Investigation of Endogenous Retroviruses in the Pathogenesis of Sporadic Amyotrophic Lateral Sclerosis (ALS) by

Day, Edmund

https://doi.org/10.34737/w121w

Investigation of Endogenous Retroviruses in the Pathogenesis of Sporadic Amyotrophic Lateral Sclerosis (ALS)

by

Edmund Frederick Day

A thesis submitted in partial fulfilment for the requirements of the award of Doctor of Philosophy by the University of Westminster

August 2022

Abstract

Human Endogenous Retroviruses (HERVs) are remnants of ancient retroviral infections that have become incorporated into the human genome over the course of our evolution as a species. These fossil viruses have been co-opted by our genome as regulators of cellular gene expression amongst various other functions. Over the past few years however they have been increasingly discovered as differentially expressed in neurological conditions such as Amyotrophic Lateral Sclerosis.

While HERV-K (HML-2) transcripts were reported as elevated in premotor cortex samples from an ALS American cohort by Li *et.al.* (2015) the RT-qPCR assays performed in this sudy, using the same primers and reaction conditions in a larger ALS UK cohort were unable to corroborate these findings. Our collaborators at Kings College London, however, were able to find a novel HERV-K3 (HML-6) transcript on locus 3p21.31c upregulated in the primary motor cortex. We were able to confirm this using a subset of their primary motor cortex samples using RT-qPCR primers but were unable to replicate the results in a larger premotor cortex cohort. Using the modified ERVMap RNA-Seq method used by Jones *et.al.* (2021) we were able to analyse a number of publicly available datasets covering Cerebellum, Frontal Cortex, Motor Cortex and Peripheral Blood Mononuclear Cells. Within these datasets a number of novel HERVs were identified as being differentially expressed across the tissue types. A single HERV-H transcript was seen to be significantly downregulated in both the frontal cortex by RNA-Seq analysis and in the premotor cortex of our ALS UK cohort.

This identification of a novel HERV-H transcript being differentially regulated in the premotor cortex and subsequent RNA-Seq analysis on the blood and the brain provides a solid basis for future research into HERVs as novel biomarkers for ALS in which a diagnostic marker for early diagnosis and target for treatment is lacking to-date.

Table of Contents

List of Figures

List of Tables

Acknowledgements

The completion of this study could not have been completed without the expertise of my Director of Studies Dr. Adele L. McCormick, without their patience and support this thesis would not exist.

I would also like to thank Ashley R. Jones, who taught me how to perform the RNA-Seq workflow and DESeq analysis on RStudio so I could perform my own analysis of publicly available RNA-Seq data. A special mention goes to his college Dr. Renata Kabijilo who help clarify coding terminology.

Finally, I would like to thank Dr. Jeremy A Garson, whose expert advice on RT-qPCR assays was invaluable.

Authors Declaration

I declare that the present work was carried out in accordance with the Guidelines and Regulations of the University of Westminster. The work is original except where indicated by reference in the text.

The submission as a whole or part is not substantially the same as any that I previously or am currently making, whether in published or unpublished form, for a degree, diploma or similar qualification at any university of similar institution.

Until the outcome of the current application to the University of Westminster is known the work will not be submitted for any such qualification at another university or similar location.

Any views expressed in this work are those of the author and in no way represent those of the University of Westminster.

List of Abbreviations

Abbreviation	Definition
HERV	Human Endogenous Retrovirus
ERV	Endogenous Retrovirus
RTE	Retro Transposable Elements
BLAST	Basic Local Alignment Search Tool
DAVID	Database for Annotation, Visualization and Integrated Discovery
SMART	Simple Module Architecture Research Tool
RT-qPCR/qPCR	Real Time Quantitative Polymerase Chain Reaction
PCR	Polymerase Chain Reaction
TRIM28	Tripartite Motif Containing 28
ORF	Open Reading Frame
RNA	Ribonucleic Acid
DNA	Deoxyribonucleic Acid
cDNA	Complementary DNA
gDNA	Genomic DNA
mRNA	Messenger RNA
rRNA	Ribosomal RNA
miRNA	Micro RNA
MS	Multiple Sclerosis
ALS	Amyotrophic Lateral Sclerosis
sALS	Sporadic ALS
MND	Motor Neuron Disease
ASD	Autism Spectrum Disorder
TDP43/TARDBP	Tar DNA Binding Protein-43
BCL11b	B-cell lymphoma 11b
LTR	Long Terminal Repeat
bp	Base Pair
Mb	Mega Base Pair, 1,000,000 base pairs
Kb	Kilo Base Pair, 1,000 base pairs
UTR	Untranslated Region
gag	Group Antigen
pol	Polymerase
env	Envelope
RT	Reverse Transcriptase
LINE-1	Long Interspersed Element-1
HTLV	Human T-lymphotropic Virus
HSV	Herpes Simplex Virus
HHV	Human Herpesvirus
HIV	Human Immunodeficiency Virus
SIV	Simian Immunodeficiency Virus
Ig	Immunoglobulin
APOBEC3	Apolipoprotein B Editing Complex

Abbreviation	Definition
CD8	Cluster of Differentiation 8
HARRT	Highly Active Antiretroviral Therapy
ART	Antiretroviral Therapy
MRI	Magnetic Resonance Imagery
FA	Fractional Anisotrophy
CNS	Central Nervous System
XMRV	Xenotropic Murine Leukemia Virus-Related Virus
HML	Human Mouse Mammary Tumor Virus-Like
TLR	Toll Like Receptor
TNF	Tumour Necrosis Factor
INF	Interferon
PBMC	Peripheral Blood Mononuclear Cell
ATP	Adenosine Triphosphate
SLE	Systemic Lupus Erythmatosis
Ct	Cycle Threshold
GOI	Gene of Interest
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
XPNPEP1	X-Prolyl Aminopeptidase 1
MIQE	Minimum Information for Publication of Quantitative Real-Time PCR Experiments
NGS	Next Generation Sequencing
RNA-Seq	RNA Sequencing
LB	Luria Broth
MRC	Medical Research Council
SD	Standard Deviation
UV	Ultraviolet
NYGC	New York Genome Centre
UCSC	University of California, Santa Cruz
KCL	Kings College London
MEGA	MEGA Encrypted Global Access
MUSCLE	Multiple Sequence Comparison by Log-Expectation
NTC	No Template Control
CLUSTAL	Cluster Analysis of the Pairwise Alignments
NCBI	National Center for Biotechnology Information
R^2	Coefficient of Determination
P-Value	Probability Value
PMD	Postmortem Delay
RIN	RNA Integrity Number
chr	Chromosome
lfcSE	Log Fold Change Standard Error
PCA	Principal Component Analysis
RMSD	Root-Mean-Square Deviation
CSF	Cerebrospinal Fluid

1.0 Introduction

1.1 Introduction of HERVs

An often-quoted figure from the human genome project states that roughly 8% of the human genome is made up of retroviral elements (Lander *et al.*, 2001). Previously thought to be parasitic elements part of supposed "junk DNA" these have been theorised to either be the origin of retroviral agents or to be the remnants of ancient retroviral infections entering the germ line somewhere in the distant evolutionary past of our species (Coffin, Varmus and Hughes, 2002; Bannert and Kurth, 2006). These remnants have undergone many mutations since their incorporation into the genome to nullify their pathogenic potential and now serve many different functions relating to gene expression (Bannert and Kurth, 2006; Rebollo, Romanish and Mager, 2012; Buzdin, Prassolov and Garazha, 2017). Human Endogenous Retroviruses (or HERVs) are a relatively young addition to our genetic code, with the HERV-K family of retrotransposons originating from several infectious events roughly 33-40 million years ago (Leib-Mösch *et al.*, 1993; Vargiu *et al.*, 2016). In addition to these documented origins around 31 different families of HERVs have been identified (Antony *et al.*, 2011; Vargiu *et al.*, 2016; Morandi *et al.*, 2017). This relatively recent inclusion into our genome means that many HERV families, such as HERV-K, HERV-W and HERV-E families, still contain full length *gag, env* and *pol* genes within their sequences (Bannert and Kurth, 2006; Hohn, Hanke and Bannert, 2013). The presence of these genes means these HERV sequences retain the potential to form infective elements therefore still retain pathogenic ability. This potential is shown in their loose association with retrovirus subtypes, with HERV family associations being shown as beta-like, gamma-like etc based on their genetic similarity to wild type virus , as shown by a study comparing just over 3000 HERV *pol* sequences (Vargiu *et al.*, 2016; Gifford *et al.*, 2018). This study also confirmed HERV-K as being the youngest retroviral element in our genome, alongside HERV-FC (HERV family F, member C) (Vargiu *et al.*, 2016). The "K" in the HERV-K family designation originally referring to the presumed use of lysine tRNA in the mechanism of reverse transcription, though this has not been found in all family members (Hanke, Hohn and Bannert, 2016; Xue, Sechi and Kelvin, 2020). These HERV elements have various beneficial roles in gene expression throughout the body (Buzdin, Prassolov and Garazha, 2017). Their function as retrotransposons are defined as movable genetic elements that act as a copy-paste

mechanism for modification of gene expression. This can either have an inhibitory or promoting effect on the targeted gene and results in the copy number of each HERV varying depending on their type and intended function, with the HERV-K family having an approximate copy number of 25,000 across the genome (Sverdlov, 1998; Belshaw *et al.*, 2005; Thomas, Perron and Feschotte, 2018). This mechanism is driven by the enzyme reverse transcriptase which converts viral RNA to cDNA and facilitates the re-insertion into the host genome (Hohn, Hanke and Bannert, 2013). Reverse transcriptase is notoriously error prone, with nucleotide errors occurring 1 in every 2000 incorporations, or as low as 1 in 70 in certain HIV strains, meaning the chance of protein coding mutations in the transcribed sequence is also high (Preston, Poiesz and Loeb, 1988; Roberts, Bebenek and Kunkel, 1988; Sebastián-Martín, Barrioluengo and Menéndez-Arias, 2018).

HERVs have had a profound effect on our evolution as a species and have had their involvement categorised in various tissues and developmental processes. In foetal neuronal tissue HERV-K has been observed with higher levels of expression than other cell lines, indicating a role in development (Rebollo, Romanish and Mager, 2012; Mortelmans, Wang-Johanning and Johanning, 2016). This has a structural relation to a murine ERV (endogenous retrovirus) which interacts with TRIM28 to methylate histone structures near the ERV site during development (Mortelmans, Wang-Johanning and Johanning, 2016). There is also evidence to suggest that HERV-K may have a neuroprotective effect. . This comes from a study in which increased HERV-K *env* protein expression resulted in expression of neurotrophin nerve growth factor and brain-derived neurotrophic factor, both of which contribute to ongoing survival of neuronal cells (Bhat *et al.*, 2014; Mortelmans, Wang-Johanning and Johanning, 2016). In addition, HERV-W Syncytin-1 transcripts have also been found in higher levels in foetal tissue, indicating their potential involvement in development alongside HERV-K (Hohn, Hanke and Bannert, 2013; Mortelmans, Wang-Johanning and Johanning, 2016). Other HERV functions relating to gene expression are related to their interaction with Long Terminal Repeats (LTRs). These are found flanking HERV sequences and have roles in inhibition, promotion, and response signals to other genomic processes (Douville *et al.*, 2011). This goes partway to explain their variation in copy number and location when flanking certain genes.

1.2 Implication of HERV-K and HERV-W in neurological and non-neurological diseases

Both HERV-K and HERV-W families are able to form Virus-Like Particles (VLPs) due to fully intact ORFs and have been implicated in neurological and other non-neurological human diseases (Brodziak *et al.*, 2012; Monde *et al.*, 2012; Mameli *et al.*, 2013; Johanning *et al.*, 2017; Dolei *et al.*, 2019; Tam, Ostrow and Gale Hammell, 2019; Dembny *et al.*, 2020). The HERV-K family of retrotransposons have been identified in many cancers with full length HERV-K *env* transcripts being expressed in breast cancer cells (Grandi and Tramontano, 2018). HERV-K expression has also been found in other cancers with upregulation of RNA transcripts being found in leukaemia, teratocarcinoma cells, germ cell tumours and melanomas (Yi, Kim and Kim, 2006; Agoni, Guha and Lenz, 2013; Hohn, Hanke and Bannert, 2013). In neurodegenerative conditions, HERV-K has been found in elevated levels in Amyotrophic Lateral Sclerosis (ALS) and Multiple Sclerosis (MS) (Li *et al.*, 2015; Hera and Urcelay, 2016; Tam, Ostrow and Gale Hammell, 2019). The HERV-W family of retrotransposons has a more clearly defined role in MS as the Syncytin-1 transcripts have been found in elevated levels in brain tissue and contributing to the pathogenesis of the disease (Fujinami and Libbey, 1999; Antony *et al.*, 2011; Morandi, Tarlinton and Gran, 2015). Both families of HERVs have been found in schizophrenic brain tissue and have also been shown to be highly expressed in cases of severe depression with HERV-W showing some involvement alongside HERV-K (Suntsova *et al.*, 2013; Slokar and Hasler, 2016; Grandi and Tramontano, 2017; Küry *et al.*, 2018). HERV-K and HERV-W families have also been observed as differentially expressed in Autism Spectrum Disorder (ASD), the same study also noted that there was a negative correlation between HERV-H expression and age in ASD with significantly higher expression of the family member in those that had severe forms of the disorder (Balestrieri *et al.*, 2012). Psychiatric conditions have also been shown to have HERV family expression, with HERV-W and H in schizophrenia and bipolar disorder being differentially expressed across multiple brain regions (Li *et al.*, 2019).

HERV-K subfamilies, grouped as beta-like retroviruses by one classification system, function much like other retrotransposons in their effect in promoting or inhibiting functions of certain genes. They have been found to have strong correlations with protein expression in neuronal cells, such as with Tar DNA Binding Protein-43 (TDP 43, upregulated in ALS), and there have been many studies on their function in diseases, such as in cancer and

autoimmune disorders (Singh, 2007; Magiorkinis, Belshaw and Katzourakis, 2013; Li *et al.*, 2015). Their role in healthy tissue has not been well categorised but HERV Long Terminal Repeats (LTR) found upstream from transcribed proteins are thought to have beneficial actions on protein expression (Cohen, Lock and Mager, 2009). Their primary beneficial role is thought to be as drivers of human evolution, especially in the more active HERV-K subfamily as it has not undergone as many silencing mutations as other HERV families due to its more recent inclusion into the human genome (Sverdlov, 1998; Hughes and Coffin, 2004; Bannert and Kurth, 2006; Shin *et al.*, 2013). Evidence of this driving force of evolution comes from the ability of HERVs to insert themselves into new positions in the genome, upregulating transcription of sequences downstream of the insertion site creating new genes or non-coding RNA sequences (Suntsova *et al.*, 2015). HERVs are also thought to have a minor effect in individuals as age increases with HERV-K family members differentially expressed in young vs old comparisons (Nevalainen *et al.*, 2018) and thought to be as a result of down regulation of DNA methylation with increasing age.

1.3 Genomic organization of HERV-K & HERV-W families

Of the 31 HERV families HERV-K and HERV-W have been shown to be among the most biologically active. The HERV-K family contains around 160 proviral sequences, with 90 intact or partially intact full length sequences and around 2000 solo long terminal repeats (LTR's that are separated from the progenitor proviral sequences) (Buzdin *et al.*, 2003; Gray *et al.*, 2019). These HERV-K family members are spread across more than 1000 different loci within the human genome with most of them being solo-LTR sequences (Xue, Sechi and Kelvin, 2020). These solo-LTRs appear at a 10-fold increase in apparent abundance compared to more intact proviral sequences (Xue, Sechi and Kelvin, 2020). This family can also be divided into 2 subgroups based on their average age of being included into the human genome, derived from phylogenetic methods, with events at 5.8 and 10.3 million years ago (named HS-a and HS-b respectively) (Buzdin *et al.*, 2003; Katsura and Asai, 2019). Alternate methods of grouping HERV-K family members is by the presence or absence of a 292bp sequence in the *pol-env* region or by their LTR sequences (Hanke, Hohn and Bannert, 2016). One phylogenetic study of these HERV sequences showed 90% of these insertions had parallels to non-human primates while 10% were human specific sequences included in Introns (Buzdin *et al.*, 2003). Of these human specific sequences (HS) the most recent

inclusion into our genome (HS-a) was shown to be the most retrotranspositionally active (Buzdin *et al.*, 2003). As the HERV-K family is a very recent inclusion into our genome it's genes have not been fully silenced by mutations in its genetic code like some older families of HERVs (Buzdin *et al.*, 2003; Subramanian *et al.*, 2011; Katsura and Asai, 2019). This results in the ability to produce virus like particles, functioning enzymatic activity and immunologically active antigens from its *env* genes.

The genome of HERV-K sequences, while varying between members, is 9.5kb in length following the general sequence layout of 5'-UTR-*gag-pro-pol-env/rec*-UTR-3' with the *gag, pro* & *pol* genes appearing on separate reading frames with *env* & *rec* being produced from alternate splicing of mRNA (Fig. 1) (Agoni, Guha and Lenz, 2013; Xue, Sechi and Kelvin, 2020). The alternate splicing of the HERV-K reading frames relies on host post translational modification machinery, made available to the HERV by its location upstream of transcribed genes (Agoni, Guha and Lenz, 2013).

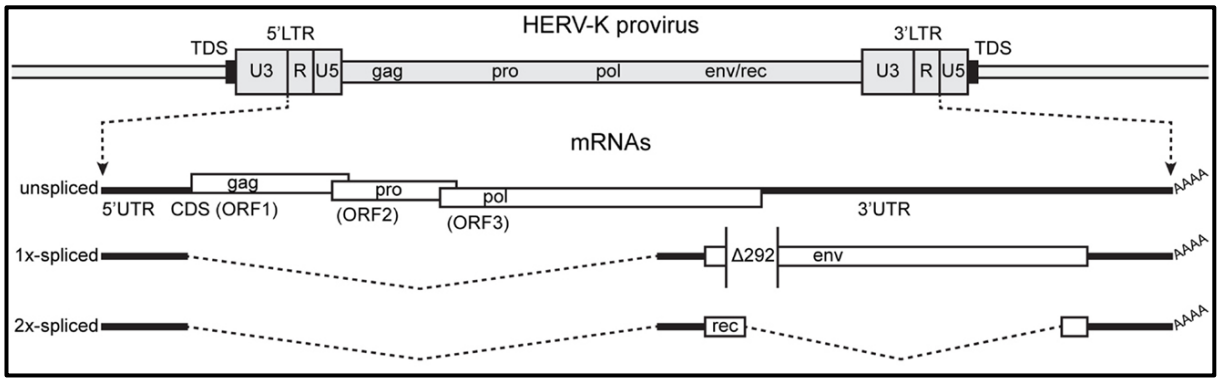


Figure 1.1 Structure of the HERV-K genome and spliced mRNAs. **(Agoni, Guha and Lenz, 2013)**.

