorphic *Alu*, L1 and SVA insertions that affected gene expression in lymphoblastoid cell lines and human induced pluripotent stem cells respectively (Goubert et al., 2020). One of these transposable element quantitative trait loci identified was an *Alu* insertion in the *ALS2* gene that we also found as a non-ref RIP in our cohort (Table 2) and was shown to in part regulate the expression of the gene (Goubert et al., 2020).

Although our N number was not sufficient, we hypothesise that germline non-ref RIPs either could play a role in the heritability of ALS through rare or *de novo* insertions that generate a phenotype or smaller effects of common variants (Savage et al., 2019). This is consistent with our analysis of reference genome SVA RIPs in Parkinson's disease using data in the longitudinal study Parkinson's Progressive Markers Initiative in which we correlated presence or absence of a specific RIP with Parkinson's disease progression (Pfaff et al., 2021). Non-ref RIPs that occur at higher frequencies could contribute to ALS susceptibility and be used to identify novel loci or occur in regions that are currently known to be involved in the disease. SNPs identified through GWAS do not necessarily identify the causative variant, but highlight regions of the genome containing variants linked to disease. For example, an *Alu* insertion in the *CD58* gene, that altered splicing and was found to be in perfect LD with the SNP associated with multiple cclerosis at this locus, has been suggested as the functional causative variant (Payer et al., 2017; Payer et al., 2019). We therefore tested the LD of two non-ref RIPs at the ALS GWAS loci, *UNC13A* and *SCFD1*, with the respective SNPs. These two RIPs variants were not found to be in LD with the risk alleles of the SNPs, but these RIPs could have an effect on risk independent of the GWAS loci. Rare retrotransposon insertions that compromise gene function, for example by exon disruption, cause a range of genetic diseases (Hancks and Kazazian, 2016) and of the insertions in the gnomAD-SV browser (https://gnomad.broadinstitute.org/), 0.21 % of *Alu*, 0.25 % of L1 and 0.36 % of SVA non-ref RIPs are loss of function variants (*Alu*, L1 and SVA insertions were extracted from VCFs downloaded from the gnomAD-SV browser and the percentage of those classified as loss of function variants calculated). The majority of these variants are very rare with 89 % of *Alu*, 92 % of L1 and 92 % of SVA insertions which cause a loss of function, having an allele frequency of < 0.0005 as reported in the VCF file. Non-ref RIPs have not been addressed in ALS genomes previously, therefore it is currently unknown if there are rare insertions causal for the disease. In addition, ongoing retrotransposition of these elements is an important consideration in sporadic ALS, as *de novo* insertions unique to specific individuals could be part of their genetic risk.

It is clear from the literature that retrotransposons have the ability to alter gene regulation and that their polymorphism is correlated with differential gene expression (Savage et al., 2019; Pfaff et al., 2021; Frohlich et al., 2022; Petrozziello et al., 2020; Pfaff et al., 2020; Pozojevic et al., 2022; Price et al., 2021). Our attempts are not to define somatic variation that occurs over the lifetime of an individual with ALS but rather the uncharacterised hereditability. To that end, we plan to extend our analysis of retrotransposon variation into the larger ALS genomic WGS dataset within ProjectMinE that currently consists of 6500 cases and 2500 controls. This would provide sufficient power to detect a potential significant association ($p < 1.7x10^{-5}$ based on 3000 RIPs tested) at an allele frequency of 0.05 with a relative risk of 1.42. Such analysis will allow further stratification of the ALS genetic factors associated with risk for, or progression of ALS. Validation of this hypothesis and strategy is not only supported by data in this study but also previous reports demonstrating that the analysis of reference SVA variation in Parkinson's disease can be correlated with disease progression (Pfaff et al., 2021) and that an increased burden of retrotransposition competent L1s is associated with Parkinson's disease (Pfaff et al., 2020). Interestingly, the bulk of these retrotransposon insertions are located in the non-coding DNA and most probably regulatory in nature, which would also support the hypothesis that the genetic risk of an individual for ALS is in part modified by the environment/life events as demonstrated in many conditions (Savage et al., 2019; Gianfrancesco et al., 2019; Marshall et al., 2021; Quinn et al., 2019).