The image above shows the genomic organisation of HERV-K provirus as it is situated inside the host's genome, also shown is the alternate splicing required to gain the *env* and *rec* mRNA sequences and the open reading frames which house the main proviral proteins.

Also reported in a study conducted by Agoni, Guha and Lenz (2013) there is a wide variability in length of the genome between members of the HERV-K family, with most of the sequences being in excess of 7,000bp in length (Agoni, Guha and Lenz, 2013). These HERV-K sequences do not seem to vary considerably in the population, with the main

identifications of polymorphisms occurring when solo-LTR's have been identified (Hughes and Coffin, 2004). These polymorphisms are likely due to their alternate inclusion into the genome, potentially recombining with other proviral sequences or being duplicated in an allelic manner at a separate site from the progenitor provirus and have been suggested to be a main driver of the evolutionary process (Hughes and Coffin, 2004; Bannert and Kurth, 2006; Vargiu *et al.*, 2016). The paper by Hughes and Coffin (2003) has also reported high relative formation of these solo-LTR's in the HERV-K family, occurring at an approximate rate of 0.002 per generation.

The genomic organisation of HERV-K as depicted in Figure 1, has a similar organisation to exogenous retroviruses they originated from, , with *pol* coding the reverse transcriptase (transcribes viral RNA to cDNA), RNaseH (breaks down RNA-DNA Hybrid to pure cDNA) and Integrase (facilitates viral genome entry into host genome) genes, *gag* coding for the viral core proteins and *env* coding for the viral envelope components. One of the more important genes when related to human health, within the viral genome is the *env* gene, coding for the viral envelope proteins. These *env* proteins have been identified in a number of conditions either acting as super-antigens, causing toxicity to the cells they are activated in or alternatively acting in a protective manner (Bhat *et al.*, 2014; Li *et al.*, 2015; Grandi and Tramontano, 2018). The *rec* protein, which is contained within the *env* sequence, performs a similar function to the HIV *rev* protein which regulates protein expression of the provirus; despite this similarity in function the *rec* and *rev* proteins contain no sequence homology (Ehlhardt *et al.*, 2006; Hanke *et al.*, 2013). While the function of *rec* has not been clearly defined its expression has been identified in synovial fluid and has been found to be elevated in certain cancers (Hanke *et al.*, 2013; Schmitt *et al.*, 2015). *Gag* functions in much the same way as in HIV, meaning its purpose lies in packaging and release of virions, which works in concert with the *pro* protease gene (Grandi and Tramontano, 2018). While these proteins do not have any use in healthy functioning of the cell, it has been shown that *gag* is still able to confer the ability to form virus like particles (Subramanian *et al.*, 2011). The last important gene from the HERV-K sequence is the *pol* gene, which encodes the viral polymerase genes, including reverse transcriptase which is needed for the synthesis of cDNA from viral RNA, so the virus can integrate into the host genome. The *pol*

gene has been used previously in phylogenetic studies to differentiate between HERV-K transcripts and to infer evolutionary relationships between the proviruses (Vargiu *et al.*, 2016).

The HERV-W family of endogenous retroviruses is, alongside HERV-K, widely researched for its effects in pathology and health. Its retroviral sequence is structured similarly to HERV-K with the differences down to the coding of individual gene families. One of the most important genes related to pathology in HERV-W related disease is the *env* protein Syncytin-1, which has been found to have an important effect in both neurological diseases and has shown to have an effect in placental tissue, in which there is elevated expression of this gene. (Oluwole *et al.*, 2007; Cohen, Lock and Mager, 2009; Grandi and Tramontano, 2017; Morandi *et al.*, 2017). HERV-W has also been shown to be the only family of ERV's that responds to L1 (Long Interspersed Element-1 or LINE-1) retrotransposons, with its interaction determined to be the primary force behind its integration into the primate genome (Grandi and Tramontano, 2017).

The frequency of the HERV-W family is on a comparable level to the inclusion of HERV-K when considering all of its elements in the genome, with 65 proviruses and 135 L1 (Long Interspersed Element-1 or LINE-1) mediated pseudogenes (Grandi and Tramontano, 2017). While this does show a lack of provirus inclusions when related to HERV-K the L1 related elements pick up the rest of the variation in the family (Grandi and Tramontano, 2017). The association with L1 in the genome is potentially the family's second important feature in pathogenesis. The facilitated retrotransposition of this interaction could lead to altered expression resulting in various conditions including autoimmune reactions against HERV-W *env* proteins and MS (Grandi and Tramontano, 2017).

HERV research has also investigated the activation of the endogenous element with infection by other viruses, of which HTLV also has some interaction with HERVs. One study into the link between HTLV infection and the expression of HERV families showed that *tax,* a viral gene product associated with viral and cellular processes, increased the expression

of HERV transcripts in Jurkat cells (Toufaily *et al.*, 2011). The primary families of HERVs activated in this way were identified as being HERV-W and HERV-H. In addition to *tax* Jurkat cells were exposed to several T-cell activating proteins which also resulted in increased HERV family transcription levels (Toufaily *et al.*, 2011). There has also been research showing the activation of HERVs concurrent with Herpesvirus infections. The research paper by Brudek et al, 2017, used lymphocyte cells taken from MS patients and infected cultures with HSV-1, HHV-6 and HHV-3 to see whether they could detect any increase in reverse transcriptase (RT) activity which is a retroviral marker (Brudek *et al.*, 2007). While all viral infections showed a detectable increase in RT activity, HHV-3 was the only variant which showed sustained expression over an extended period of time (Brudek *et al.*, 2007).

1.4 Similarities between HERVs and Exogenous Retroviruses

A link between retroviral infection and Neurodegenerative disease was first discussed in the 1970's with the identification of reverse transcriptase activity in post mortem brain tissue of those suffering from ALS, alongside a similarity in symptoms to other diseases of the nervous system, such as poliomyelitis (Viola *et al.*, 1975; Norris, 1977, obtained from NCBI archive). Neurodegenerative symptoms in retroviral disease have been observed more recently as well, with other human viruses like HIV and HTLV exhibiting some ALS like symptoms, which vary in severity depending on specific cases (Ando *et al.*, 2015; Bowen *et al.*, 2016). These symptoms shown by retroviral infections tend to respond well to antiretroviral therapy, with complete clearance of ALS like symptoms reported upon treatment with antiretrovirals, though longevity in these cases also varies with disease (Li *et al.*, 2015; Küry *et al.*, 2018). A clinical communication by Garcia-Montojo *et al.* (2021), published as part of an ongoing trial for the use of an antiretroviral drug in the treatment of ALS , showed that the amount HERV-K (HML-2) transcripts detected continuously fell over the 24 week administration of Triurimeq. These trials are also given mention in a review paper on the subject of using HERVs as therapeutic targets, citing the abundance of potential diseases which HERVs play a role in and the importance a potential antiretroviral therapy could have on those affected by HERV related conditions (Giménez-Orenga and Oltra, 2021).

Similarities between endogenous and exogenous retroviruses also lies within the innate and humoral immune response. There has been IgG response documented towards HERV-K in blood samples from pregnant mothers as well as patients undergoing treatment for various cancers, indicating that the immune system mounts an immune response to HERV-

K in both these conditions (Alfahad and Nath, 2013; Mortelmans, Wang-Johanning and Johanning, 2016; Gröger and Cynis, 2018). The IgG response in these cases gives evidence to the humoral immune response, similar in all infections, but the innate immune response to HERVs also contains some elements which are also used in combating exogenous retroviral infections. The method employed to combat exogenous viral infection with endogenous elements is to use HERV-K *env* proteins to out-compete exogenous viruses for cell surface entry receptors (Frank and Feschotte, 2017). The APOBEC3 protein utilised by the innate immune system works to repress endogenous retroviruses by a deaminase protein, introducing many mutations into its genetic code and preventing it from exercising its pathogenic potential (Lee, Malim and Bieniasz, 2008). Studies have shown inhibition of HERV-K expression in the presence of APOBEC3, with the hypermutation outcome being similar to that of a response to HIV (Lee, Malim and Bieniasz, 2008). However, it is worth pointing out however that only 2 of the 16 HERV-K proviruses in the study responded to the APOBEC3 mechanism, which the researchers determined to be due to viral tropism (Lee, Malim and Bieniasz, 2008).

Another study showed a similarity in immune response to endogenous and exogenous retroviruses in the specific T-Cell response to HERV-K. The study showed a cross reactivity between HERV-K specific CD8+ T-Cells, and a varied population of HIV/SIV virions (Jones *et al.*, 2012). These CD8+ cells targeting HERV-K were taken from HIV+ individuals, cloned and then tested against a diverse population of samples from both HIV Type 1 & 2, along with SIV samples (Jones *et al.*, 2012). The CD8+ cells proved to be effective in eliminating all immunodeficiency virus types *in vitro*, hinting at a possible role for HERV-K in antiretroviral defence (Jones *et al.*, 2012). This also goes part way in explaining why, in a separate study, HERV-K expression was shown to increase concurrent to HIV infection, to the point where pseudo-virions were detectable when tagged with a marker against the HERV *env* proteins (Bhardwaj *et al.*, 2014). HERV-K has also shown some other protective effects when the body is exposed to HIV infection with the HERV-K *env* proteins shown to have protective effects against the neurotoxicity of HIV-1 *Vpr* (Monde *et al.*, 2012; Bhat *et al.*, 2014). Other areas in which exogenous and endogenous viruses respond in a similar way is in the application of Highly Active Antiretroviral Therapy or HARRT. It has already been stated in

this review that there is a relationship between the application of ART and a decline in HIV related ALS symptoms, giving a credible link to a possible retroviral influence in ALS. This is also shown in a study where HERV-K expression was recorded as being significantly lower in those patients responding to HARRT (van der Kuyl, 2012). A different study showed that expression of HERV-K and HIV was increased in patients whose HARRT was proving to be unsuccessful (Hohn, Hanke and Bannert, 2013). The researchers of the study theorised that this was due to an interaction between the *rec* and *rev* response region (a nucleotide region present in viral mRNA which recruits transporter proteins and allows the viral mRNA to be transported to the host cell cytoplasm) in HERV-K and HIV (Hohn, Hanke and Bannert, 2013). This interaction between the endogenous and exogenous virus shows a possible link in the ALS like symptoms in some HIV infections and the involvement of HERV-K in the sporadic form of the disease. Another paper to identify a link between HIV and HERV-K observed that the expression of HERV-K provirus pseudo-virions was increased in HIV infected patients vs non-infected patients but qPCR data in the study showed that the expression was not always related to CD4+ cells (Bhardwaj *et al.*, 2014). The greatest difference in expression was in the monocyte fraction of cells measured, while this may not show a direct interaction between the two events it can be assumed that there is an indirect activation through the immune system.

1.5 Amyotrophic Lateral Sclerosis (ALS) Pathology of the Central Nervous System

The brain, along with the spinal cord, forms the central nervous system of the Human body and is responsible for controlling all of our physiological functions. The brainstem, cerebellum and cerebral hemispheres form the main sections of the brain with the cerebral hemisphere containing the cerebral cortex amongst other features such as the basal ganglia (Siegel and Sapru, 2019). The cerebral cortex makes up the outer layer of the brain, several millimetres thick and subdivided into the frontal, parietal and temporal lobes (Siegel and Sapru, 2019). This layer is made up of primarily cellular grey matter and white matter consisting mostly of myelinated axons. The grey matter area is further subdivided into 6 distinct layers (Figure 1.2 D), with the motor cortex lacking the cell packed granular layer of the primary sensory areas (James Knierim, 2000). Instead of the granular layer the most distinct layer of the motor cortex is the 5^{th} layer which contains pyramidal Betz cells whose projections span the preceding 4 layers of the cerebral cortex and axons which run

through the corticospinal tract (Figure 1.2 D) (James Knierim, 2000). Cells originating in the first 2 cortical layers connect to other areas of the cortex, the 3rd layer connecting to the opposite hemisphere of the brain and the 6th layer connecting to the thalamus (Siegel and Sapru, 2019). The frontal lobe is the largest of the 3 lobes of the cerebral cortex and contains the principle components for motor control of voluntary movements and formulation of the motor components of speech (Siegel and Sapru, 2019).

ALS or Motor Neurone Disease (MND) is a progressive neurodegenerative disease effecting the central nervous system, focused on the motor cortex resulting in a progressive decline in the body's ability to control voluntary muscles (Rowland and Shneider, 2001; Kiernan *et al.*, 2011). This disease is known to affect the elderly in the greatest proportion, with incidences increasing as the individual ages past 60 (Logroscino *et al.*, 2010). This disease is classified into two separate presentations in the disease, familial and sporadic, with sporadic making up 90% of reported cases in the US (Kiernan *et al.*, 2011). ALS is characterised by the loss of upper and lower motor neurons, resulting in progressive paralysis of muscles in the body, resulting in losses to speech, movement and respiration (Marini *et al.*, 2018).

The motor cortex is located towards the posterior portion of the frontal cortex consisting of the precentral gyrus which houses the primary motor cortex, responsible for the control of voluntary movements (Figure 1.2) (Siegel and Sapru, 2019). The neural cells within the primary motor cortex are arranged somatotopically, meaning different areas are associated with different parts of the body e.g. neurons controlling, the biceps and arm muscles are all located together (James Knierim, 2000; Siegel and Sapru, 2019). Immediately rostral of the primary motor cortex are the premotor and supplemental motor area (Figure 1.2). The supplemental motor areas primary function is to coordinate voluntary movements, eliciting more complex patterns of movement than the primary motor cortex (Siegel and Sapru, 2019). The premotor cortex, the area of interest in this study, works in tandem with the supplementary motor area in coordinating muscle movement. It sends axons to both the primary motor cortex and the corticospinal tract directly, controlling more complex postures than the primary motor cortex and those movements that are guided by sight (James Knierim, 2000; Siegel and Sapru, 2019). The premotor cortex also controls reactions

to internal and external stimuli along with memorised motions, including reaching for objects (Purves *et al.*, 2001).

In grading the progression of ALS in the brain the involvement of the premotor cortex region corresponds to stage 2 of neurodegeneration when measured by TAR DNA Binding Protein 43 (TDP-43) involvement, the first stage presenting as abnormalities in the betz cells in layer V of the primary motor cortex (Eisen *et al.*, 2017). As the disease progresses in the primary motor cortex magnetic resonance imaging (MRI) can detect areas of upper motor neuron degeneration and, due to the somatotopic layout of the primary motor cortex, show impairment of specific regions (such as hands and feet) which correlate with visible symptoms (Costagli *et al.*, 2016). This susceptibility mapping (looking at areas which increase in magnetic signal) is used to look at the dysregulation of iron in regions of the primary motor cortex for limb involvement (Costagli *et al.*, 2016; Bhattarai *et al.*, 2019). A recent paper found that increased susceptibility to MRI in the primary motor cortex was found in all limb-onset ALS patients, but not in those which had cervical spinal limb-onset variant of the disease.

During stage 2 lesions develop in the premotor areas marking the beginning of cerebellar dysfunction in the disease (Eisen *et al.*, 2017). Degeneration of neuronal cells in the motor cortex area can be measured by Fractional Anisotropy (FA), the measurement of directional movement of water molecules with isotropic movement (FA-0) correlating to cerebrospinal fluid and anisotropic movement (FA-1) relating to fibre bundles in the brain (Alba-Ferrara and de Erausquin, 2013). In a study looking at neurodegeneration in ALS it found that the degeneration of fibrous matter in the premotor cortex reduced the FA of the white matter below and was potentially related to degeneration in the posterior limb of the internal capsule and the disintegration of the white matter along the corticospinal tract (Zhang *et al.*, 2018). This degeneration has also been observed in the sporadic form of the disease, both in structure, seen in cortical thinning of the region (Li *et al.*, 2015) resulting in hypoactivation (Cosottini *et al.*, 2012), and function with decreased glucose metabolism (Marini *et al.*, 2018).

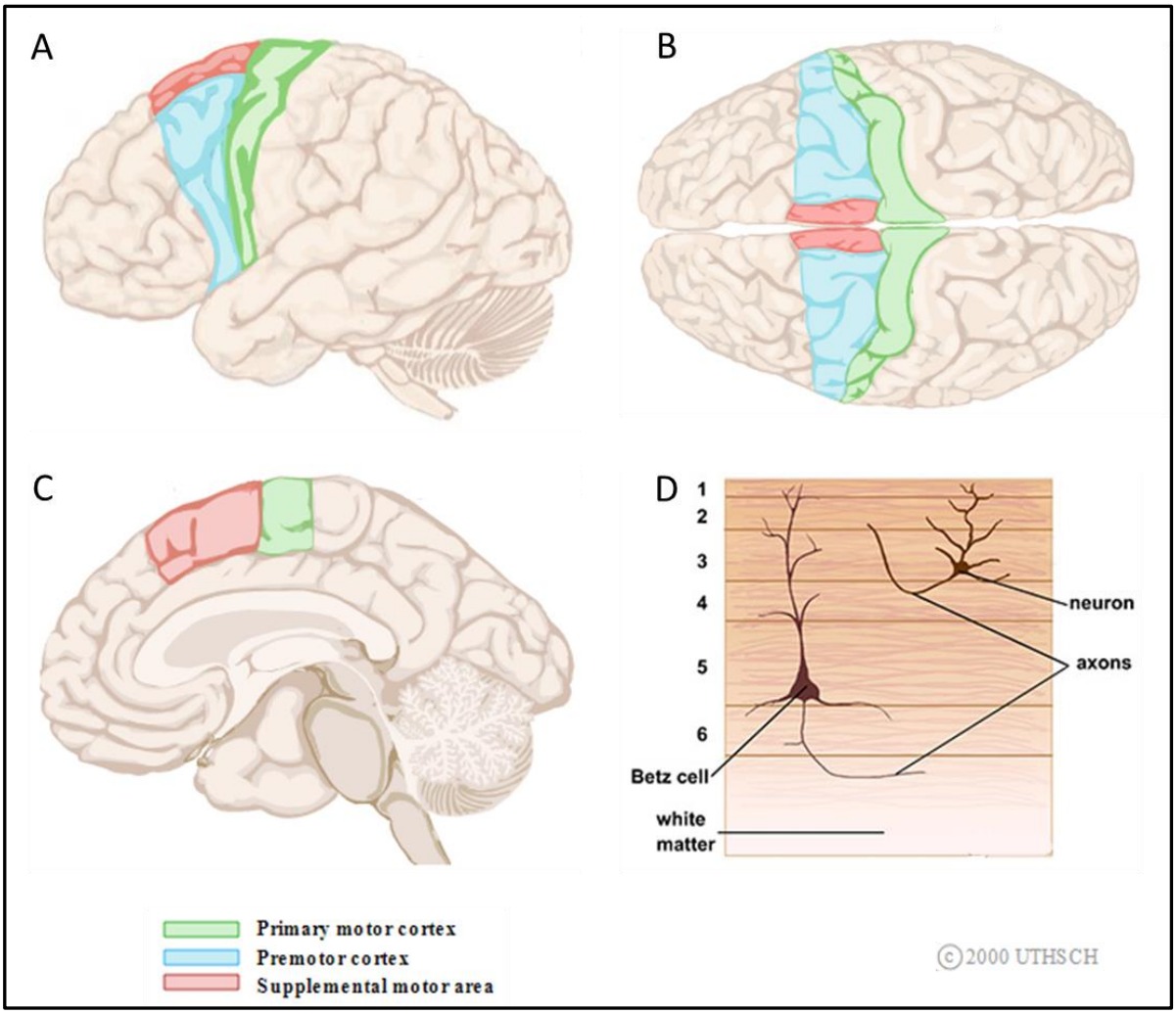


Figure 1.2. Motor Cortex Location and its Cytoarchetecture (James Knierim, 2000)
The figure above displays areas of the motor cortex within the human brain, showing the area from A) Lateral, B) Dorsal and C) Medial views. Also shown in D is the Cytoarchetecture of the motor cortex layer showing the 6 layers of the cerebral cortex and the location of the pyramidal Betz cells and non-pyramidal cells within these layers (James Knierim, 2000).

While the progressive loss of motor neurons is characteristic of ALS multiple regions of the brain and spinal cord are involved in the pathogenesis of the disease. The cerebellum is located at the posterior of the brain and has classically been linked to motor cortex actions, though it has been shown to have functions in other neurological areas (Schmahmann and Caplan, 2006; Fernández, Sierra-Arregui and Peñagarikano, 2019). Structurally the cerebellum consists of 10 lobules, generally numbered with roman numerals, separated into 2 separate lobes with each region (lobule) performing a separate function and lobules VI and VII related to motor control (Manto *et al.*, 2012; Fernández, Sierra-Arregui and Peñagarikano, 2019). Studies have shown that in both sporadic and familial ALS that neurodegeneration of the motor system extends beyond the motor cortex and into the

cerebellum (Borba *et al.*, 2019). Specifically, in sporadic ALS there has been a distinct observable loss in grey matter in lobule VI by MRI (Borba *et al.*, 2019). This has also been seen in another study which not only used MRI, but FA, and Functional Connectivity (FC) and showed decreased FC and changes in grey matter volume in the cerebellum in ALS patients (Qiu *et al.*, 2019).

Alongside the motor cortex, degeneration in the spinal cord is also a common feature amongst ALS cases. The spinal cord can be divided into 5 broad sections based on their location relative to the body, Cervical which is situated closest to the brain and runs down the neck, Thoracic running down the upper back, lumbar which covers the lower back and sacral which runs the lowest part of the spine to the tailbone (Nógrádi and Vrbová, 2013). Changes in the spinal cord can be observed in cross sections and image study of cervical and upper thoracic regions of the spinal cord, specifically in the atrophy of grey matter (El Mendili *et al.*, 2014; Paquin *et al.*, 2018). Cellular changes can be seen in the form of mitochondrial dysfunction localised to specific areas of the spinal cord; one study showed that there was a significant decrease in complex IV activity in the lumbar and cervical regions in post-mortem samples of ALS patients compared to controls (Delic *et al.*, 2018). Changes can also be seen in the protein level of gene expression with one study showing significant changes in 292 out of 6810 identified proteins in ALS, with proteins involved in mRNA splicing shown to be enriched (Oeckl *et al.*, 2020).

1.5.1 TAR DNA Binding Protein 43 (TDP-43) and B-cell lymphoma 11b (BCL11b) Involvement in ALS

The familial form of ALS has been extensively categorised following the discovery of the first ALS associated gene, Superoxide Dismutase 1 (SOD1, superoxide metabolism), and the creation of a mouse line expressing this gene (Ajroud-Driss and Siddique, 2015). Since this discovery other genes have been found to be involved with ALS or causing ALS like symptoms, such as Sequestosome-1 (SQSTM1) and Ubiquitin 2 (UBQLN2) which are involved in protein degradation, the involvement of Profilin 1 (PFN1) in actin polymerisation and C9ORF72 which is involved in RNA transcription (Ajroud-Driss and Siddique, 2015; Mejzini *et al.*, 2019). This is not the full genetic picture of genetic

involvement in ALS however, with other genes such as Microtubule Associate Protein Tau (MAPT), Nucleoporin GLE-1, Nuclear factor kappa-light-chain-enhancer of activated B cells (NFkB), along with many others also implicated (Nguyen, Van Broeckhoven and van der Zee, 2018). Sporadic ALS on the other hand is less clear in its aetiology, with many genetic and environmental factors thought to be involved in the progression of the disease. There have been a variety of loci identified as risk factors via genome wide studies with some overlap in genes that play an important role in both familial and sporadic ALS (Douville *et al.*, 2011). TAR DNA Binding Protein 43 (TDP-43) is one of the proteins which are affected in both conditions and it has been shown to have links to HERV activity. TDP-43 works in a regulatory role in HIV retroviral infection and the paper by Li *et.al.* shows that the activity of this protein is concurrent to the expression levels of HERV-K. Introducing an activated TDP-43 sequence into human neuronal and HeLa cells showed increased expression of HERV-K indicating a positive relationship (Li *et al.*, 2015). This was further shown by siRNA inhibition of the TDP-43 gene and its subsequent negative effect on HERV-K transcription *in vitro*. The study also investigated the potential for binding loci effecting the HERV sequence. Bioinformatics analysis of the genome showed that there were 5 binding loci identified next to HERV-K sequences, which were confirmed to interact with the TDP-43 molecule by co-immunoprecipitation (Li *et al.*, 2015). The link between HERV-K and TDP-43 has also been observed in aggregation of ALS-linked TDP-43 mutant proteins, which significantly increased HERV-K viral proteins in neurons (Manghera, Ferguson-Parry and Douville, 2016). It was also shown in the study by Manghera, Ferguson-Parry and Douville, 2016, that while astrocytes were able to clear the accumulation of HERV-K viral proteins through autophagy, neurons were incapable of clearing these viral proteins. In addition to these studies there has been more recent acknowledgement of the regulatory behaviour of TDP-43 for retrotransposons like HERV-K. In a paper by Romano, Klima and Feiguin (2020) in a Drosophilia model of expression, silencing TDP-43 resulted in a strong upregulation of HERV family elements (annotated as RTEs in the paper) which was corrected when TDP-43 expression was restored. In addition to this a positive correlation between HERV-K and TDP-43 was also seen when a study looked at the abundance of antibodies against HERV-K *env* epitopes and TDP-43 in plasma of ALS patients (Simula *et al.*, 2021). This study found a strong positive correlation between these two antibody targets which increased with disease progression, and could indicate that these cells are

over expressing TDP-43 in order to combat the increased HERV-K expression in cells from the disease state (Simula *et al.*, 2021). This over-expression of TDP-43 could then contribute to an increase in oxidative stress on the motor neuron leading to pathology observed in ALS (Zuo *et al.*, 2021).

An additional protein, B-cell lymphoma 11b (BCL11b), has also been suggested as a modifier of HERV expression in ALS. Originally discovered as a key regulator of differentiation and survival of T-lymphocytes during development, BCL11b has been shown to be involved in the growth of neuronal cells and the process by which these cells direct axonal growth to reach the correct targets (Lennon *et al.*, 2017). BCL11b is highly expressed in the pyramidal betz cells originating in layer V of the motor cortex (Figure 1.2 D) whose axons project down the spinal cord; however expression is absent from those cells responsible for intercortical connections (Lennon *et al.*, 2017). BCL11b involvement in neurodegenerative diseases has been observed by its downregulation contributing to pathological effects seen in Huntingdon's disease and its effect on the expression of brain-derived neurotrophic factor (BDNF) in Alzheimer's disease. Its relevance to the regulation of HERVs in ALS however comes from its interaction with HIV when the virus infects the central nervous system (CNS) (Desplats *et al.*, 2013; Lennon *et al.*, 2016). In latent HIV patients showing neurological impairment with neurodegenerative symptoms similar to ALS there was a high expression of BCL11b (Desplats *et al.*, 2013). This increased expression of BCL11b in latent HIV cases was associated with decreased expression of pro-inflammatory proteins, thereby mitigating changes to the transcriptome brought on by HIV infection of the CNS (Desplats *et al.*, 2013). Its dual role in supressing activation of latent retroviral infection in HIV cases can provide an insight into its potential role in suppressing HERV transcription (Lennon *et al.*, 2016). BCL11b has been shown to silence the HIV long terminal repeat preventing the transcription of the HIV *tat* protein in a similar manner to TDP-43 (Lennon *et al.*, 2016). This works in tandem with BCL11b and as there is some evidence that TDP-43 can alter HERV-K expression in ALS (Li *et al.*, 2015) this could suggest that BCL11b may provide a similar function (Lennon *et al.*, 2016). Further suggestive evidence of BCL11b's potential involvement in ALS comes from an expression study performed by Andrés-Benito *et al.* (2017) where it was found to be significantly upregulated in its function as a DNA/RNA transcriptional regulator. This would potentially

be in a similar function to its role in pathological conditions, interacting with positive transcription elongation factor b (P-TEFb) to regulate the activity of RNA polymerase in the cell and suppressing viral transcription (Cherrier *et al.*, 2013).

1.5.2 HERV Involvement in ALS

Even though there has been some evidence of retroviral involvement in neurodegenerative conditions since the mid 70's it is only recently that more solid research has been done in the area confirming the presence of upregulated HERV RNA transcripts in patient post-mortem brain tissue (Viola *et al.*, 1975; Norris, 1977; Li *et al.*, 2015). The primary and most recent work linking ALS with HERV-K expression has been published by Li *et.al.* in their 2015 paper "Human endogenous retrovirus-K contributes to motor neuron disease" which identifies a link between the endogenous retrovirus and the disease. Other papers have also drawn links to the involvement of retroviruses in the aetiology of sALS, one which identified Reverse Transcriptase activity (a retroviral marker) in 50% of serum samples from ALS patients compared to 7% in non-ALS controls and eliminated XMRV from being involved in the disease (McCormick *et al.*, 2008). In addition to this work Steele *et.al.* in their 2005 paper found increased serum reverse transcriptase activity in ALS patients, compared with their non-blood relative spousal controls, with no familial history of the disease. This paper also noted that blood relatives of the ALS patients also had high serum RT activity when compared with the controls, and hints at the involvement of an endogenous retrovirus in this disease (Steele *et al.*, 2005). Endogenous retroviral expression has also been observed in the frontal cortex of *C9orf72* positive ALS patients (Prudencio *et al.*, 2017). The gene present at C9orf72 (open reading frame 72 of chromosome 9) is abundant in nerve cells of the CNS and is involved in RNA binding. The mutation associated with ALS in the C9orf72 region is a hexanucleotide G_4C_2 repeat expansion, resulting in a reduction of the "healthy" protein and aggregation of the mutated sequence (Liu, Russ and Lee, 2020). This gene is thought to contribute to ALS in one of three ways, by loss of function of the protein, the addition of repeat RNA sequences in its loci and dipeptide repeat regions introduced by repeat associated translation (Balendra and Isaacs, 2018). The accumulation of the repeat RNA sequences as a toxic contributor to the pathology in ALS has been called into question however as a paper by Liu, Russ and Lee (2020) showed the mutation was accompanied by mild expression changes. The study

proposed that the pathologic mechanism of the C9orf72 mutation was instead mediated by the removal of TDP-43 positive cells in the region and a reduction of the proteins expression (Liu, Russ and Lee, 2020). The study by Prudencio *et al.*, (2017) confirmed that repetitive elements, including endogenous retroviruses, had increased expression, correlating to RNA polymerase activity in postmortem brain tissue of ALS patients.

The mechanism for HERV-K involvement and neurotoxicity via the *env* gene has also been explored by Li *et.al.* 2015 in which increased expression of HERV-K env was measured by qPCR in post-mortem brain tissue of sporadic ALS patients compared to matched controls and caused toxicity in neuronal cells that had been transfected with the HERV-K env gene *in vitro*. In addition, transcripts of the *pol* gene, have also been shown to be upregulated in ALS brain tissue, giving further evidence as to their involvement in some cases of ALS, and seem to be especially increased during chronic illness (Mortelmans, Wang-Johanning and Johanning, 2016). Support of a potential involvement of HERV-K *env* in ALS can be seen in the superior frontal cortex of younger ALS patients presenting with TDP-43 related frontotemporal dementia. In the paper by Phan *et al.* (2021) researchers found an increase in expression of HERV-K *env* transcripts in both serum and post mortem brain tissue samples with the reverse transcriptase protein of HERV-K localised to TDP-43 deposits in the tissue samples. However, there have been more recent studies looking at HERV-K differential expression in ALS which seem to disprove this initial finding. In a differential expression study involving patient samples from both brain and spinal cord HERV-K (HML-2) transcripts were found to have no significant differences between ALS and Controls despite showing high levels of variation between samples and tissue types (Mayer *et al.*, 2018). A paper by Garson *et al.*, 2019, looking at HERV-K (HML-2) and HERV-W gene transcript expression in premotor cortex of ALS patients in a UK cohort also found no statistically significant differential expression when comparing ALS and Controls by RT-qPCR. Another paper by Ishihara *et al.* (2022) also found no significant difference in HERV-K expression in post mortem motor cortex tissue samples when using digital droplet RT-qPCR in a Japanese dataset.

While HERV-K has had more research conducted on its involvement with ALS, HERV-W has also been shown to be upregulated in ALS cases (Küry *et al.*, 2018). Links between HERV-K

and HERV-W have been observed in a study looking at immune responses to HERV proteins from these families in ALS. The study found antibodies directed against both HERV-K and HERV-W *env* proteins, finding significant elevation in expression (Arru *et al.*, 2018a). HERV-W has been associated with the expression of SOD-1 in familial cases of ALS, indicating some involvement with the neurological symptoms associated with the disease (Ajroud-Driss and Siddique, 2015; Li *et al.*, 2015). It is proposed that the activation of this HERV in relation to SOD-1 damages cellular processes by oxidative stress after activation (Li *et al.*, 2015). In a study conducted in 2007 the researchers showed that HERV-W *env* proteins could also be detected in affected muscle tissue (Oluwole *et al.*, 2007). This suggests there could be a relationship between the expression of these ERV's, though any definitive link between them has yet to be explored. There is some evidence to suggest that environmental factors can activate transcriptional activity of HERV-W, with its sequences being elevated in other viral infections (Brudek *et al.*, 2007; Grandi and Tramontano, 2017; Xue *et al.*, 2018). While this activity in muscle cells has been reported, HERV-W's involvement in other neurological diseases, such as MS, Schizophrenia and severe Depression disorder, have been better documented in the literature (Magiorkinis, Belshaw and Katzourakis, 2013; Slokar and Hasler, 2016; Küry *et al.*, 2018).

1.6 Neurotoxicity of HERV Elements

The modulation of genetic elements via the insertion of proviral promotors upstream of a gene sequence is not the only factor in the progression of neurological disorders. While the activation of certain genes is an important part of the aetiology of the disease, individual proteins from HERVs can cause damage to their host cells in disease states or cases of overregulation (Grandi and Tramontano, 2018; Küry *et al.*, 2018). This is alongside the protective effect HERV *env* proteins can have on neuronal cells such as, during HIV infection (Bhat *et al.*, 2014).

Referring to what was mentioned earlier, in the study by Li *et.al.* activation of HERV-K has been shown to cause toxicity in neuronal cells. This was observed when both the full proviral transcript and sole *env* proteins were expressed with a similar rate of cell death suggesting that the *env* transcripts were the cause of the cellular toxicity (Li *et al.*, 2015). This was confirmed as causal by observing the result of neuronal damage and measuring

for increased HERV-K expression as a result. Further experiments involving the addition of *env* and promoter sequences into transgenic mice for *in vivo* observations resulted in the death of neuronal cells compared to controls (Li *et al.*, 2015). In cells which had not fully succumbed to *env* expression also showed morphological changes, with the number of dendritic branches decreased (Li *et al.*, 2015). The relationship between this *env* toxicity and ALS has also been hinted at, with the specific loss of mass in the motor cortex observed in transgenic animals with HERV-K expression upregulated with no analogous loss of brain tissue from other areas of the brain and CNS (Li *et al.*, 2015).

Indirect effects of HERVS on cellular toxicity take the form of HERV-W *env* protein which encodes Syncytin-1, a protein which has been shown to be overexpressed in glial cells in MS (Antony *et al.*, 2011; Bhat *et al.*, 2014). This protein works by modulating the immune response and creating an environment of ongoing inflammation around neuronal cells in the disease leading to cellular toxicity. It modulates the immune system into ongoing inflammatory action via it's interaction with Toll-Like Receptor 4 signalling mechanism, triggering a cytokine release (Nath *et al.*, 2015; Slokar and Hasler, 2016). This feeds into an expression loop with inflammatory mediators such as TNF-α activating the normal expression pathway for Syncytin-1 (Hera and Urcelay, 2016). Peptides from HERV-K's *env* region have also been observed to modulate inflammatory mediators. A paper by Arru *et al.* (2021) showed 2 peptide regions which together activated IL6, IFN-γ and IFN-α, key modulators of the inflammatory response in human immunity. An earlier paper by the same author also confirmed that *env* regions in both HERV-K and HERV-W were able to modulate the immune system in ALS (for HERV-K) and MS (for HERV-W) (Arru *et al.*, 2018a). HERV-K *env* was also identified in the serum section of blood samples extracted from frontotemporal dementia presenting ALS patients (Phan *et al.*, 2021). Other CNS related cells such as astrocytes and oligodendrocytes when exposed to serum from MS patients undergo apoptosis, which hints at the toxicity of this *env* protein to non-neuronal cells (Hera and Urcelay, 2016; Buzdin, Prassolov and Garazha, 2017; Grandi and Tramontano, 2017).

The interaction of these *env* proteins from these separate HERVs with their host cell are the primary examples of direct HERV toxicity in Neuronal cells, and other methods of

toxicity lie within the activation of the hosts innate and humoral immune responses (Grandi and Tramontano, 2018; Gröger and Cynis, 2018).