4. Materials and methods

4.1. Identification of RIPs in whole genome sequencing data of 30 individuals from Project MinE

The samples from Project MinE used in this study were sequenced to ~ 25X coverage with 150 bp reads on the HiSeq X respectively. Table S2 summarises the phenotype data for 30 individuals used in the analysis. The input bam files for TEBreak and MELT required alignment using Burrow Wheeler Aligner (BWA-MEM) (Li, 2013), therefore the bam files for 30 individuals (15 controls and 15 ALS) from Project MinE were sorted by name and converted to fastq files using SAMtools (Li et al., 2009). The fastq files were aligned to GRCh37/hg19 using BWA-MEM with parameters –M –Y, converted to bam format and sorted using SAMtools and duplicates marked using Picard (https://broadinstitute.github.io/picard/). The subsequent bam file was used as input into TEBreak (installed 11/2016) and MELT-SINGLE (MELTv2.1.5) pipelines. Non-reference RIPs were identified and the insertions resolved using the tebreak.py and resolve.py scripts obtained by downloading the TEBreak folder from https://github.com/adamewing/tebreak. The list of RIPs generated using TEBreak in the output text file were filtered using the following parameters; a minimum of 4 split reads, a minimum of 4 discordant read pairs, a minimum element match of 0.90 and a minimum reference match of 0.95 (using the general_filter.py script). The insertions were then annotated to identify those that had been previously reported in the literature, this used a file compiled for the TEBreak program (nonref.collection.hg19.bed.gz) and the annotate.py script. The number of variant and reference reads are reported to distinguish heterozygotes from homozygotes (using genotyper.py script). The BWA aligned bam file was also used as input into MELT-SINGLE using the default parameters (for details on running MELT-SINGLE see MELT documentation https://melt.igs.umaryland.edu/manual.php) to generate a VCF file of the non-ref RIPs per individual analysed. MELT-SINGLE performs the pre-processing of the bam file and the four steps of the MELT-SPLIT pipeline for each family of non-ref RIPs (*Alu*, L1 and SVAs) on a single genome at a time. Analogous to TEBreak the RIPs were annotated for their location in relation to genes and if they had been previously identified. However, the priors file for MELT only included those non-ref RIPs identified in the 1000 Genomes Project in which MELT was used to characterise retrotransposon variation (Sudmant et al., 2015). The non-ref RIPs generated by MELT were filtered to remove those with two or fewer split reads and an assess score of less than three. The minimum number of split reads between the two tools differs because the number recommended specifically for each tool was used.