1.7 HERV expression in Peripheral Blood Mononuclear Cells (PBMCs) as a potential biomarker of ALS

Amyotrophic Lateral Sclerosis (ALS) is increasingly recognised as a multi-system disease with changes in gene expression and cellular activation not isolated to the progressive loss of motor neurons (Vijayakumar *et al.*, 2019). Also affected within the Central Nervous system (CNS) are astrocytes, cells that help maintain and support the environment within the brain, which have been observed to undergo morphological and gene expression changes in response to the progressive neurodegeneration (Yamanaka and Komine, 2018). As the motor neurons begin to die off changes in muscle expression pathways can also be observed, notably those relating to cellular metabolism and cell growth (Vijayakumar *et al.*, 2019).

Peripheral blood mononuclear cells (PBMCs) are defined as cells in circulating blood which contain a round nucleus including monocytes, lymphocytes (T-Cells, B-Cells & Natural Killer Cells) and dendritic cells (Kleiveland, 2015). From these cell types the largest population is lymphocytes, making up to 70%-90% of PBMCs present in blood, monocytes are the next largest at 10%-20% while dendritic cells are rare, only comprising around 1-2% of PBMC content (Kleiveland, 2015). PBMCs are useful in tissue expression studies as it has been shown that over 80% of gene expression in a given tissue is mirrored in PBMCs (Liew *et al.*, 2006). PBMCs also have the advantage of being readily sampled from patients where some tissue biopsies, i.e. regions of the brain including the premotor cortex, would prove to be difficult to obtain or potentially harmful to the patient (Liew *et al.*, 2006). In ALS a mirror of mitochondrial dysfunction in muscle and brain tissue can be observed in PBMCs, with cytochrome-c-oxidase (the last enzyme in the electron transport chain in ATP synthesis) activity decreasing with increasing severity in disease state (Ehinger *et al.*, 2015); this provides more evidence of ALS as a multi-system disease (Ehinger *et al.*, 2015).

As mentioned in section 1.4 there are incidences of ALS-like symptoms occurring in retroviral infections such as HTLV-1 and concurrent altered expression of HERV-K (HML-2)

transcripts in HIV (Garcia-Montojo *et al.*, 2018). In HTLV, those patients presenting with ALS-like symptoms had generally high levels of virus detected in PBMCs, and a separate study showed that HTLV-1 *tax-rex* sequences can be detected in PBMC's of 40% of ALS patients (Alfahad and Nath, 2013; Ando *et al.*, 2015). With HERVs being considered as a new biomarker in ALS disease it is important to discover whether these viral transcripts will be detectable in PBMCs, allowing for an alternative approach to measure HERV expression outside of the CNS and preserving precious brain donor tissue as gene expression in blood and the brain have been found to be highly correlated (Rollins *et al.*, 2010). Research has confirmed that HERVs are transcriptionally active in PBMCs, with expression levels changing as the body ages and different HERV families being associated with different developmental stages; HERV-E for example not being expressed in early childhood and HERV-W *env* expressed during foetal development as its protein, syncytin1 is responsible for the tight binding of epithelial cells in the placenta (Balestrieri *et al.*, 2015; Grandi and Tramontano, 2017). In regards to neurological conditions HERV expression has already been detected in increased levels in PBMCs in Autism, as well as in multiple sclerosis (MS) (Balestrieri *et al.*, 2012; Tam, Ostrow and Gale Hammell, 2019). The use of PBMCs as an biomarker for ALS has also been suggested recently as a paper showed that there was an increase in HERV-K *env* peptide expression on the surface of B-Cells and NK cells which could modulate the inflammatory response (Arru *et al.*, 2021).

As gene expression in PBMCs has been shown to be an effective mirror for gene expression in those tissues that would be difficult or damaging to reach for monitoring disease progression it provides a useful means for detecting novel biomarkers in ALS and other conditions (Liew *et al.*, 2006; Ehinger *et al.*, 2015; Tortarolo *et al.*, 2017). If HERVs are identified as a potential novel biomarker and pathogenic determinant in ALS, then measurement of their expression in PBMCs would provide a less invasive method of monitoring disease progression (Dolei *et al.*, 2019). PBMCs also offer the best opportunity to detect mirroring of CNS HERV expression as, unlike other RNAs, they cannot be detected outside of isolated cells (Bhardwaj *et al.*, 2014; Hosaka *et al.*, 2019)

1.8 Association of HERVs in Multiple Sclerosis (MS) & Schizophrenia

MS or Multiple Sclerosis is a chronic disease of the central nervous system involving the demyelination of neuronal cells and the build-up of scar tissue, from which it gets its name

(Antony *et al.*, 2011; Hera and Urcelay, 2016). MS effects women more than men and can occur between the ages of 20-50 with an unknown environmental or genetic trigger which activates the condition (Antony *et al.*, 2011). While the exact cause of this disease has yet to be determined the involvement of HERVs has been extensively documented (Fujinami and Libbey, 1999; Morandi, Tarlinton and Gran, 2015; Morandi *et al.*, 2017). The main HERV pathogenic component contributing to MS is HERV-W's *env* protein syncytin-1, a surface protein highly expressed in disease state and a known modulator of the immune system involved in cellular toxicity.

The only *env* gene of HERV-W that has a full open reading frame for syncytin-1 is located on chromosome 7, located within a relatively intact proviral sequence (Hera and Urcelay, 2016). The primary difference between this HERV-W sequence and its pathogenic counterpart seems to be a 12bp insertion in a transmembrane section of the protein (Hera and Urcelay, 2016). This difference is suspected to arise from either genetic recombination or originating from a separate HERV-W sequence located elsewhere in the Genome (Hera and Urcelay, 2016). One suspected candidate occurs on the X Chromosome in location Xq22.3, and differs from the main *env* sequence by a truncated N-terminal region (Hera and Urcelay, 2016). This protein still retains the ability to be transcribed, even with its truncated end and retains the mutational potential to encode the full protein with the removal of the stop codon from its reading frame (Hera and Urcelay, 2016). An additional genetic study looking at insertional variations in *env* candidates also identified a polymorphism in this locus, which was related to an increase risk factor in women for MS (García-Montojo *et al.*, 2014; Hera and Urcelay, 2016). This polymorphism was also shown to be able to produce the pathogenic version of syncytin-1 (Hera and Urcelay, 2016).

In addition to the involvement of HERVs in MS the involvement of other viruses in the onset of the disease has also been of some interest to the scientific community (Mameli *et al.*, 2013). Amongst these viruses of interest was EBV, a known pathogen of immune cells, with some links to the onset of MS in those with genetic risk factors for the disease (Mameli *et al.*, 2013; Morandi *et al.*, 2017). With continuing research into HERV-W and MS it was shown that HERV-W transcripts were increasingly expressed during EBV induced Infectious Mononucleosis (Mameli *et al.*, 2013). Alongside this association it has been stated in a

recent study that a surface glycoprotein of EBV, gp350, can activate HERV-W expression in targeted cells (Grandi and Tramontano, 2017). It has also been shown that those with genetic risk factors who have suffered Infectious Mononucleosis have a higher risk of developing MS (Mameli *et al.*, 2013; Nath *et al.*, 2015). From these observations, there looks to be an association of EBV with HERV-W expression in terms of activation of proviral sequences in MS and requires further investigation.

The expression of MS related HERV-W is not just regulated to the brain as its expression can also be detected in Peripheral Blood Mononuclear Cells (PBMC's) (Antony *et al.*, 2011; Nardo *et al.*, 2011; Küry *et al.*, 2018). These *env* transcripts are generally found in monocytes in areas of recent demyelination where they can also be detected in nearby blood vessels (Morandi, Tarlinton and Gran, 2015). While HERV-W and HERV-H activity has been observed in these cells there has also been studies into other immune related cells. Monocytes taken from patients with active MS have also been shown to have gliotoxic attributes *in vitro* when harvested from peripheral blood (Morandi, Tarlinton and Gran, 2015). While expression has been detected in B-Cells, Monocytes, Macrophages and Natural Killer cells no expression has been detected in T-Cells (Morandi *et al.*, 2017). Expression of additional HERV families in MS has also been categorised with *gag* & *env* transcripts from HERV-E and HERV-K along with *env* from HERV-W were shown to be significantly expressed compared to the control group, though differences between disease status groups were small (Bhetariya, Kriesel and Fischer, 2017). Another family member, HERV-Fc1, has been found to have increased extracellular expression in MS patient peripheral blood, with a 4-fold increase reported (Laska *et al.*, 2012).

While Schizophrenia may not initially seem related to MS it does share some similarities in terms of its proposed aetiology, with unknown genetic and environmental factors being just one of these links (Nath *et al.*, 2015). Other factors that these diseases share are a similarity in age of onset and geographic distribution (Nath *et al.*, 2015). One of the primary commonalities between MS and Schizophrenia is the HERV families associated with the neurological condition (Slokar and Hasler, 2016). In studies relating to the expression of HERVs in this disease shown that HERV-W and HERV-K10 were significantly upregulated in the frontal cortex of schizophrenic patients (Slokar and Hasler, 2016). It was also found that

these transcripts were detected more readily in recent-onset patients rather than chronic sufferers indicating that HERV transcripts play a role more in the onset of the disease rather than its continuation (Slokar and Hasler, 2016). It has been proposed that the mechanism for HERV-W involvement in schizophrenia is through Syncytin-1's mediation of the inflammatory response (Wang, Huang and Zhu, 2018). Syncytin-1 has been shown to activate IL-1β and IL-6, both of which have been found to be overexpressed in schizophrenic patient brain tissue (Wang, Huang and Zhu, 2018). It has also been shown to induce other inflammatory markers seen in schizophrenia, such as C-Reactive Protein (Wang, Huang and Zhu, 2018).

Other HERV involvement in schizophrenia comes from gene regulation, either downregulating or enhancing expression depending on the gene involved. Two such genes, Gamma-Aminobutyric Acid Type B Receptor Subunit 1 (GABBR1) and Proline Dehydrogenase 1 (PRODH) are controlled by HERV-K and are downregulated and promoted respectively (Suntsova *et al.*, 2013). It has been stated in the study by Slokar & Hasler (2016) that the upregulation of one family of HERV, in the case of this research paper HERV-W, can mean the downregulation of another in schizophrenia, which the study observed to be ERV-9 resulting in a disruption in the balance of protein expression in the cell. This gives further evidence to a multifactorial role of HERV elements in the pathogenesis of this condition.

1.9 Association of HERVs in non-neurological diseases

As HERV sequences have been found acting as promotors and enhancers of genes it is perhaps unsurprising that they would, in a state of dysregulation, also play a role in disease. The function of individual HERVs in disease state varies depending on the condition and genetic make-up of the retro element, from individual *env* proteins being involved in cellular toxicity or interactions with host signalling via surface expression, to a more general role in transcriptional modification (Shin *et al.*, 2013; Nath *et al.*, 2015; Grandi and Tramontano, 2018; Dervan *et al.*, 2021). This varied role in disease states is shown in table 1, where we can see a brief overview of some HERVs normal gene function and the disease

state which follows dysregulation/mutation. While this table does not include all known disease functions, it does show how varied diseases can be within a family of HERVs.

Table 1.1. A Summary of the information contained in tables 1, 2 & 3 from Cohen, Lock and Magor's 2009 paper "Endogenous retroviral LTRs as promoters for human genes: A critical assessment".

The table below shows only those entries associated with disease or disease progression (Cohen, Lock and Mager, 2009). Also includes information obtained from Suntsova et al., 2013, Mameli *et al.*, 2009, Bashratyan *et al.*, 2017 & Krzysztalowska-Wawrzyniak *et al.*, 2011. Genes: ERV1 - Endogenous Retrovirus 1, MaLR - Malate Response Regulator, HERV – Human Endogenous Retrovirus.

Gene Name	HERV Type & Location	Function (Disease)	Human non-LTR expression	LTR Expression
TMPRSS3 (transmembrane protein, serine 3)	MLT1C/MaLR chr21:42683443–42683623	Serine protease (deafness, cancer)	Widespread, upregulated in cancer	PBL
CYP19A1 (aromatase)	MER21A/ERV1 chr15:49417965–49418479	Oestrogen synthesis (cancer)	Skin, adipose, brain, gonad	Placenta
IL2RB (interleukin-2 receptor B)	THE1D/MaLR chr22:35900862–35901238	Lymphocyte proliferation	Lymphocytes, cancer	Placenta
PTN (pleiotrophin)	LTR2B/HERV-E chr7:136603664–136604133	Neural development (neurodegenerative diseases)	CNS, testis, uterus, placenta, cancer	Placenta
Unknown, HERV-K113 insertion.	HERV-K113 chr1919:21633273–21633575	Systemic Lupus Erythematosus/ Rheumatoid Arthritis	Unknown	Unknown
Complement Protein C4, Variant	HML-2/HERV-K Chr6:32.01M-32.04M (Intron 9)	Compliment System (Diabetes when HERV-K insertion present)	Blood	Unknown
Syncytin-1 (Enervin)	HERV-W chr7:92468380-92477915	Cell-Cell Fusion, Placental Attachment (Multiple Sclerosis)	Unknown	Placenta
PRODH (proline dehydrogenase)	HML-2/HERV-K chr22:18920392-18928317	Neurotransmitter Synthesis, (Schizophrenia)	Unknown	CNS
DNAJC15 (DNAJ domain-containing)	LTR7Bd/HERV-H chr13:42531808–42532236	Regulates Hsp70 (cancer)	Widespread in normal and cancer	Widespread, additional cancer cell lines

Cancer cell lines make use of various HERV elements depending on the cell of origin and, inevitably, the genes with which the HERV is associated. While evidence that these endogenous elements directly result in a cell progressing into a malignancy is scarce their function in cancer covers a wide variety of processes, either in the form of gene promotion or as a result of HERV protein expression (Yi, Kim and Kim, 2006; Agoni, Guha and Lenz, 2013; Downey *et al.*, 2015; Johanning *et al.*, 2017). Functioning as transcriptional modifiers in cancer we see a couple of entries in table 1, with HERV-H's modulation of a Hsp70 regulatory gene being a good example (Cohen, Lock and Mager, 2009). Extensive methylation of the genome seen in some cancers also influence genes controlling HERV expression, removing the cells normal control over these LTR's (Katoh and Kurata, 2013; Grandi and Tramontano, 2018). Cellular control of LTR's comes in the form of suppression of transcription utilising innate immune defences against retroviral infections, by suppressing the provirus the cell gains control over the downstream genetic element (Sverdlov, 1998). This reduction in transcriptional control can cause the expression of these proviral sequences, leading to their involvement in expression of downstream genes. A potential pathway of HERV involvement in cancer progression lies in their potential for chromosomal rearrangement, as an example HERV-K related insertional changes in the genome have been implicated as a risk factor in lung cancer development (Hohn, Hanke and Bannert, 2013; Gonzalez-Cao *et al.*, 2016). Other HERV-K proteins involved in cancer development are *Rec* & *Np9* which are known transcriptional and immune system modifiers with *Np9* acting as a regulator of the apoptosis related protein p53 (Schmitt *et al.*, 2015). *Rec*'s involvement in cancer is less clearly defined though it has been shown to interact with the androgen receptor in human cells along with the mitotic associated human small glutamine-rich tetratricopeptide repeat-containing protein (hSGT) (Hanke *et al.*, 2013). Other HERV families have also been shown to have involvement in cancers, with HERV-H being prominent in colorectal cancer (Zhang, Liang and Zheng, 2019). In addition, a HERV-H element located on the X-chromosome is responsible for a number of gastrointestinal cancers, detected in multiple stages of disease progression (Zhang, Liang and Zheng, 2019).

While we have mentioned alternate splice variants and their involvement in cancer, HERV-K *env* proteins are also shown to have pathogenic properties relating to both autoimmune disease and cancer due to the effect they have on many different biological processes. This

includes functioning as immune system suppressors and activators, being involved in apoptotic processes, inducing abnormal cell-cell interactions, acting as inflammatory mediators and having some cytotoxic elements as well (Downey *et al.*, 2015; Grandi and Tramontano, 2018). As tumour cells rely on cell-cell fusion to spread, HERV involvement has been theorised to aid in their ability to attach to numerous cell types; this is primarily due to the function of *env* proteins in viral infection, acting as an attachment point for the viruses to gain entry into cells (Grandi and Tramontano, 2018). The HERV-K family *env* protein also plays a role in the pathogenesis of breast cancer, with its upregulation being linked to many negative factors of the disease, effecting the metastatic spread and being an indicator of p53 mutation (Zhao *et al.*, 2011; Grandi and Tramontano, 2018). This protein has also been detected as significantly expressed in breast cancer tissue sections compared to control samples taken from surrounding healthy tissue (Zhao *et al.*, 2011). This *env* region has also been used in a therapeutic trial targeting HERV-K showing a result in slowing the metastatic spread of breast cancer (Zhou *et al.*, 2015). In melanoma, the expression of HERV-K (HML2) *env* protein has the ability to activate humoral immunity, which was detected in around 20% of cases, and associated with a poor prognosis in those who have anti HML-2 antibodies (Hahn *et al.*, 2008; Grandi and Tramontano, 2018). Further evidence of HERV-Ks role in cancer comes from a study disrupting HERV-K genes using CRISPR/Cas-9 technology, which showed that disruption of the HERV-K *env* gene interfered with RNA binding, alternate splicing of epidermal growth factor receptor and other cellular proteins (Ibba *et al.*, 2018).

HERV-K expression has also been implicated in prostate cancer cell lines, with transcripts from both positive and negative strands expressed in cell lines (Agoni, Guha and Lenz, 2013; Johanning *et al.*, 2017). While the study by Agoni, Guha and Lenz (2013) showed definite involvement of HERV-K in prostate cancer it did not expand on the function of these retroviruses. The detection of HERV-K *env* RNA transcripts in prostate tumours has been hinted at in other studies as well (Hohn, Hanke and Bannert, 2013; Katoh and Kurata, 2013; Magiorkinis, Belshaw and Katzourakis, 2013). It has also been described in research studies that the RT inhibitor Abacavir, when used on tumours with detectable elevated HERV-K transcription, provided an anti-proliferative effect on cancer cells, providing evidence of a link between its expression and prostate cancer pathology (Grandi and Tramontano, 2018).

Within the scope of the differential regulation of HERV-K transcripts in prostate cancer there seems to also be a relationship between increasing age and the detection of HERV-K transcripts in prostate cancer, with higher expression of HERV-K *gag* transcripts also seen in African patients (Wallace *et al.*, 2014). Transcripts of HERV-H *env* have also been found to be expressed in prostate cancer, where antibodies targeting the *env* of the HERV family were found to be significantly expressed alongside HERV-K (Manca *et al.*, 2022). Outside of HERVs H & K, other HERV families have also been found to be involved in cancer progression. HERVs W, F, R and S have also been found differentially expressed in cancer cell lines (Zhang, Liang and Zheng, 2019).