4.2. PCR validation of target RIPs and amplification in the MNDA UK DNA bank cohort

PCR assays were designed to validate presence/absence of 12 *Alu*, 11 L1 and 6 SVA insertions (randomly selected) by genotyping PCR. Primers were either placed in the genomic flank of the RIP to amplify the empty/filled site or an additional primer was placed in either the 5′ or 3′ end of the particular RIP to amplify the empty site and the 5′ or 3′ end of the element (see Table S4 for primer sequences). GoTaq Hot Start polymerase (Promega) was used under standard conditions to amplify the empty/filled sites for *Alu* and truncated L1 insertions and the 5′ and 3′ ends of the RIPs. KOD Hot Start polymerase (Novagen) under standard conditions with the addition of betaine (1 M final concentration) was used to amplify the empty/filled site of the SVA RIPs. DNA samples (1 ng input for PCR assays) from the MNDA UK DNA bank cohort (ref DNA0042) were used for validation of the RIPs identified in the TEBreak and MELT analyses as DNA was available from 18 individuals whose whole genomes were included in the analysis. Each non-ref RIP identified *in silico* was validated by PCR analysis in at least 9 of these DNA samples to incorporate a range of genotype calls from TEBreak and MELT. Each non-ref RIP was categorised as one of the following: true positive – presence of RIP called by TEBreak or MELT and confirmed present by PCR; true negative – absence of RIP called by TEBreak or MELT and confirmed absent by PCR; false positive – presence of RIP called by TEBreak or MELT but absent in PCR; false negative – absence of RIP called by TEBreak or MELT but present in PCR. False positive rate, false negative rate, sensitivity, specificity and accuracy for *Alu*, L1 and SVA insertions were calculated individually for TEBreak and MELT, and for the two tools combined with 95 % confidence intervals using the *epi.*tests function in the epiR package in R. Five of the non-ref RIPs identified in either an ALS-associated gene or haploblock were amplified in additional samples from the MNDA UK DNA bank cohort to determine the allele frequencies in controls and ALS samples. Association of these particular variants with ALS was tested using a Chi-squared test (PLINK v1.07) (Purcell et al., 2007).

4.3. Pathway analysis of genes containing RIPs which were identified in ALS and control genomes

Coordinates of genes from the UCSC genome browser were exported to Galaxy and intersected with the merged list of non-ref RIPs identified in the TEBreak and MELT analyses to identify genes containing non-ref RIPs. The gene list generated was analysed using DAVID (Huang da et al., 2009a,b) to identify those tissues, pathways and cellular components enriched for genes containing a non-ref RIP. The Bonferroni corrected p values are reported in Fig. 4 for the GO term cellular component direct and UP tissue categories.

4.4. Linkage disequilibrium analysis of non-reference RIPs with SNPs associated with ALS through GWAS

The genotypes of the SNPs in the genes *SCFD1* (rs10139154) and *UNC13A* (rs12608932), that were associated with ALS through genome wide association studies (GWAS), were obtained for the UK Project MinE samples that had been genotyped for the L1 insertion in *SCFD1* and the *Alu* insertion in *UNC13A* using PCR. PLINK v1.07 was then used to calculate LD between rs10139154 and the *SCFD1* L1 and rs12608932 and the *UNC13A Alu.*

4.5. Analysis of the distribution of RIPs in ALS associated loci in ALS and controls genomes

A list of ALS associated genes was compiled from two reviews (White and Sreedharan, 2016; Zufiria et al., 2016) and the chromosomal start and stop coordinates (GRCh37/hg19) were obtained from the UCSC genome browser (https://genome.ucsc.edu/). The coordinates of haploblocks associated with ALS were generated by intersecting the chromosomal loci of trait associated SNPs from the GWAS catalogue track on UCSC genome browser that reached genome wide significance ($p < 5x10^{-8}$) for association with ALS risk with haploblock loci for the European population (Berisa and Pickrell, 2016). The coordinates of non-ref RIPs were intersected with those of the ALS-associated genes and haploblocks**.**

Author contributions

Conceptualization, JPQ, VJB and ALS; methodology, ALS, AI, AR, GGS, JLGP, VJB and JPQ; software, ALS, AI and AR; validation, ALS; data interpretation, ALS, AI, GGS, SK, VJB, AAC and JPQ; data curation, ALS, AAK and AI; writing—original draft preparation, ALS; writing—review and editing, AI, GGS, JLGP, SK, VJB, AAC and JPQ; funding acquisition, JPQ, VJB, GGS and AAC. All authors have read and agreed to the published version of the manuscript.