There has been shown to be some relationship between HERV-K and the development of diabetes mellitus (Bashratyan *et al.*, 2017). In mouse models of type 1 diabetes, *env* and *Gag* genes play a potential role in the progression of this disease, with *env* being shown to induce an autoimmune response in the pancreatic islets associated with their expression on excreted microvesicles (Bashratyan *et al.*, 2017). Other links between HERV-K and diabetes come from its relationship with complement protein C4 in which this complement factor has a HERV-K insertion in its 9th intron (collectively known as HERV-K(C4)) and the study showed that low copy numbers of HERV-K(C4) were linked to the patient suffering from type 1 diabetes (Mason *et al.*, 2014). The researchers theorised a couple of ideas on how HERV-K expression could be related to the presentation of type-1 diabetes in patients, such as effecting the complement pathway by affecting C4 RNA transcription (Mason *et al.*, 2014).

There has also been evidence to show a relationship between HERVs and certain autoimmune diseases such as Rheumatoid Arthritis and Systemic Lupus Erythematosus (SLE) (Ehlhardt *et al.*, 2006; Brodziak *et al.*, 2012). As there are multiple genetic and environmental factors that also contribute to these diseases, we can say that HERVs play a supporting role in these conditions. HERVs function in these diseases, aside from gene regulation come in 2 major forms, molecular mimicry and acting as superantigens (Ehlhardt *et al.*, 2006). Molecular mimicry has been shown between SLE elements and a *gag* protein from HRES-1 (HTLV-Related Endogenous Sequence-1) where it shows cross reactivity with an snRNP (Gröger and Cynis, 2018). On the side of superantigen activity, Epstein-Barr Virus

has been shown to activate HERV-K18 and has been implicated in juvenile idiopathic arthritis (Mameli *et al.*, 2013). An indirect effect of immune system modulation in autoimmune disease is the contribution of HERV-K expression to Pulmonary Arterial Hypertension (Saito *et al.*, 2017). The related study showed that elevated HERV-K expression in circulating macrophages could initiate and sustain vascular changes that resulted in the condition (Saito *et al.*, 2017). Concurrent to these changes HERV-K expression was found to upregulate an inflammatory cytokine, IL-6, though the researchers stated that it was difficult to infer direct cause and effect based on the study (Saito *et al.*, 2017). As HERVs have had clear involvement in inflammation and autoimmune disease it is not surprising that their increased expression has been noted as having a negative effect in Hepatitis C (Weber *et al.*, 2021). In the study by Weber *et.al.* (2021) the research team tested the response of HERV-K (HML-2) elements to antiretroviral therapy and found a decreased expression with ongoing treatment, linking the expression to the levels of albumin found in the cirrhotic tissue.

1.10 Molecular Approaches to Analysing Gene Expression.

1.10.1 Quantitative Real-Time Polymerase Chain Reaction (RT-qPCR) and the Importance of Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE)

Quantitative Real-Time Polymerase Chain Reaction (RT-qPCR) is a method of quantifying gene expression and was developed in the early 1990's (Higuchi *et al.*, 1992; Heid *et al.*, 1996; Huggett *et al.*, 2005). This quantitative method uses the change in intensity of fluorescent reporters to identify the amplification of nucleic acids in real-time (Jia, 2012; Stephenson and Stephenson, 2016). There are 2 principle methods for the use of fluorescent dyes in RT-qPCR reactions, hybridisation probes (e.g. TaqMan probes), which use sequences specific to the gene target with a florescent reporter and quencher attached, and non-specific DNA binding dyes (SYBR green chemistry) (Jia, 2012). There are advantages and disadvantages to each of these methods for RT-qPCR, for SYBR green chemistry the non-specific binding of the dye to any dsDNA means it can be used to monitor the amplification of any dsDNA sequence and reduces assay cost (Smith and Osborn, 2009). The primary disadvantage is the same as its strength, binding to any dsDNA sequence

means it can also bind to non-specific sequences in the reaction mix (Smith and Osborn, 2009). TaqMan probes are highly specific to their gene target however and can be labelled with different fluorescent dye colours, enabling multiplexing, running multiple reactions in the same mix (Smith and Osborn, 2009). The RT-qPCR process is divided into 4 stages, Ground, Exponential, Linear and plateau. The initial linear ground stage is the beginning steps of the reaction where the primer targets are being amplified but the fluorescence is too low to be detected. This is followed by the exponential phase where the fluorescent signal has risen above background levels significantly enough to be detected and the reaction product concentration is doubling with each cycle. This stage is where the Cycle threshold (Ct) of the reaction is recorded. The final stages of the RT-qPCR reaction are the linear and plateau where the reaction reagents are running low and amplification efficiency starts to decline, the end point of the reaction is the plateau where no more product is being produced and therefore no change in fluorescence is detected (Jia, 2012).

The quantification of target sequences by RT-qPCR can either be absolute, requiring calibration curves of known dilutions to compare against or relative, comparing to a stably expressed reference gene (Schmittgen and Livak, 2008). Relative quantification of gene expression by RT-qPCR is used to determine change in expression values between patient sample sets as opposed to determining copy number as in the absolute method (Livak and Schmittgen, 2001; Schmittgen and Livak, 2008). The main method for calculating relative quantification in expression studies is the ΔΔCt method, a mathematical model for calculating the relative change in expression of target genes (Livak and Schmittgen, 2001). This mathematical model is given in figure 2 and provides a simple method for the calculation of relative expression changes between 2 patient groups. It can be noted that the order of ΔCt in the second row of Figure 2 is interchangeable but has an effect on how results for the model are interpreted (Schmittgen and Livak, 2008).

$$2^{-\Delta\Delta Ct}$$

$$\Delta\Delta Ct = \Delta Ct(Disease\ Patient\ Sample) - \Delta Ct(Control\ Patient\ Sample)$$

$$\Delta C = Ct\ Gene\ of\ Interest\ - Ct\ Reference\ Gene$$

Figure 1.3. ΔΔCt Equation for the Relative Quantification of a Gene of Interest.

The equation given in the top row of the figure is the ΔΔCt method for relative quantification of a gene of interest with the following 2 rows describing the method of obtaining a value for ΔΔCt from control and disease state patient samples with a gene of interest. (Schmittgen and Livak, 2008)

An alternate method for the relative quantification of RT-qPCR data exists in the form of the Pfaffl mathematical model (Figure 1.4) (Michael W. Pfaffl, 2001). This model accounts for the variability between experimental amplification efficiencies of gene targets used in RT-qPCR assays. Accounting for these differences in primer efficiencies increases the accuracy of expression data; it enhances reproducibility by standardising each reaction run between tissue types or region of tissues which may have variations in sample loading or differing ranges of cDNA input (Michael W. Pfaffl, 2001).

$$Relative\ Quantification\ Ratio = \frac{(E_{GOI})^{\Delta Ct\ GOI}}{(E_{RG})^{\Delta Ct\ RG}}$$

Figure 1.4. Pfaffl Equation for Relative Quantification of Gene of Interest.

The equation given in the image above is the Pfaffl mathematical model for relative quantification of a gene of interest (GOI). This equation factors in the experimentally derived amplification efficiency ($E_{GOI/HKG}$) of both the GOI and the reference genes (HKG) (Michael W. Pfaffl, 2001).

Housekeeping genes, more recently referred to in the literature as reference genes, are those related to the maintenance of cellular processes so are assumed to be expressed across both disease and non-disease state (Eisenberg and Levanon, 2013). An ideal reference gene candidate for quantifying expression should be stably expressed across all samples being studied, regardless of tissue type or experimental conditions (Penna *et al.*,

2011). The historical gene of choice for reference comparison in multiple tissue types has been Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) due to its relatively high expression across multiple tissue types (de Jonge *et al.*, 2007). While this gene is highly expressed in most tissues individual expression rates can vary wildly between tissue types requiring validation of individual assays for ideal reference genes (Dean, Udawela and Scarr, 2016; Kuang *et al.*, 2018). An additional issue in using GAPDH is the presence of its pseudogenes, these lack introns and are similar in size to GAPDH (Sun *et al.*, 2012). The presence of these pseudogenes means that assays using GAPDH primers commonly mis-prime to pseudogenes (Sun *et al.*, 2012). This is most obvious in brain tissue samples where genes can vary wildly between different regions of the brain (Koppelkamm *et al.*, 2010; Dean, Udawela and Scarr, 2016). A study looking for stably expressed reference genes in ALS patient brain tissue discovered a novel reference gene, XPNPEP1 which had much higher stability in samples compared to the commonly used reference gene GAPDH (Durrenberger *et al.*, 2012). This study highlighted that while GAPDH is commonly used for its relatively high transcription in many tissues any study choosing to use it should verify it against a panel of reference genes before use (Durrenberger *et al.*, 2012).

The Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) provides a standardised framework for the inclusion of RT-qPCR data in research publications (Bustin *et al.*, 2009). The need for a gold standard in RT-qPCR publications arose due to a lack of consensus in what information should be published, leading to a lack of confidence in data obtained from some research studies (Bustin *et al.*, 2009). The aim of the document was therefore to provide a guideline for the publication of RT-qPCR data, modelled on similar papers providing guidelines for publication of Microarrays and proteomics all co-ordinated under Minimum Information for Biological and Biomedical Investigations project (Brazma *et al.*, 2001; Taylor *et al.*, 2007, 2008). The MIQE guidelines allow a list of information for data that should be included in a publication to ensure accuracy and allows for easy reproducibility of published data. This includes validation of reference genes and the recommendation that RT-qPCR data be normalised against more than a single candidate reference gene, mitigating any errors that could be encountered like variation of reference genes between differing regions of the brain (Bustin *et al.*, 2009; Koppelkamm *et al.*, 2010). Alongside the validation of reference genes the guidelines

suggest the inclusion of all data regarding kits and other assay validation processes such as primer efficiency data (Bustin *et al.*, 2009).

Since the publication of the MIQE guidelines however there have still been some papers which have only included a single reference gene without adequately justifying their choice in the text. These include the study by Li et al., 2015 where the use of a single reference gene, GAPDH has not been adequately justified. More recent studies involving the use of just a single reference gene include analysis of Long interspersed nuclear element (LINE-1) retrotransposons in autism brain tissue, the analysis of Tet methylcytosine dioxygenase 2 (Tet2) in mouse brain, and a study looking at biomarkers for human brain pericytes only normalising against GAPDH (Gontier *et al.*, 2018; Shpyleva *et al.*, 2018; Smyth *et al.*, 2018). These papers highlight a concern that researchers have not undertaken sufficient validation of their expression assays, resulting in a lack of confidence in their results. There are however papers which do follow MIQE compliance in their RT-qPCR data, a good example of which is the paper by Kuang *et.al.* which outlines some technical considerations when dealing with qPCR assays such as sample acquisition, RNA purification and optimising RT-qPCR performance (Kuang *et al.*, 2018).

1.10.2 Next Generation Sequencing Platforms and RNA-Seq of HERV elements.

Next Generation Sequencing (NGS) describes a number of different nucleotide sequencing platforms which are successors to the previous Sanger Sequencing model (Behjati and Tarpey, 2013). NGS provides a cheaper, faster, high throughput method of obtaining nucleotide sequences compared to the previous technology, and have a higher accuracy compared to Sanger Sequencing (Behjati and Tarpey, 2013). While NGS sequencing technologies differ in the method in which sequence reads are obtained they all sequence small fragments of nucleotide sequences multiple times in parallel generating massive amounts of sequence data which is then constructed into a contiguous sequence by comparing to a reference genome using bioinformatics tools (van Dijk *et al.*, 2014). In addition, a significant improvement over the previous method lies in the collection of nucleotide reads without the need for electrophoresis, with technologies collecting data directly from the sequencing reaction (van Dijk *et al.*, 2014). It also has the advantage of being able to collect sequencing data without bias, that is without needing prior knowledge

of the genome region being targeted, which allows the analysis of full genomes and discovery of novel gene mutations. Detecting mosaicism in the genome, or 2 or more populations of cells with different genotypes, is a good measure of the detection limits of Sanger and NGS technologies. While Sanger sequencing has a detection limit of 20%, not being able to detect differences in alleles below this value, NGS technologies can detect differences at or below 2%, making them much more sensitive in detecting minority variants present at low levels (Gajecka, 2016; Jamuar, D'Gama and Walsh, 2016).

The four main platforms for NGS are Illumina (MiSeq, HiSeq etc.), IonTorrent (PGM), Oxford Nanopore (MinION) and Pacific Biosystems (PacBio) (Quail *et al.*, 2012; Rhoads and Au, 2015; Lu, Giordano and Ning, 2016). The Illumina Sequencing platform works on sequencing by synthesis approach which generates multiple amplified copies of a sequence attached to an acrylamide flow cell by bridge amplification. This is followed by the addition of fluorescently tagged nucleotides which are washed over the flow cell one at a time then excited by laser light, those nucleotides which have attached give off light and are recorded by the MiSeq system (Ju *et al.*, 2006; Quail *et al.*, 2012). IonTorrent PGM works by attaching sequences to beads and filling proton detecting wells with them, then adding each of the four bases sequentially until the sequence fragments have been fully bound. The sequences are detected by the protons released as a by-product of the polymerase reaction on addition of a new nucleotide (Quail *et al.*, 2012). In the PacBio platform double stranded DNA sequences are circularised with the addition of hairpin loops at either end and added to a zero-mode waveguide (ZMW) sequencing unit (Rhoads and Au, 2015). The ZMW has a single polymerase molecule attached to the bottom and supplied with the 4 fluorescently labelled nucleotides, sequences being recorded by the individual light flashes given off as the polymerase molecule adds the fluorescently labelled nucleotides to the template sequence, recorded in real time (Rhoads and Au, 2015). Finally, the MinION system provided by Oxford Nanopore works by feeding double stranded DNA onto a motor protein by an adaptor and feeds single stranded DNA into the nanopore which records the base by measuring the change in ion current detected in the flow cell (Lu, Giordano and Ning, 2016). A study looking into the effectiveness of MiSeq, IonTorrent and PacBio platforms showed that all NGS methods showed high accuracy when dealing with GC rich regions while the IonTorrent displayed low efficacy when processing AT rich regions resulting in poor

coverage of around 30% of a *Plasmodium falciparum* genome (Quail *et al.*, 2012). In a study comparing MinION to the PacBio platform the researchers found similar read quality and length between the platforms (Lu, Giordano and Ning, 2016).

These NGS platforms can be subdivided into what is coming to be known as second generation, generating short read lengths (Illumina and IonTorrent up to 600 bases), and third generation, generating long reads from nucleotide data (PacBio and MinION >60 kilobases) (illumina, 2011; Rhoads and Au, 2015; Lu, Giordano and Ning, 2016; Scientific, 2017). For a sequence read to be useful it must be long enough for the sequence to be mapped specifically to the nucleotide template (Whiteford *et al.*, 2005). This means as the length of a read decreases the likelihood that the generated sequence information will map to multiple places in the reference genome increases (Whiteford *et al.*, 2005). While a minimum read length of 50 nucleotides (nt) produces 1000nt contiguous sequences that can cover 80% of chromosome 1 of the human genome this can have trouble when dealing with repeat elements such as HERVs which exist in fragments throughout the human genome (Whiteford *et al.*, 2005; Bhardwaj *et al.*, 2015; Mayer *et al.*, 2018). While long read NGS systems are more useful for building *de novo* genome sequences they also have advantages when dealing with repeat regions in the genome. When dealing with repeat sequences long read platforms allow for sequence reads that completely span low complexity regions, allowing mapping to specific chromosomes within the genome (Pollard *et al.*, 2018). Long reads also have a useful function in transcriptomics, allowing RNA transcripts obtained from tissues to be viewed full length, enabling the examination of splice variants (Wang *et al.*, 2016). One of the principle attractions of short read sequencing over long read however is its relative price, with long read platforms tending to be more expensive for generating sequence data.

High throughput RNA sequencing (RNA-Seq), utilises many NGS platforms for the characterisation of the transcriptome from tissues under analysis (Li *et al.*, 2014). In RNA-Seq the NGS library preparation differs slightly from DNA in that mRNA needs to be purified from extracted patient total RNA before fragmentation and cDNA synthesis (Atamian and Kaloshian, 2012). Beyond the identification of novel gene transcripts and mRNA splice variants, RNA-Seq allows for the quantification of gene expression across the whole

transcriptome and allows the expression of individual alleles to be identified (Kimberly R Kukurba and Montgomery, 2015). A study looking into the variation between sequencing platform for the gathering of RNA-Seq data in Pacific Biosystems, Illumina and Ion Torrent NGS platforms showed high similarity in efficiency of sequence reads but variable quality when detecting splice variants (Li *et al.*, 2014). The use of RNA-Seq in researching HERV activity in certain diseases has been useful to characterise its expression pattern in the transcriptome of effected tissues. Compared to RT-qPCR expression data, RNA-Seq allows mapping of HERV cDNA transcripts to their chromosomal locus, a recent paper researching changes in HERV expression in breast cancer managed to identify the majority of transcripts being expressed as coming from proviruses contained in introns from chromosomes 9, 10, 12, 14 and 19 (Montesion *et al.*, 2017). An additional study looking at HERV-H and HERV-W expression in human brain samples found a HERV-W transcript located at chr7q21.2, highly expressed across 3 brain regions as well as increased general HERV transcription in bi-polar and schizophrenic patients (Li *et al.*, 2019). RNA-Seq is also able to identify different Repeat elements in generated NGS data, with Long Interspersed Nuclear Elements, Short Interspersed Nuclear Elements and Long Terminal Repeat (including endogenous retroviruses) which was observed in blood and skin samples from Parkinson's patients (Billingsley *et al.*, 2019). These studies show the value of using RNA-Seq for researching the expression of endogenous retroviruses that are present in high copy number and have repetitive nucleotide sequences that are GC rich.

The use of the technology has also been suggested for use in ALS to track biomarkers in the blood or spinal fluid of patients. In a review paper by Kiaei and Kiaei (2021) the authors noted that RNA-Seq as a tool has great potential to track biomarkers in both spinal fluid and blood, with upwards of 890 differentially expressed genes found in PFN1 mouse models of ALS. In another recent paper which tested a bioinformatics pipeline to detect potential biomarkers, found virus-like sequences in the blood of ALS patients compared with controls (Melnick *et al.*, 2021). This gives some indication that HERVs (as proviral sequences) may be detected by the process as they were not specifically using repeat element databases for their search. Finally a recent paper by Jones *et al.* (2021) using RNA seq analysis found HERV-K3 (HML-6) on chromosome 3 upregulated in the primary motor cortex of ALS brain tissue compared to controls. This gives evidence that RNA-Seq is already

being implemented as an important tool in the discovery of novel HERV transcripts being differentially expressed in ALS tissue samples.

2.0 Materials & Methods

2.1 Materials

2.1.1 Bacterial Strains, Plasmids and Bacteriological Media

Table 2.1. *E.coli* Strain and Genotype

Strain	Genotype	Supplier
JM109 *E.coli* Competent Cells	*end*A1, *rec*A1, *gyr*A96, *thi*, *hsd*R17 (r_k^-, m_k^+), *rel*A1, *sup*E44, Δ(*lac-pro*AB), [F´ *tra*D36, *pro*AB, *laqI*^{q}ZΔM15]	Promega (USA)

Table 2.2. Culture Media Used for Bacterial Growth

Media	Composition	Supplier
Luria Broth (LB)	10g/L Tryptone, 10g/L NaCl, 5g/L Yeast Extract	Appleton Woods, UK
LB Agar	Luria Broth combined with 15g/L of Agar	Appleton Woods, UK
Super Optimal Broth with Catabolite Repression (SOC) Medium	2% Tryptone, 0.5% Yeast Extract, 10 mM NaCl, 2.5 mM KCl, 10 mM MgCl2, 10 mM MgSO4, and 20 mM Glucose	ThermoFisher, UK
5-Bromo-4-Chloro-3-Indolyl β-D-Galactopyranoside (X-GAL)	Molecular Formula: $C_{14}H_{15}BrClNO_6$	ThermoFisher, UK
Isopropyl β-D-1-Thiogalactopyranoside (IPTG)	Molecular Formula $C_9H_{18}O_5S$	ThermoFisher, UK
Ampicillin (AMP)	Molecular Formula: $C_{16}H_{18}N_3NaO_4S$	ThermoFisher, UK

2.1.2 pGEM-T Easy Vector System

The pGEM-T Easy Vector (Promega, WI, USA) was used for cloning in PCR products for sequencing purposes to check primer specificity as it is a pre-linearized Vector with 3´-T and 5'- A overhangs for ligation using T4 DNA Ligase. The plasmid encodes LacZ gene for blue/white colony selection and M13 primer binding sites bridge sequence insert site with M13 Forward Primer binding at nucleotide 2941 and M13 Reverse Primer binding at nucleotide 161 which can be used to sequence PCR products that have been successfully cloned into the pGEM-T Easy Vector to confirm primer specificity.

2.1.3. Primers Sequences for Reverse Transcription Quantitative Polymerase Chain Reaction (RT-qPCR) and Sanger Sequencing

Table 2.3 Primers Used in RT-qPCR Assays

The table below displays primer information for gene targets used in RT-qPCR assays using SYBR Green Chemistry. Primer information contained below details those primers used to determine reference genes to be used in gene expression assays. Primer sequences are not shown for all gene targets due to intellectual property rights as they were obtained commercially.

Primer Target	Primer Sequences/GenBank Accession Number and Anchor Nucleotide	Amplicon Size	Source, Supplier (Supplier Location)
X-Prolyl Aminopeptidase 1 (XPNPEP1)	Accession Number: NM_001167604 Anchor nucleotide: 2003	112bp	Qiagen (Germany)
Ribosomal Protein L13a (RPL13A)	Accession number: NM_012423 Anchor Nucleotide: 727	≈180bp	Primer Design (UK)
Ubiquitin C (UBC)	Accession number: NM_021009 Anchor Nucleotide: 452	≈150bp	Primer Design (UK)
Tyrosine 3-Monooxygenase/Tryptophan 5-Monooxygenase Activation Protein Zeta (YWHAZ)	Accession number: NM_003406 Anchor Nucleotide: 2585	≈145bp	Primer Design (UK)
Cytochrome C1 (CYC1)	Accession number: NM_001916 Anchor Nucleotide: 929	≈140bp	Primer Design (UK)
Succinate Dehydrogenase Complex Flavoprotein Subunit A (SDHA)	Accession number: NM_004168 Anchor Nucleotide: 1032	≈120bp	Primer Design (UK)
Eukaryotic Translation Initiation Factor 4A2 (EIF4A2)	Accession number: NM_001967 Anchor Nucleotide: 900	≈115bp	Primer Design (UK)
β-Actin (ACTB)	Accession number: NM_001101 Anchor Nucleotide: 1194	≈95bp	Primer Design (UK)
M13 Sequencing Primers	Accession No: X65308 Anchor Nucleotide: 2941	Dependent on Size of Insert	Eurofins (Belgium)

Table 2.4. Known Primer Sequences Used in RT-qPCR Assays

The table below displays primer information for gene targets used in RT-qPCR assays using SYBR Green Chemistry. Included in the primers listed below are Human Endogenous Retrovirus family HERV-K and HERV-W *env* gene region targets used in gene expression assays.

Primer Target	Primer Sequences/GenBank Accession Number and Anchor Nucleotide	Melting Temperature	GC %	Amplicon Size	Source, Supplier (Supplier Location)
HERV-K *gag* (Group-Specific Antigen)	Forward: 5'-AGCAGGTCAGGTGCCTGTAACATT-3'	64.41°C	50	214bp	Li *et.al* (2015), Eurofins (Belgium)
	Reverse: 5'-TGGTGCCGTAGGATTAAGTCTCCT-3'	62.95°C	50		
HERV-K *pol* (Polymerase)	Forward: 5'-TCACATGGAAACAGGCAAAA-3'	56.05°C	40	140bp	Li *et.al* (2015), Eurofins (Belgium)
	Reverse: 5'-AGGTACATGCGTGACATCCA-3'	59.10°C	50		
HERV-K *env* (Envelope)	Forward: 5'-CTGAGGCAATTGCAGGAGTT-3'	58.46°C	50	164bp	Li *et.al* (2015), Eurofins (Belgium)
	Reverse: 5'-GCTGTCTCTTCGGAGCTGTT-3'	60.04°C	55		
HERV-K *RT* (Reverse Transcriptase)	Forward: 5'-TTCAACCCATGGGGCCTCT-3'	60.48°C	50	182bp	Primer Design Research Work (Thesis Section 3.1), Eurofins (Belgium)
	Reverse: 5'-AAACCTGGTGGCTGGTTCTTT-3'	60.34°C	50		
HERV-W *env* (Envelope)	Forward: 5'-GTATGTCTGATGGGGGTGGAG-3'	59.58°C	57.14	115bp	Lever *et.al.* (2017) Eurofins (Belgium)
	Reverse: 5'-CTAGTCCTTTGTAGGGGCTAGAG-3'	59.11°C	52.17		
Glyceraldehyde 3-Phosphate Dehydrogenase (GAPDH)	Forward: 5'-TGCACCACCAACTGCTTAGC-3'	61.17°C	55	87bp	Li *et.al* (2015), Eurofins (Belgium)
	Reverse: 5'-GGCATGGACTGTGGTCATGAG-3'	61.02°C	57.14		
TAR DNA-Binding Protein 43 (TDP-43)	Forward: 5'-GTACGGGGATGTGATGGATG-3'	57.83°C	55.00	85bp	Douville & Nath (2011), Eurofins (Belgium)
	Reverse: 5'-CTGCGCAATCTGATCATCTG-3'	57.05°C	50.00		
BAF Chromatin Remodelling Complex Subunit BCL11b (BCL11b)	Forward: 5'-AACCCGCAGCACTTGTCC-3'	60.59°C	61.11	189bp	Bartram et al. (2014) Eurofins (Belgium)
	Reverse: 5'-ATTTGACACTGGCCACAGGT-3'				

Table 2.4. (Continued) Known Primer Sequences Used in RT-qPCR Assays

Primer Target	Primer Sequences/GenBank Accession Number and Anchor Nucleotide	Melting Temperature	GC %	Amplicon Size	Source, Supplier (Supplier Location)
HERV-K3 *env*	Forward: 5'-GGTTCTCCCAATAAAGTGGTAATG-3'	57.55	41.67	176	Primer Design Research Work (Thesis Section 2.2.17), Eurofins (Belgium)
	Reverse: 5'-GTGAAAGCTCCCTGCAAATG-3'	57.64	50.00		
HERV-K3 *pol*	Forward: 5'-CTCACATGTTCCTACAGGTTTG-3'	56.98	45.45	82	Primer Design Research Work (Thesis Section 2.2.17), Eurofins (Belgium)
	Reverse: 5'-ACCTCGTGGATTACATCCT-3'	55.06	47.37		
HERV-K22 *pol* (Reverse Transcriptase)	Forward: 5'-GCCGGCCATATAGAACCATCA-3'	60.00	52.38	135	Primer Design Research Work , Eurofins (Belgium)
	Reverse: 5'-TTGAAGGGGCCCCATAGGTT-3'	60.85	55.00		
HERV-H *env* (Envelope)	Forward: 5'- ATCCTTGGCTACCTTCCCCT -3'	59.95	55.00	193	Primer Design Research Work, Eurofins (Belgium)
	Reverse: 5'- GCAGCCGTCAGAGGTTGTAA -3'	60.32	55.00		

2.1.4 Human Post-Mortem Brain Tissue

Frozen Postmortem premotor cortex brain tissue samples (n=40) were obtained from the MRC neurodegenerative disease brain bank, London, UK in 100mg sections, derived from Sporadic Amyotrophic Lateral Sclerosis (sALS) patients (n=20) and non-ALS controls (n=20) that were matched for age and sex as close as possible and used in RT-qPCR assay validation to measure relative expression of HERV-K and HERV-W transcripts in postmortem brain tissue.

Ethical approval to use the post-mortem premotor cortex brain tissue samples was obtained from Westminster University FST Research Ethics Committee (Reference: ETH1718-1476). The Medical Research Council (MRC) neurodegenerative disease brain bank has ethical consent to collect and transfer human tissue samples for research purposes granted by the Wales Research Ethics Committee (REC) (Reference: 08/MRE09/38+5) and REC (Reference: 18/WA/0206).

Additional, postmortem premotor cortex brain tissue samples (used in Garson et al, 2019 paper) were obtained from the MRC neurodegenerative disease brain bank, London, UK derived from Sporadic ALS patients and non-ALS controls and matched for age and sex as close as possible and used in RT-qPCR assays to measure expression of HERV-H, HERV-K22 and HERVK-3 transcripts based on RNA seq findings. Ethical approval to use the post-mortem premotor cortex brain tissue samples was obtained from Westminster University FST Research Ethics Committee (Reference: ETH1819-0060).

Table 2.5 Summary Information for Post-Mortem Premotor Cortex Brain Tissue Samples

Shown in the table below is summary clinical information for post-mortem tissue samples that were used in the HERV-K and HERV-W gene expression studies using RT-qPCR. Information received about patient samples includes ranges of age at time of death, delay in retrieving post-mortem tissue after death and ALS Status. Full data is shown in Supplementary Table 6

Variable	Summary Statistic/Variable	Value
Age at Time of Death (yrs)	Median	70
	Range	43-92
Postmortem Delay (hr)	Median	43
	Range	2-78
Clinical Status	ALS	19 (70% Female, 30% Male)
	Control	20 (70% Female, 30% Male)
Sex	Male	12
	Female	27

Table 2.6 Summary Information for Additional Non-ALS or ALS Associated, Control Post-Mortem Premotor Cortex Tissue Samples

Shown in the table below is summary clinical information for postmortem tissue samples used in research work detailed in the methods section below from the Garson *et.al.* (2019) paper. Full data is shown in Supplementary Table 7

Variable	Summary Statistic	Value
Age at Time of Death (yrs)	Median	75
	Range	55-89
Postmortem Delay (hr)	Median	41
	Range	24-95
Sex	Male	3
	Female	5

Table 2.7 Summary Information for additional Post-Mortem Premotor Cortex Brain Tissue Samples used in Garson et.al. 2019 and not used in initial RT-qPCR assay validation and HERV-K and HERV-W RT-qPCR assays

Shown in the table below is summary clinical information for postmortem brain tissue samples used in RT-qPCR assays to measure relative expression of HERV-H, HERV-K22 and HERV-K3 transcripts. Full data is shown in Supplementary Table 8

Variable	Summary Statistic	Value
Age at Time of Death (yrs)	Median	68
	Range	47-89
Postmortem Delay (hr)	Median	44
	Range	3-98
Clinical Status	ALS	35 (65% Male, 35% Female)
	Control	20 (60% Male, 40% Female)
Sex	Male	35
	Female	20
RIN	Median	6.6
	Range	3.8-8.2

Table 2.8. Summary Information for Post-Mortem Primary Motor Cortex Brain Tissue Samples Used in RNA Sequencing Analysis

The samples listed in the table below were supplied by the MRC neurodegenerative disease brain bank. These samples were sent by our collaborators at KCL to Source Bioscience (UK) for RNA extraction, quantification and sequencing using their Illumina Hi-Seq Next Generation Sequencing (NGS) platform for RNA-seq analysis using ERVMap RNA Seq pipeline and were selected to match patient samples used in our initial RT-qPCR validation and HERV-K and HERV-W RT-qPCR experiments. Full data is shown in Supplementary Table 9.

Variable	Summary Statistic	Value
Age at Time of Death (yrs)	Median	70
	Range	43-90
Postmortem Delay (hr)	Median	39
	Range	2.5-95
Clinical Status	ALS	11 (72% Male, 28% Female)
	Control	14 (71% Female, 29% Male)
Sex	Male	7
	Female	18
RIN	Median	6.5
	Range	4.6-8

Table 2.9. Summary Information for the Publicly Sourced RNA-Seq Peripheral Blood Mononuclear Cell (PBMC) Dataset.

The following table shows a summary of clinical information for RNA-Seq samples obtained from the publicly available PBMC dataset published in Zucca *et al.*, 2019 from an Italian cohort. All RNA RIN values were >8.0. Full data is shown in Supplementary Table 10

Variable	Summary Statistic	Value
Age at Time of Death (yrs)	Median	62
	Range	36-86
Sex	Male	11
	Female	11
Clinical Status	ALS	15 (53% Female, 47% Male)
	Control	7 (57% Male, 43% Female)

Table 2.10 Summary Information for the Publicly Sourced RNA-Seq cerebellum and frontal cortex sample dataset (Prudencio *et.al.* 2017).

Full data is shown in Supplementary Table 11.

Variable	Summary Statistic	Value
Age at Time of Death (yrs)	Median	61
	Range	42-82
Postmortem Delay (hr)	Median	12
	Range	2-30
Clinical Status	ALS	18 (61% Female, 39% Male)
	Control	9 (67% Male, 33% Female)
Sex	Male	13
	Female	14

Table 2.11 Summary Information for the Publicly Sourced RNA-Seq medial motor cortex tissue sample dataset obtained from New York Genomic Centre in Partnership with Target ALS

Raw RNA-Seq files and metadata were requested from Target ALS and downloaded from NYGC web portal. Full data is shown in Supplementary Table 12.

Variable	Summary Statistic	Value
Age at Time of Death (yrs)	Median	60
	Range	32-74
Postmortem Delay (hr)	Median	8.6
	Range	3-32
Clinical Status	ALS	34 (68% Male, 32% Female)
	Control	6 (50% Male/Female)
Sex	Male	26
	Female	14
RIN	Median	6.1
	Range	3.4-7.9

Table 2.12 Clinical Information for the Publicly Sourced RNA-Seq lateral motor cortex brain tissue samples obtained from New York Genomic Centre.

Full data is shown in Supplementary Table 13.

Variable	Summary Statistic	Value
Age at Time of Death (yrs)	Median	64
	Range	32-78
Postmortem Delay (hr)	Median	8
	Range	3-28
Clinical Status	ALS	39 (58% Male. 42% Female)
	Control	6 (50% Male/Female)
Sex	Male	26
	Female	19
RIN	Median	6.2
	Range	3.9-7.6

2.2 Methods

2.2.1 In-Silico Design of Primer Sequences for Amplification of HERV-K transcripts.

Full length 5'LTR-*gag-pol-env*-3'LTR HERV-K (HML-2) sequences were obtained from searching GenBank (NCBI, USA), with 23 unique sequences obtained. In addition, 93 HERV-K *gag-pol-env* sequences were also obtained from a paper by Subramanian et al., (2011), which used a bioinformatics-based method for identifying HERV-K sequences in the human genome assembly GRCh37/hg19 (Feb. 2009). This gave a set of 116 HERV-K sequences. These sequences were then aligned using Multiple Sequence Comparison by Log-Expectation (MUSCLE) algorithm (Edgar, 2004) provided as part of the Molecular Evolutionary Genetics Analysis Software v7.0 (MEGA7, Kumar et al., 2016) package and the European Molecular Biology Lab (EMBL, Cambridge, UK) service. Additional short length HERV-K sequences identified in GenBank (NCBI, USA) were also included to aid in the evaluation of new primer targets, which varied in number according to the genomic region i.e.: n=30 sequences for HERV-K *gag*, n=31 for HERV-K *pol* and n=308 sequences for HERV-K *env* region. HERV-K primer sequences from the paper by Li *et.al* (2015), were aligned against known HERV-K *gag-pol-env* sequences from GenBank and Subramanian *et.al.* (2011) in order to determine how well they aligned to HERV-K HML-2 family members. New candidate primer pairs were chosen that targeted HERV-K RT (Reverse Transcriptase) using conserved areas within the genomic region, with primer candidates altered to select for desirable characteristics, such as melting temperature, GC content, primer length and self-complementarity (Dieffenbach, Lowe and Dveksler, 1993). Optimal features for primers included a length between 18 and 24 bases, a GC content of around 50%, and a thermal window of 59-61^{o}C so that the annealing step occurs simultaneously and ensures efficient reaction conditions (Dieffenbach, Lowe and Dveksler, 1993). New primer set(HERV-K RT) identified and provided in Table 2.4 of the materials section.

2.2.2 Extraction of total RNA from Frozen Human Premotor and primary motor Cortex Brain Tissue.

Working in a biological safety cabinet II, and on dry ice, each brain tissue sample (approximately, 50-75 mg) was added to a sterile cryovial and homogenized in 1 ml Qiazol Lysis Reagent (Qiagen, Hilden, Germany). The Tissue was homogenized for 30 seconds

using the TissueRuptor II (Qiagen) with speed setting 8 with a new probe used for each sample. Lysates were placed at room temperature and homogenized for a further 10 seconds at room temperature to ensure complete homogenization and lysis of the tissue. Phenol-chloroform extraction was carried out by adding 1 volume chloroform (200µl) to the lysates and centrifuging at 4^{o}C at 12,000xg for 15 minutes. The clear upper phase was then extracted without disturbing the interphase yielding 400-500ul of lysate and transferred to a new cryovial tube. An equal volume of 70% ethanol was added to the clear separated upper phase lysate and RNA extracted using RNeasy Lipid Tissue Mini Kit (Qiagen) with on-column DNase treatment for removing genomic DNA (gDNA) following the manufacturer's instructions. RNA was eluted using 45µl nuclease free water (ThermoFisher), with an additional 8µl aliquot of RNA put aside for analysis of RNA integrity and yield and placed for long term storage at -80^{o}C.