Funding

This work was funded by the MNDA, Grant No: 41/523. ALS, AI, JPQ, GGS, VJB and AAC are funded by the Motor Neurone Disease Association and JPQ by the Andrzej Wlodarski Memorial Research Fund. This is in part an EU Joint Programme - Neurodegenerative Disease Research (JPND) project. The project is supported through the following funding organisations under the aegis of JPND - https://www.jpnd.eu *(United Kingdom, Medical Research Council* (MR/L501529/1; MR/R024804/1) *and Economic and Social Research Council* (ES/L008238/1)*).* This study represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. AI is funded by National Institute for Health Research (NIHR) Biomedical Research Centres at South London. JLGP acknowledges funding from MINECO-FEDER [grant number SAF2017-89745-R], the European Research Council [ERC-STG-2012–309433], the Howard Hughes Medical Institute (IECS-55007420) and a private donation from Ms Francisca Serrano (Trading y Bolsa para Torpes, Granada, Spain).

Ethical approval for use of DNA samples from the UK MND DNA Bank: Trent Research Ethics Committee 08/H0405/60.

7. Data availability statement

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

CRediT authorship contribution statement

Abigail L. Savage: Conceptualization, Methodology, Software, Validation, Writing – original draft. **Alfredo Iacoangeli:** Methodology, Software, Writing – review & editing. **Gerald G. Schumann:** Methodology, Writing – review & editing, Funding acquisition. **Alejandro Rubio-Roldan:** Methodology, Software. **Jose L. Garcia-Perez:** Methodology, Writing – review & editing. **Ahmad Al Khleifat:** Writing – review & editing. **Sulev Koks:** Writing – review & editing. **Vivien J. Bubb:** Conceptualization, Funding acquisition. **Ammar Al-Chalabi:** Writing – review & editing, Funding acquisition. **John P. Quinn:** Conceptualization, Methodology, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors acknowledge the Project MinE consortium for access to the whole genome sequencing data and the use of the research computing facility at King's College London, Rosalind (https://rosalind.kcl.ac.uk), which is delivered in partnership with the National Institute for Health Research (NIHR) Biomedical Research Centres at South London & Maudsley and Guy's & St. Thomas' NHS Foundation Trusts, and part-funded by capital equipment grants from the Maudsley Charity (award 980) and Guy's & St. Thomas' Charity (TR130505). AAC is an NIHR Senior Investigator. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, King's College London, or the Department of Health and Social Care. Sequence data and samples used in this research were in part obtained from the UK National DNA Bank and UK MND DNA bank for MND Research, funded by the MND Association and the Wellcome Trust. We would like to thank people with MND and their families for their participation in this project. Samples were provided by the Centre for Integrated Genomic Medical Research on approval by the UK MND Collection and we acknowledge sample management undertaken by Biobanking Solutions funded by the Medical Research Council at the Centre for Integrated Genomic Medical Research, University of Manchester. The authors would also like to acknowledge the use of the Chadwick HPC server at the University of Liverpool.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gene.2022.146799.

References

Al-Chalabi, A., Hardiman, O., 2013. The epidemiology of ALS: a conspiracy of genes, environment and time. Nat Rev Neurol. 9 (11), 617–628.

Al-Chalabi, A., Lewis, C.M., 2011. Modelling the effects of penetrance and family size on rates of sporadic and familial disease. Hum Hered. 71 (4), 281–288.

Al-Chalabi, A., Fang, F., Hanby, M.F., Leigh, P.N., Shaw, C.E., Ye, W., et al., 2010. An estimate of amyotrophic lateral sclerosis heritability using twin data. J Neurol Neurosurg Psychiatry. 81 (12), 1324–1326.

Batzer, M.A., Deininger, P.L., 2002. Alu repeats and human genomic diversity. Nat Rev Genet. 3 (5), 370–379.

Bennett, E.A., Keller, H., Mills, R.E., Schmidt, S., Moran, J.V., Weichenrieder, O., et al., 2008. Active Alu retrotransposons in the human genome. Genome Res. 18 (12), 1875–1883.

Berisa, T., Pickrell, J.K., 2016. Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics. 32 (2), 283–285.

Bragg, D.C., Mangkalaphiban, K., Vaine, C.A., Kulkarni, N.J., Shin, D., Yadav, R., et al., 2017. Disease onset in X-linked dystonia-parkinsonism correlates with expansion of a hexameric repeat within an SVA retrotransposon in TAF1. Proc Natl Acad Sci U S A. 114 (51), E11020–E11028.

Brown Jr., R.H., Al-Chalabi, A., 2017. Amyotrophic Lateral Sclerosis. N Engl J Med. 377 (16), 1602.

Chaisson, M.J., Huddleston, J., Dennis, M.Y., Sudmant, P.H., Malig, M., Hormozdiari, F., et al., 2015. Resolving the complexity of the human genome using single-molecule sequencing. Nature. 517 (7536), 608–611.

Chuong, E.B., Rumi, M.A., Soares, M.J., Baker, J.C., 2013. Endogenous retroviruses function as species-specific enhancer elements in the placenta. Nat Genet. 45 (3), 325–329.

Chuong, E.B., Elde, N.C., Feschotte, C., 2016. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science. 351 (6277), 1083–1087.

Chuong, E.B., Elde, N.C., Feschotte, C., 2017. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 18 (2), 71–86.

Conley, A.B., Jordan, I.K., 2012. Cell type-specific termination of transcription by transposable element sequences. Mob DNA. 3 (1), 15.

Conley, A.B., Piriyapongsa, J., Jordan, I.K., 2008. Retroviral promoters in the human genome. Bioinformatics. 24 (14), 1563–1567.

Douville, R., Liu, J., Rothstein, J., Nath, A., 2011. Identification of active loci of a human endogenous retrovirus in neurons of patients with amyotrophic lateral sclerosis. Ann Neurol. 69 (1), 141–151.

Elbarbary, R.A., Lucas, B.A., Maquat, L.E., 2016. Retrotransposons as regulators of gene expression. Science.

Ewing, A.D., 2015. Transposable element detection from whole genome sequence data. Mob DNA. 6, 24.

Faulkner, G.J., Billon, V., 2018. L1 retrotransposition in the soma: a field jumping ahead. Mob DNA. 9, 22.

Faulkner, G.J., Kimura, Y., Daub, C.O., Wani, S., Plessy, C., Irvine, K.M., et al., 2009. The regulated retrotransposon transcriptome of mammalian cells. Nat Genet. 41 (5), 563–571.

Feusier, J., Watkins, W.S., Thomas, J., Farrell, A., Witherspoon, D.J., Baird, L., et al., 2019. Pedigree-based estimation of human mobile element retrotransposition rates. Genome Res. 29 (10), 1567–1577.

Frohlich, A., Pfaff, A.L., Bubb, V.J., Koks, S., Quinn, J.P., 2022. Characterisation of the Function of a SINE-VNTR-Alu Retrotransposon to Modulate Isoform Expression at the MAPT Locus. Front Mol Neurosci. 15, 815695.

Gardner, E.J., Lam, V.K., Harris, D.N., Chuang, N.T., Scott, E.C., Pittard, W.S., et al., 2017. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res. 27 (11), 1916–1929.

Genomes Project C, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., et al., 2015. A global reference for human genetic variation. Nature. 526 (7571), 68–74.

Gianfrancesco, O., Bubb, V.J., Quinn, J.P., 2019. Treating the "E" in "G x E": Trauma-Informed Approaches and Psychological Therapy Interventions in Psychosis. Front Psychiatry. 10, 9.

Goerner-Potvin, P., Bourque, G., 2018. Computational tools to unmask transposable elements. Nat Rev Genet. 19 (11), 688–704.

Goubert, C., Zevallos, N.A., Feschotte, C., 2020. Contribution of unfixed transposable element insertions to human regulatory variation. Phil. Trans. R. Soc. B375.

Hanby, M.F., Scott, K.M., Scotton, W., Wijesekera, L., Mole, T., Ellis, C.E., et al., 2011. The risk to relatives of patients with sporadic amyotrophic lateral sclerosis. Brain. 134 (Pt 12), 3454–3457.