2.2.3 Quantification and RNA Integrity Number (RIN) Determination of Total RNA Extracted from Post-mortem Brain Tissue Using the Qubit and Bioanalyser.

Extracted total RNA was analysed for RNA integrity using Agilent 2100 Bioanalyzer and RNA6000 Nano Kit (Agilent Technologies, USA). The 12 sample well bioanalyzer chips were used to process 10 patient samples per batch with the 2 remaining wells used for reference RNAs obtained from Invitrogen (Product codes AM6050 & AM7962). Typically, 1.5µl RNA samples were prepared for RIN analysis by diluting 1:1 with nuclease free water, then heated at 70^{o}C for 2 min in a thermal cycler and processed according to the manufacturer's instructions. Following RIN analysis samples were quantified on a Qubit 2.0 Fluorometer using Qubit RNA Broad Range assay (ThermoFisher, UK). Qubit assays were set up according to the manufacturer's instructions, with room temperature Qubit RNA BR Reagent (200x) added to BR Buffer in a 1:200 Reagent to Buffer ratio to create the reaction working solution (i.e. 20ul Reagent to 1980ul BR Buffer) and vortex mixed for 30 seconds. RNA standards/samples were added to the appropriate volume of working solution (Table 2.13) and incubated at room temperature for 3 minutes before readings were taken. At the start of quantification, a known concentration of total RNA (300ng/µl) pooled from multiple donors was used to test assay viability before and after the reading of new reference standards. Following this patient derived RNA samples were assessed for RNA concentration with a target of 40ng/µl difference in readings required for duplicate or

triplicate values. Successful readings for patient samples were recorded with mean concentrations derived from duplicates for use in determination of the amount of total RNA to use for cDNA synthesis.

Table 2.13. Qubit Assay Reaction Volumes
Displayed in the table below are volumes for reagents used in Qubit Broad Range RNA assays.

Reagent	Standards Volume	Sample Volume
Working Solution	190µl	199µl
RNA Standard (Kit)	10µl	
Patient Sample/Pooled RNA		1µl
Total in Each Tube	200µl	200µl

2.2.4 Nanodrop Quantification of total RNA and plasmid DNA

Quantification and initial determination of nucleic acid purity was performed in addition to Qubit assay (section 2.2.3) using NanoDrop 1000 Spectrophotometer (ThermoFisher, UK). Prior to quantification the pedestal was cleaned using 70% ethanol followed by dH_2O and left to dry. A blank measurement was taken by adding 1µl of the nuclease free water that was used for elution of the nucleic acid and placed onto the nanodrop pedestal, and was used to blank the instrument Following the water blank the pedestal surface was cleaned and 1µl of nucleic acid sample added, the NanoDrop lid was closed and the first reading taken. This process was repeated until all samples had been read in duplicate and the values recorded where an A260 reading of 1.0 OD is equivalent to ≈40µg/ml single stranded RNA. For double stranded DNA an A280 reading of 1.0 OD is equivalent to ≈50µg/ml.

2.2.5 cDNA Synthesis using SuperScript III First Strand Synthesis Kit (Invitrogen)

Invitrogen SuperScript™ III Reverse Transcriptase First Strand Synthesis Kit was used to synthesise cDNA from extracted patient total RNA, following the manufacturer's instructions. Routinely, 1µg of extracted patient derived total RNA was used per cDNA reaction in a final 20ul reaction volume containing 10ul of 2x Reverse Transcription (RT) Reaction mix and added to 2µl RT enzyme in 0.2ml MicroAmp capped reaction tube. Reaction mixes were prepared in a UV hood in a separate laboratory to the addition and dilution of Patient total RNA to prevent contamination of the cDNA master mix as well as the patient samples. Prior to addition of the RNA to the cDNA Reaction mix patient total RNA was diluted to 125ng/µl, with 8ul being added to the reaction mix to make the final

cDNA reaction volume of 20µl and the concentration of RNA at 50ng/µl. In addition to the Reverse Transcription reaction mixes 2 controls were prepared, a water control was added for each experiment consisting of H_20 in place of 125ng/ul of RNA and a no-RT control was made with RNA minus the kit reverse transcriptase to test for contaminants.

Following the manufacturer's instructions, cycling conditions for first strand cDNA synthesis included an initial activation temperature of 25^oC for 10 minutes, followed by a cDNA synthesis step of 50^oC for 30 minutes and finishing with a denature step of 85^oC for 5 minutes. Samples were transported to the cDNA laboratory for the addition of 1µl RNaseI (Invitrogen), followed by a 20-minute incubation at 37^oC to remove single stranded RNA from the synthesised cDNA. Synthesised cDNA was then diluted 1:2 with nuclease free water and 4µl-5µl aliquots were pipetted into 0.2ml DNase/RNase free PCR tubes and then stored at -20^oC for future use. RNA was also tested for the presence of contaminating gDNA by SYBR Green RT-qPCR comparing to CT values from cDNA expression assays using the method outlined in section 2.2.6, replacing cDNA with matching concentration of total RNA.

2.2.6 SYBR Green RT-qPCR to measure HERV-K and HERV-W transcripts in post-mortem brain tissue.

Quantitative Real time PCR was performed using Applied Biosystems (USA) QuantStudio5 96 well Thermal Cycler with experiment setup performed using QuantStudio™ Design and Analysis Software v1.4.3 (Life Technologies, CA, USA). Pre-PCR setup steps were performed in separate laboratories to prevent contamination by PCR amplicons and environmental gDNA during reaction plate setup. The qPCR work flow was divided into 3 laboratories, the first for the preparation of the RT-qPCR master mix and addition of the reaction mixture to MicroAmp Optical 96-Well Reaction Plates (Applied Biosystems, USA), second for the addition of cDNA to the RT-qPCR master mix, and the final lab for addition of sealed MicroAmp Optical 96-well reaction plate into the thermal cycler. UV hoods used for preparation of PCR master mixes and reaction plates and addition of cDNA were cleaned using RNase Away and sterilised by a 30-minute UV cycle prior to use.

Typically, 20µl RT-qPCR assays were prepared using 10µl of 2x SYBR Green Fast Mix (Applied Biosystems, USA), 2µl of 25ng/µl patient cDNA with 10µM forward and reverse primer sets

used for amplification of GAPDH, HERV-W *env* and HERV-K *gag, pol, env* & RT genomic regions (supplied by Eurofins, Germany) and 2µl of primer mix used for primers targeted towards XPNPEP1 (Qiagen. Germany) gene transcript which was supplied as a 10x stock concentration. Control reactions were prepared using 2µl nuclease free water (ThermoFisher, UK) in place of cDNA template in the RT-qPCR to serve as non-template control (NTC). The qPCR Reaction conditions were performed using QuantStudio™ Design and Analysis Software v1.4.3 (Life Technologies, CA, USA) using settings SYBR Green Fast Chemistry, the details of which are given in tables 2.14 & 2.15 Individual assay setups are detailed below.

Table 2.14. RT-qPCR Reaction Conditions

The table below displays reaction conditions during the amplification of patient cDNA samples during RT-qPCR assays.

Cycle Step	Temperature and Length	Cycle Count
Initial Activation	Initial increase to temperature at a rate of 2.74°C/second and held at 95°C for 20 Seconds	1
Denature	Increase from 60°C anneal step at 2.74°C/second then 95°C for 1 Second	45
Anneal & Extend	Decrease from denature step at 2.12°C/second and held at 60°C for 20 Seconds	

Table 2.15. RT-qPCR Melt Curve Conditions

Following amplification, a dissociation step was performed to determine the melt curve profile of amplicons generated during cDNA amplification.

Melt Curve Step	Temperature and Length	Cycle Count
Denature	Increase from 60°C anneal step at 2.74°C/second then 95°C for 1 Second	1
Extension	Decrease from denature step at 2.12°C/second and held at 60°C for 20 Seconds	1
Dissociation	Increase from 60°C anneal step at 0.15°C/second, acquiring fluorescence at each temperature increment then 95°C for 1 Second.	1
Hold	Decrease from 95°C at a rate of 1°C/second to 10°C and held until cancelled by user.	1

2.2.7 AmpliTaq Hot Start DNA Polymerase to Produce XPNPEP1 targeted PCR Amplicons for Cloning into pGEM-T Easy Vector for Sequencing of PCR Amplicons.

The XPNPEP1 10x primer mix (Qiagen, Germany) (as described in the Materials 2.1.3) was used to amplify cDNA template using Amplitaq Hot Start DNA PCR Supermix (ThermoFisher UK) to introduce 5'A overhangs into amplicon sequences for insertion into pGEM-T Easy Vector so that the PCR inserts could be sequenced to determine the specificity of the primer sets as the primers were obtained commercially and the sequences were withheld. Typically, 25µl reactions were set up according to manufacturer's instructions using 1x primer mix for XPNPEP1 amplification which were added to separate Hot Start PCR master mixes. The PCRs were run on a Veriti Thermal Cycler (Applied Biosystems, USA) according to thermal cycling conditions detailed in Table 2.16 and PCR amplicons were then run on an agarose gel (methods section 2.2.13) to determine the size of the PCR amplicons generated prior to cloning into pGEM T easy vector.

Table 2.16. AmpliTaq Hot Start Polymerase Reaction Conditions

Cycle Step	Temperature and Length	Cycle Count
Initial Activation	95°C for 15 Minutes	1
Denature	95°C for 30 Seconds	30
Anneal	60°C for 30 Seconds	
Extension	72°C for 1 Minute	
Hold	15°C for ∞	1

2.2.8 Preparation of IPTG XGAL Amp Agar Plates

Luria Broth (LB) Agar was prepared using 16g LB agar added to a 500ml flask followed by 250ml dH_2O and sealing the top tightly. The contents were then mixed by swirling the contents in the flask by hand until the powder had mixed fully with the dH_2O. Once fully mixed the remaining 250ml of dH_2O was added to the mixture, lid sealed and mixed. The lid was then partially unsealed and secured with autoclave tape prior to autoclave sterilisation at 121^0c for 15 minutes.

IPTG X-Gal Amp agar plates were prepared by heating a Duran glass bottle containing 500ml of solid LB agar (that had previously been autoclaved) at 100°C until liquid and followed by cooling until the LB agar reached 50°C. The LB agar was then supplemented with 200µl 50mg/ml X-GAL, 2.5ml 20mg/ml Ampicillin and 500µl 0.5M IPTG making a final concentration of 100µg/µl ampicillin, 0.5mM IPTG and 20µg/ml X-Gal. The IPTG X-Gal Amp LB Agar mixture was then poured aseptically using a Bunsen flame into several sterile bacteriological petri dishes and allowed to set at room temperature. Excess moisture was then removed in a drying oven for 45 minutes, and the LB agar stored at 4^0C for 2 months if not used immediately.

2.2.9 JM109 High competency cell Cloning and Blue-White Colony Selection for Sequencing of XPNPEP1 PCR Amplicons

As the XPNPEP1 primer sets were pooled and obtained commercially from Qiagen (Germany), the PCR amplicons were cleaned up using Monarch PCR & DNA Cleanup Kit (New England Biolabs, USA) following the manufacturer's instructions and cloned into the pGEM-T Easy vector (Promega, USA) so that the PCR amplicons could be sequenced using M13 forward and reverse primer sets to determine that the primers were specific for the target region.

Following PCR clean up and determination of DNA concentration using the nanodrop (section 2.2.4) Ligations were performed using T4 DNA ligase and 10x ligase buffer (ThermoScientific, UK) with a 1:3 plasmid-to-insert ratio and incubated overnight at 4^{o}C, along with a positive control consisting of 2µl of control DNA and a background plasmid control, followed by heat inactivating at 70°C for 5 minutes. The following day JM109 High Competency *E.coli* cells were removed from -80^{o}C storage and left to thaw on wet ice. 2µl of the ligation products were added to separate pre-chilled 15ml centrifuge tubes with 1 tube containing 0.1ng of pUC18 plasmid for determining transformation efficiency. 50µl of JM109 *E.coli* cells were added to each pre-chilled tube and left on wet ice for 20 minutes. The JM109 *E.coli* cells were then heat-shocked at 42^{o}C for 50 seconds, immediately followed by placing the 15ml centrifuge tubes back onto wet ice for cooling for at least 2 minutes. After cooling, 950µl of prewarmed S.O.C. Medium (Invitrogen) was added to 50µl transformed JM109 *E.coli* cells and incubated at 37^{o}C for 1 hour on a shaking incubator at 200rpm and then 100µl competent cells from each transformation reaction were plated onto IPTG X-Gal Amp agar plates (see section 2.2.8) for overnight incubation at 37^{o}C.

Blue/White colony screening was performed on overnight cultures, with light blue/white colonies selected for growth as amplicons inserted into the expression plasmid were less than 250bp in size and did not fully disrupt Galactosidase activity. Six colonies were picked from each transformation plate and streaked onto fresh LB agar plates to obtain pure cultures and incubated overnight at 37^{o}C. The following day the overnight cultures were removed, and a single colony was inoculated into 3ml LB broth supplemented with ampicillin at a final concentration of 100ug/ml and incubated overnight at 37^{o}C on a shaking incubator. The following day plasmid DNA was isolated by centrifuging overnight cultures, lysing bacterial cells and purifying plasmid DNA using QIAprep Spin Miniprep Kit (Qiagen, Germany) following the manufacturer's instructions. Samples were quantified using Nanodrop Spectrophotometer (see section 2.2.4) and the vector inserts were Sequenced using M13 primers for sequencing across pGEM-T insert site using Sanger sequencing (section 2.2.14).

2.2.10 Reference Gene Selection Assay and Analysis (qBase+, NormFinder, BestKeeper & RefFinder)

In order to determine the most stably expressed reference genes in ALS and non-ALS derived post-mortem brain tissue for normalisation of gene expression levels, total RNA was extracted from post-mortem premotor cortex tissue samples (obtained from MRC neurodegenerative disease brain bank) n=5 ALS patients and n=5 non-ALS controls, which were matched as close as possible for gender , Age and RIN values for total RNA that was extracted from these samples. All RNA samples obtained from ALS and controls were reverse transcribed into cDNA using Superscript III RT as outlined in section 2.2.5 and were run in triplicate in the SYBR Green Fast RT-qPCR assay to measure mRNA expression levels of each reference gene in which the primers were supplied by Primer Design (UK) as part of a 6 reference gene panel targeting: RPL13A (Ribosomal Protein L13A), UBC (Ubiquitin C, Polyubiquitin Precursor), YWHAZ (Tyrosine-3-Monooxegenase/Tryptophan-5-Monooxygenase Activation Protein Zeta), CYC1 (Cytochrome C1), EIF4A2 (Eukaryote Translation Factor 4A2), and SDHA (Succinate Dehydrogenase Complex Flavoprotein Subunit A). In addition, a primer mix targeting β-Actin was also obtained from Primer Design (UK), primers targeting GAPDH were obtained from Eurofins, Germany and primers for XPNPEP1, were obtained from Qiagen (Germany). The RT-qPCRs were performed in triplicate in 0.2ml MicroAmp Optical 96-Well Reaction Plates in which 3 primer targets were run per plate including non-template controls for each primer set with GAPDH used as an inter-plate control and thermal cycling conditions performed as described previously in section 2.2.6.

Quantitative PCR data analysis was performed in order to determine the most stably expressed reference gene using qbase+ software, version 3.0 (Biogazelle, Zwijnaarde, Belgium - www.qbaseplus.com) along with the online tool RefFinder (Xie *et al.*, 2012) which combines NormFinder, BestKeeper, GeNorm and ΔCt selection methods. Normfinder relies on a mathematical model for the evaluation of reference gene stability as opposed to the pairwise association measured by the other methods mentioned above (Andersen, Jensen and Ørntoft, 2004). This mathematical model differs from the pairwise methods by combining the inter and intra group variation values of a candidate reference gene, the

mean of the variance between the sample expression values of a candidate reference gene plus the standard deviation of the distribution of these values is then taken as the final stability measure value (Andersen, Jensen and Ørntoft, 2004). BestKeeper compares pairs of reference genes against a stability index, which is generated from its most stably expressed genes, using a root value of their geometric means where the root value is the number of reference genes used for the index (Pfaffl *et al.*, 2004). The genes are then put through pairwise correlation analysis, comparing the difference between their means and then comparing the reference gene to the BestKeeper index value and ranking in order of stability (Pfaffl *et al.*, 2004). The ΔCt method involves viewing the change in Ct (ΔCt) between a pair of reference genes across patient samples, consistently indicating a stably expressed gene as well as less stably expressed genes (Silver *et al.*, 2006). The pairwise means of these comparisons with other genes can then be interrogated by analysing their standard deviation, with stable genes expressing close to 1 (Silver *et al.*, 2006). The final method, GeNorm, is a more complicated pairwise model which starts by finding the single control normalisation error by measuring the fold expression difference between 2 samples when normalised to the first or second reference gene; the Internal control gene stability measure (M) is then measured as the arithmetic mean of all pairwise correlations of a reference gene compared to other candidates (Vandesompele *et al.*, 2002). Stability is ranked from the lowest M value (most stable) to highest M value (least Stable) with a separate calculation for the best number of candidate reference genes to include in the assay, given as value "V" measuring the pairwise variation of reference genes (Vandesompele *et al.*, 2002).

2.2.11 Determining Amplification Efficiency of Primers used in RT-qPCR Assays.

Primer efficiency assays were conducted using RT-qPCR SYBR Green Fast chemistry with cDNA derived from one ALS sample and one non-ALS control sample and used for all primer targets (HERV-K *gag, pol*, RT and *env*, HERV-W *env* as well as HERV-H, HERVK-22 and HERV-K3 transcripts). Extraction of total RNA was obtained from post-mortem brain tissue as described in section 2.2.2 and cDNA synthesis performed using Superscript III RT as

described in section 2.2.5. Following cDNA synthesis, a 6 series dilution of cDNA at 1:4 dilution steps using nuclease free water (ThermoFisher) starting with undiluted cDNA and finishing with a 1:1024 dilution, performed in a 96 well MicroAmp plate, on cDNA from both the ALS and control sample. A maximum of 5 primer targets were used per 96 well plate on cDNA from both ALS and control and the appropriate controls were included such as No Template Control (NTC) reactions. Quantitative PCR analysis was performed using QuantStudio™ Design and Analysis Software v1.4.3 (Life Technologies, CA, USA) and primer efficiency graphs plotted using Microsoft Excel with replicate Ct values under 0.3 standard deviation (SD) used for estimation of primer efficiency. The Ct means for each dilution were plotted against a log transformed value of the dilution factor (i.e. 0.25 for 1:4 dilution transformed to log value in excel using formula =LOG (0.25,10)) with the R^2 and slope values calculated by plotting a linear line on the graph. Efficiency was determined by entering the slope value into the equation "Efficiency (%) = (10^(-1/Slope)-1)x100" (Zhao *et al.*, 2018). Primer efficiency percentages between 90% and 110% were deemed acceptable as these represent an efficient doubling of reaction product with each amplification cycle (Kirschneck *et al.*, 2017)

2.2.12 Statistical Analysis of RT-qPCR Expression Data

Initial estimation of differential expression values were calculated in Microsoft Excel (Microsoft, Washington, USA) using the 2^-(ΔΔCt) method for relative quantification of RT-qPCR data ((Schmittgen and Livak, 2008). For the calculation of the initial ΔCt value the data was normalised against GAPDH and XPNPEP1 reference genes separately to allow for comparison to the results from Li *et al.* (2015). A mean of ΔCt values for control samples was used as a calibrator for the ΔΔCt step.