Hancks, D.C., Kazazian Jr., H.H., 2016. Roles for retrotransposon insertions in human disease. Mob DNA. 7, 9.

Huang da, W., Sherman, B.T., Lempicki, R.A., 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37 (1), 1–13.

Huang da, W., Sherman, B.T., Lempicki, R.A., 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 4 (1), 44–57.

Jacques, P.E., Jeyakani, J., Bourque, G., 2013. The majority of primate-specific regulatory sequences are derived from transposable elements. PLoS Genet. 9 (5), e1003504.

Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L., Bourque, G., et al., 2013. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet. 9 (4), e1003470.

Kazazian Jr., H.H., Moran, J.V., 2017. Mobile DNA in Health and Disease. N Engl J Med. 377 (4), 361–370.

Konkel, M.K., Walker, J.A., Hotard, A.B., Ranck, M.C., Fontenot, C.C., Storer, J., et al., 2015. Sequence Analysis and Characterization of Active Human Alu Subfamilies Based on the 1000 Genomes Pilot Project. Genome Biol Evol. 7 (9), 2608–2622.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., et al., 2001. Initial sequencing and analysis of the human genome. Nature. 409 (6822), 860–921.

Lauc, G., Huffman, J.E., Pucic, M., Zgaga, L., Adamczyk, B., Muzinic, A., et al., 2013. Loci associated with N-glycosylation of human immunoglobulin G show pleiotropy with autoimmune diseases and haematological cancers. PLoS Genet. 9 (1), e1003225.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25 (16), 2078–2079.

Li, W., Jin, Y., Prazak, L., Hammell, M., Dubnau, J., 2012. Transposable elements in TDP-43-mediated neurodegenerative disorders. PLoS One. 7 (9), e44099.

Li, W., Lee, M.H., Henderson, L., Tyagi, R., Bachani, M., Steiner, J., et al., 2015;7(307): 307ra153.. Human endogenous retrovirus-K contributes to motor neuron disease. Sci Transl Med.

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM2013.

Marino-Ramirez, L., Lewis, K.C., Landsman, D., Jordan, I.K., 2005. Transposable elements donate lineage-specific regulatory sequences to host genomes. Cytogenet Genome Res. 110 (1–4), 333–341.

Marshall, J.N., Lopez, A.I., Pfaff, A.L., Koks, S., Quinn, J.P., Bubb, V.J., 2021. Variable number tandem repeats - Their emerging role in sickness and health. Exp Biol Med (Maywood). 246 (12), 1368–1376.

Mathis, S., Goizet, C., Soulages, A., Vallat, J.M., Masson, G.L., 2019. Genetics of amyotrophic lateral sclerosis: A review. J Neurol Sci. 399, 217–226.

Mayer, J., Harz, C., Sanchez, L., Pereira, G.C., Maldener, E., Heras, S.R., et al., 2018. Transcriptional profiling of HERV-K(HML-2) in amyotrophic lateral sclerosis and potential implications for expression of HML-2 proteins. Mol Neurodegener. 13 (1), 39.

Mejzini, R., Flynn, L.L., Pitout, I.L., Fletcher, S., Wilton, S.D., Akkari, P.A., 2019. ALS Genetics, Mechanisms, and Therapeutics: Where Are We Now? Front Neurosci. 13, 1310.

Mills, R.E., Bennett, E.A., Iskow, R.C., Devine, S.E., 2007. Which transposable elements are active in the human genome? Trends Genet. 23 (4), 183–191.

Notwell, J.H., Chung, T., Heavner, W., Bejerano, G., 2015. A family of transposable elements co-opted into developmental enhancers in the mouse neocortex. Nat Commun. 6, 6644.

Payer, L.M., Burns, K.H., 2019. Transposable elements in human genetic disease. Nat Rev Genet. 20 (12), 760–772.

Payer, L.M., Steranka, J.P., Yang, W.R., Kryatova, M., Medabalimi, S., Ardeljan, D., et al., 2017. Structural variants caused by Alu insertions are associated with risks for many human diseases. Proc Natl Acad Sci U S A. 114 (20), E3984–E3992.

Payer, L.M., Steranka, J.P., Ardeljan, D., Walker, J., Fitzgerald, K.C., Calabresi, P.A., et al., 2019. Alu insertion variants alter mRNA splicing. Nucleic Acids Res. 47 (1), 421–431.

Pereira, G.C., Sanchez, L., Schaughency, P.M., Rubio-Roldan, A., Choi, J.A., Planet, E., et al., 2018. Properties of LINE-1 proteins and repeat element expression in the context of amyotrophic lateral sclerosis. Mob DNA. 9, 35.

Petrozziello, T., Dios, A.M., Mueller, K.A., Vaine, C.A., Hendriks, W.T., Glajch, K.E., et al., 2020. SVA insertion in X-linked Dystonia Parkinsonism alters histone H3 acetylation associated with TAF1 gene. PLoS One. 15 (12), e0243655.

Pfaff, A.L., Bubb, V.J., Quinn, J.P., Koks, S., 2020. An Increased Burden of Highly Active Retrotransposition Competent L1s Is Associated with Parkinson's Disease Risk and Progression in the PPMI Cohort. Int J Mol Sci. 21 (18).

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 Parkinsons Dis. 7 (1), 44.

Piriyapongsa, J., Marino-Ramirez, L., Jordan, I.K., 2007. Origin and evolution of human microRNAs from transposable elements. Genetics. 176 (2), 1323–1337.

Pozojevic, J., Algodon, S.M., Cruz, J.N., Trinh, J., Bruggemann, N., Lass, J., et al., 2022. Transcriptional Alterations in X-Linked Dystonia-Parkinsonism Caused by the SVA Retrotransposon. Int J Mol Sci. 23 (4).

Price, E., Gianfrancesco, O., Harrison, P.T., Frank, B., Bubb, V.J., Quinn, J.P., 2021. CRISPR Deletion of a SVA Retrotransposon Demonstrates Function as a cis-Regulatory Element at the TRPV1/TRPV3 Intergenic Region. Int J Mol Sci. 22 (4).

Project Min EALSSC, 2018. Project MinE: study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. Eur J Hum Genet. 26 (10), 1537–1546.

Prudencio, M., Gonzales, P.K., Cook, C.N., Gendron, T.F., Daughrity, L.M., Song, Y., et al., 2017. Repetitive element transcripts are elevated in the brain of C9orf72 ALS/FTLD patients. Hum Mol Genet. 26 (17), 3421–3431.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., et al., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 81 (3), 559–575.

Quinn, J.P., Savage, A.L., Bubb, V.J., 2019. Non-coding genetic variation shaping mental health. Curr Opin Psychol. 27, 18–24.

Rebollo, R., Romanish, M.T., Mager, D.L., 2012. Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet. 46, 21–42.

Renton, A.E., Chio, A., Traynor, B.J., 2014. State of play in amyotrophic lateral sclerosis genetics. Nat Neurosci. 17 (1), 17–23.

Saleh, A., Macia, A., Muotri, A.R., 2019. Transposable Elements, Inflammation, and Neurological Disease. Front Neurol. 10, 894.

Salvador-Palomeque, C., Sanchez-Luque, F.J., Fortuna, P.R.J., Ewing, A.D., Wolvetang, E.J., Richardson, S.R., et al., 2019. Dynamic Methylation of an L1 Transduction Family during Reprogramming and Neurodifferentiation. Mol Cell Biol. 39 (7).

Savage, A.L., Schumann, G.G., Breen, G., Bubb, V.J., Al-Chalabi, A., Quinn, J.P., 2019. Retrotransposons in the development and progression of amyotrophic lateral sclerosis. J Neurol Neurosurg Psychiatry. 90 (3), 284–293.

Savage, A.L., Lopez, A.I., Iacoangeli, A., Bubb, V.J., Smith, B., Troakes, C., et al., 2020. Frequency and methylation status of selected retrotransposition competent L1 loci in amyotrophic lateral sclerosis. Mol Brain. 13 (1), 154.

Schauer, S.N., Carreira, P.E., Shukla, R., Gerhardt, D.J., Gerdes, P., Sanchez-Luque, F.J., et al., 2018. L1 retrotransposition is a common feature of mammalian hepatocarcinogenesis. Genome Res. 28 (5), 639–653.

Schmidt, D., Schwalie, P.C., Wilson, M.D., Ballester, B., Goncalves, A., Kutter, C., et al., 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell. 148 (1–2), 335–348.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., et al., 2015. An integrated map of structural variation in 2,504 human genomes. Nature. 526 (7571), 75–81.

Sundaram, V., Cheng, Y., Ma, Z., Li, D., Xing, X., Edge, P., et al., 2014. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res. 24 (12), 1963–1976.

Tam, O.H., Ostrow, L.W., Gale, H.M., 2019a. Diseases of the nERVous system: retrotransposon activity in neurodegenerative disease. Mob DNA. 10, 32.

Tam, O.H., Rozhkov, N.V., Shaw, R., Kim, D., Hubbard, I., Fennessey, S., et al., 2019b. Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation. Oxidative Stress, and Activated Glia. Cell Rep. 29 (5), 1164–77 e5.

van Es, M.A., Hardiman, O., Chio, A., Al-Chalabi, A., Pasterkamp, R.J., Veldink, J.H., et al., 2017. Amyotrophic lateral sclerosis. Lancet. 390 (10107), 2084–2098.

Volkman, H.E., Stetson, D.B., 2014. The enemy within: endogenous retroelements and autoimmune disease. Nat Immunol. 15 (5), 415–422.

Wang, L., Jordan, I.K., 2018. Transposable element activity, genome regulation and human health. Curr Opin Genet Dev. 49, 25–33.

Wang, L., Rishishwar, L., Marino-Ramirez, L., Jordan, I.K., 2017. Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements. Nucleic Acids Res. 45 (5), 2318–2328.

Wang, L., Norris, E.T., Jordan, I.K., 2017. Human Retrotransposon Insertion Polymorphisms Are Associated with Health and Disease via Gene Regulatory Phenotypes. Front Microbiol. 8, 1418.

Wang, J., Song, L., Grover, D., Azrak, S., Batzer, M.A., Liang, P., 2006. dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. Hum Mutat. 27 (4), 323–329.

Wang, J., Vicente-Garcia, C., Seruggia, D., Molto, E., Fernandez-Minan, A., Neto, A., et al., 2015. MIR retrotransposon sequences provide insulators to the human genome. Proc Natl Acad Sci U S A. 112 (32), E4428–E4437.

Wang, H., Xing, J., Grover, D., Hedges, D.J., Han, K., Walker, J.A., et al., 2005. SVA elements: a hominid-specific retroposon family. J Mol Biol. 354 (4), 994–1007.

Weber, M.J., 2006. Mammalian small nucleolar RNAs are mobile genetic elements. PLoS Genet. 2 (12), e205.

White, M.A., Sreedharan, J., 2016. Amyotrophic lateral sclerosis: recent genetic highlights. Curr Opin Neurol. 29 (5), 557–564.

Zufiria, M., Gil-Bea, F.J., Fernandez-Torron, R., Poza, J.J., Munoz-Blanco, J.L., Rojas-Garcia, R., et al., 2016. ALS: A bucket of genes, environment, metabolism and unknown ingredients. Prog Neurobiol. 142, 104–129.