Following 2^-(ΔΔCt) the Pfaffl method was used for the analysis of RT-qPCR data, with relative quantification (RQ) values calculated in Microsoft Excel (Microsoft, Washington, USA) (Michael W. Pfaffl, 2001). For the Pfaffl method a geometric mean on of the GAPDH and XPNPEP1 RQ values was used for normalisation of expression data as recommended by the reference gene selection process as detailed in section 2.2.10 and the literature (Vandesompele *et al.*, 2002; Wierschke *et al.*, 2010; Dean, Udawela and Scarr, 2016). A calibrator value for each reference gene and gene of interest was used for the initial ΔCt in the equation consisting of a mean of control Ct values.

Statistical analysis was performed using GraphPad Prism version 8.0.0 (GraphPad Software, San Diego, California USA, www.graphpad.com) along with generated graphs. Calculated values from the 2^-(ΔΔCt) and Pfaffl methods for ALS and Control samples were separately added to the software tables as required by the analysis. ALS vs Control and Male vs Female data were analysed using the Mann-Whitney non-parametric t-test with comparisons of independent variables such as Age, Post-mortem delay and RNA integrity values analysed using linear regression. In addition to these statistical tests significant differential expression results were confirmed using IBM SPSS Statistics for Windows, Version 24.0. (Armonk, NY: IBM Corp.) to employ Binomial Regression using Disease Status as the dependent variable and patient metadata as the covariates.

2.2.13 Agarose Gel Electrophoresis for Visualisation of PCR Amplicons

Preparation of 2% Agarose gels was performed using 1g Molecular Biology Grade Agarose powder (Appelton Woods, UK) dissolved in 50ml of 1x Tris-Borate EDTA (TBE) buffer (ThermoFisher, UK). The mix was heated in a microwave until the agarose powder was dissolved and allowed to cool to 50^{o}C before 5ul of SYBR Safe DNA gel stain (ThermoFisher, UK) was added. The agarose solution was then poured into a sealed gel casting tray with the sample well inserted and allowed to set for 20 minutes at room temperature. The sample comb and masking tape was removed, and the gel placed in the gel tank and immersed in 1 x TBE buffer and the DNA samples were mixed with 1 x loading dye and loaded in the relevant wells along with a 100bp DNA ladder (Generuler, ThermoFisher USA) to aid in estimation of band sizes. The gel tank was closed with a lid and the electrodes were placed into the power pack and the samples were run through the gel for 30-60 minutes at 110 volts and then visualised using UV Transilluminator and pictures taken for estimation of band size.

2.2.14 Sanger Sequencing of PCR Amplicons to Determine Primer Specificity

Following RT-qPCR, the PCR amplicons were cleaned up to remove any residual dNTPs or unincorporated primers that may inhibit the sequencing reaction. PCR amplicons were purified using Monarch PCR & DNA Cleanup Kit (New England Biolabs, USA) and following the manufacturer's instructions using 17µl nuclease free water (ThermoFisher, USA) as elution buffer. The purified DNA was quantified using the Nanodrop at 260nm and purified amplicons with DNA falling between 20-40ng/µl were selected for Sanger sequencing in

order to determine that the primers were specific for the desired target region. Typically, 5ul of PCR product was mixed in a 0.5ml sterile Eppendorf with 5ul of 5uM stock of sequencing primer and sent externally to Eurofins (Germany, https://www.eurofins.com/). FASTA sequences obtained from Sanger Sequencing reactions were submitted to NCBI's nucleotide BLAST service for determining sequence specificity to genomic region.

2.2.15 Analysis of RNA-Seq Data Using a Modified ERVMap Protocol

For the UK cohort of post-mortem primary motor cortex tissue samples, obtained from the MRC neurodegenerative disease brain bank, RNA was extracted from n=11 ALS and n=14 Non-ALS control tissue samples, reverse transcribed to cDNA and ran through a short read NGS library preparation prior to Next generation sequencing via the Illumina HiSeq platform provided by Source Bioscience (UK). Additional publicly available RNA-Seq datasets for brain regions were obtained from Prudencio *et.al.* (2017) via NCBI's sequence read archive and lateral & medial motor cortex regions from the NYGC. An additional sample set was obtained from publicly available Peripheral Blood Mononuclear Cell RNA-Seq data from an Italian Cohort and subject to the same pipeline to detect HERV transcription between n=15 ALS and n=7 non-ALS Control samples.

The RNA-Seq data generated from the 150bp paired end reads was interleaved into a single sequencing file per sample and processed using a modified ERVMap protocol on server running Scientific Linux 6.6. Initially sequence files were trimmed of Illumina adapter sequences and filtered for low quality reads utilising the BBDuk.sh script provided by the BBMap_38.73 software package. Following this the latest Human genome assembly (GRCh38) was downloaded from Ensembl.org and Burrows-Wheeler Aligner used to index the genome prior to mapping Illumina reads.

Mapping the RNA-Seq data to the indexed reference genome was accomplished using samtools to generate a .bam file of Burrows-Wheeler aligned RNA-Seq data. The part of this alignment stage is "soft clipping" the sequences to filter those sequences relevant to HERVs and repeat elements using the perl script from the ERVmap process. Following this samtools version 1.9 was used to sort the aligned reads and index their positions prior to quantifying the reads aligned to ERV loci. Bedtools version 2.29.0 was used to generate

count files for the RNA-Seq data using a bed file of ERV locations provided by ERVMap tools. This process provides data on the expression of ERV loci within the sample set.

To provide relative quantification data the sequence reads from endogenous retrovirus mapping needs to be normalised against the reads mapped to cellular genes. This data is generated using STAR version 2.7 utilising the trimmed and filtered RNA-Seq files generated earlier in the protocol. The STAR aligner initially generates a reference genome combining the GRCh38 human genome assembly with an annotation file before aligning RNA-Seq data to the annotated genome. Quality controlled RNA-Seq sample files generated from BBMap earlier in the protocol were aligned to the annotated STAR reference genome to map cellular gene expression. Gene counts were generated using the python computing language htseq-counts script for each sample and combined into a count's matrix using ERVmap perl scripts.

The count matrices for ERV and cellular gene expression were processed in RStudio using DESeq2 to generate differential expression data for ERVs between ALS and non-ALS control samples using cellular gene expression to generate size factors for normalisation of expression. Further analysis of RNA-Seq data was performed in RStudio using data analysis tools available for the R programming language including co-expression analysis for correlation between the expression of TDP-43 and BCL11b where significant ERV expression was detected, this process is summarised in Figure 2.1. To verify detected HERVs had open reading frames, the ExPASy (Swiss Institute of Bioinformatics) translate service was used to identify any potential protein coding regions. If individual protein sequences were detected the protein sequence was analysed by NCBI (USA) protein BLAST to identify whether the protein coded for annotated/known viral or human proteins. Protein sequences were validated for target specificity and functional groups using Simple Module Architecture Research Tool (SMART, http://smart.embl-heidelberg.de/) and HMMER (https://www.ebi.ac.uk/Tools/hmmer/) online tools.

Additional *in silico* analyses were performed on HERV loci identified as being differentially regulated by the DESeq2 algorithm. A manual search of cellular genes within 1Mb up and downstream of the individual HERV locus with the position and names of genes relative to the insertion point recorded in a table. Using this data the genes were analysed using the gene set enrichment analysis software (Mootha *et al.*, 2003; Subramanian *et al.*, 2005) for

sets of genes which were over-represented (enriched) in the log2fold change data for the entire cellular gene set. Following this the enriched genes for the ERVs found to have significantly enriched gene sets were analysed for functional relationships using the Database for Annotation, Visualization and Integrated Discovery (DAVID, Huang, Sherman and Lempicki, 2009b, 2009a).

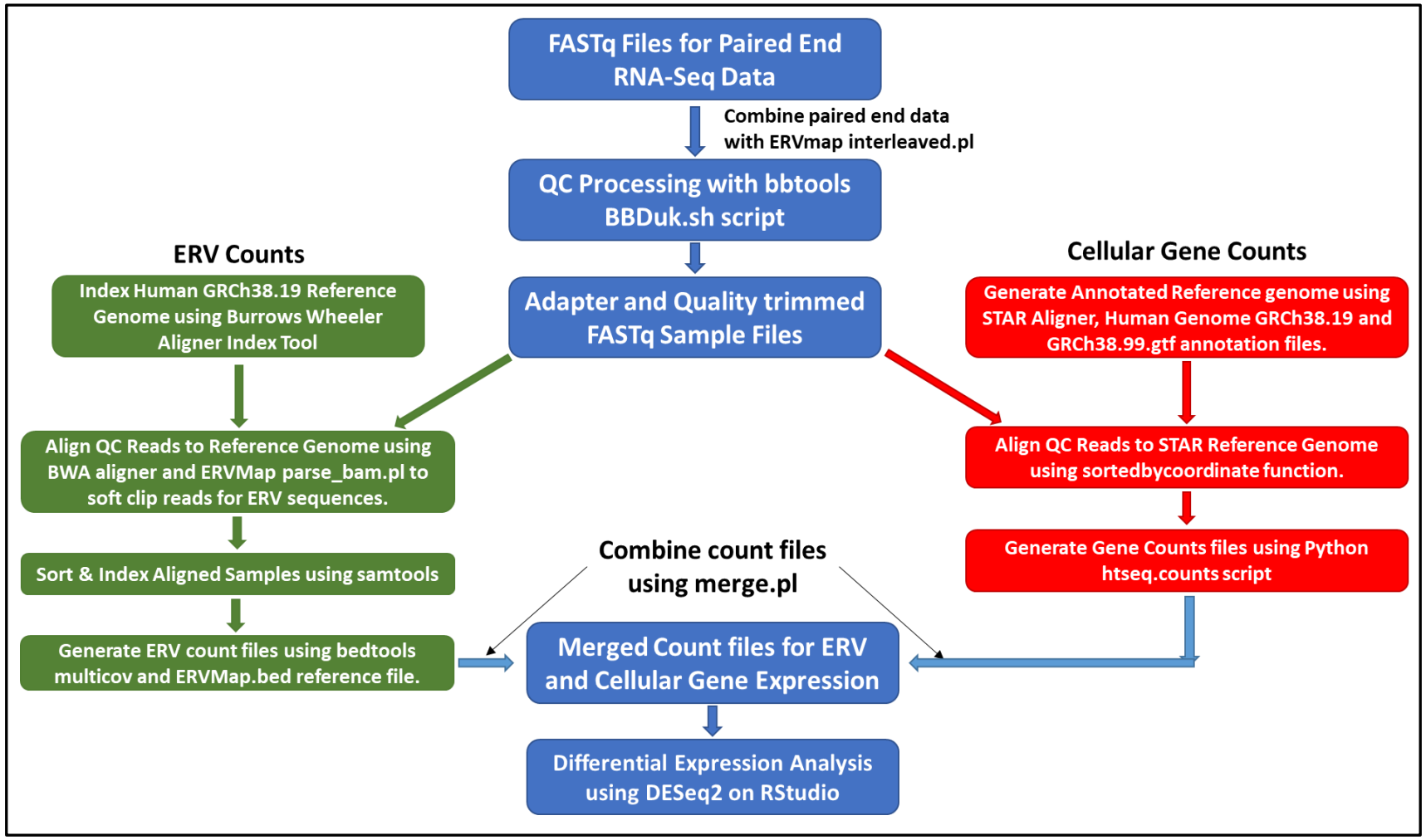


Figure 2.1. RNA-Seq Analysis Flow Chart for Generating ERV Expression Data from FASTq Files using a modified ERVMap Pipeline.

2.2.16 Analysis of Open Reading Frames from Significantly Expressed Endogenous Retroviruses Identified from DESeq2 Differential Expression Analysis

Endogenous retroviruses identified as being significantly expressed from the DESeq2 differential expression analysis were matched with their chromosomal location using the ERVMap.bed file included in the mapping step of the analysis pipeline. This nucleotide sequence location was then entered into the University of California Santa Cruz (UCSC) genome browser available at https://genome-euro.ucsc.edu/ to confirm the annotation for the specific ERV and obtain the nucleotide sequence for the region. The nucleotide sequence was then analysed for open reading frames in Unipro UGENE: a unified bioinformatics toolkit (Unipro, Russia), and compared to a consensus sequence for the ERV obtained from Dfam, a repeat sequence database available at https://www.dfam.org/. These sequences were also aligned against each other using MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms to determine similarity to the consensus sequence. Open reading frames from significantly expressed ERVs were analysed for similarity to human ERV protein sequences by using the National Center for Biotechnology Information (NCBI, USA) protein BLAST search tool. Those ERV open reading frames from significant ERVs which were identified as similar to human sequences had their amino acid sequences transferred to 3D models by SWISS model available at https://swissmodel.expasy.org/ along with their matching open reading frame from the consensus sequence. These 3D models were then aligned using PyMOL version 2.4 (Schrödinger, Inc. NY, USA) to compare the 3D models for similarity in structure and to identify variation between the consensus sequence and the sequence for the ERV from UCSC.

2.2.17. Designing HERV-K3 (HML-6) Primer Sets Targeting Proviral Sequence Present in the human chromosome at Locus 3p21.31c for RT-qPCR analysis

In order to locate unique sequences within HERV-K3 (HML-6) that was identified by RNA seq analysis by our collaborators at KCL to be differentially expressed in postmortem primary motor cortex tissue samples described by Jones *et al.* (2021) at locus 3p21.31c a multiple sequence alignment of multiple HERV-K3 family members from across the genome had to be studied in order to find a unique insertion/deletion in order to selectively detect this HERV at this particular locus. . A separate paper by Pisano *et al.* (2019) had already

performed a comprehensive alignment of HERV-K3 (HML-6) sequences across the genome and annotated differences in the HERV-K3 sequences compared to the consensus sequence. This paper highlighted that there were 2 additional HERV-K3 sequences identified in the 3p21.31 locus, 3p21.31a between chr3:46087646-46095966, 3p21.31b between chr3:46468034-46475121 and the sequence identified by Jones *et.al.* 3p21.31c between chr3:46426676-46433564. In order to target the specific sequence identified by Jones *et.al.* (2021) the forward primers were designed to target insertions or deletions unique to the 3p21.31c sequence. To this end the HERV-K3 *pol* primer set was designed so that its forward primer flanked a 30bp deletion in the target sequence compared to the dfam consensus sequence and the HERV-K3 *env* forward primer flanking the end of a 1.2kb insertion compared to the consensus sequence. These insertion/deletions were unique to the 3p21.31c HERV-K3 sequence which was confirmed using the UCSC in silico PCR tool (https://genome.ucsc.edu/cgi-bin/hgPcr).

2.2.18. Gradient PCR for Determining Optimal Annealing Temperature of HERV-K3 Primer Sets

Gradient PCR was performed using DreamTaq Hot Start DNA PCR Supermix (ThermoFisher UK) in order to determine the optimal annealing temperature of HERV-K3 primer sets. This was performed at 2^{o}C intervals using duplicate 20µl reactions set up utilising a master mix for aliquotting into separate 0.2ml PCR tubes. 10µM primers were diluted to a reaction concentration of 250nM in the master mix. The PCR reactions were run on a Veriti Thermal Cycler (Applied Biosystems, USA) according to thermal cycling conditions detailed in Table 2.17 utilising the device's variflex feature for setting up variant zones of temperature on the heating plate.

Table 2.17. DreamTaq Hot Start Polymerase Reaction Conditions for Gradient PCR

Run Step	Temperature & Time		Cycles
Initial Denaturation	95°C for 3 Minutes		1 Cycle
Denature	95°C for 30 Seconds		40 Cycles
Annealing (Temperature by Zone)	Zone 1 52°C	For 30 Seconds	
	Zone 2 54°C		
	Zone 3 56°C		
	Zone 4 58°C		
	Zone 5 60°C		
	Zone 6 62°C		
Extention	72°C for 1 minute		
Final Extention	72°C for 15 minutes		1 Cycle
Infinite Hold	15 °C for ∞		1 Cycle

The optimal annealing temperature was then determined following gel electrophoresis of the PCR products s following the method detailed in 2.2.13 , based on the size and band intensity of the PCR amplicons generated.

2.2.19 HERV-K3 RT-qPCR Utilising TaqMan Chemistry

As with section 2.2.6; Quantitative Real time PCR utilising the TaqMan chemistry used Applied Biosystems (USA) QuantStudio5 96 well Thermal Cycler with TaqMan experiment setup was performed using QuantStudio™ Design and Analysis Software v1.4.3 (Life Technologies, CA, USA) with quencher set to none due to the HERV-K3 probes utilising a non-fluorescent quencher.

Additionally, as with the SYBR Green chemistry, 20µl RT-qPCR assays were prepared using 10µl of 2x TaqMan Fast Mix (Applied Biosystems, USA) and 2µl of 25ng/µl patient cDNA. Primer set concentration differed to previous primer sets with 10µM forward and reverse primer sets for HERV-K3 *pol* (supplied by Eurofins, Germany) being diluted in master mix to a reaction concentration of 200nM after assessing optimal concentration using TaqMan primer/probe setup according to manufacturers instructions (Table 2.18). The TaqMan Chemistry also utilises a probe for highly specific amplification of a target sequence diluted to a reaction concentration of 250nM (determined following assessment of optimal concentration). Following the method in 2.2.6 the control reactions were prepared using 2µl nuclease free water (ThermoFisher, UK) in place of cDNA template in the TaqMan RT-qPCR to serve as non-template control (NTC). The qPCR Reaction conditions were performed using QuantStudio™ Design and Analysis Software v1.4.3 (Life Technologies, CA,

USA) using settings for TaqMan Fast Chemistry, the details of which are given in Table 2.19 below.

Table 2.18. TaqMan Assay Reaction Volumes

Displayed in the table below are volumes for reagents used in TaqMan Differential Expression assays.

Reagent	Volume for 1 Reaction
2x TaqMan Master Mix	10µl
Forward Primer (10µM)	0.8µl
Reverse Primer (10µM)	0.8µl
TaqMan Probe (10µM)	0.4µl
Nuclease Free Water	6µl

Table 2.19. TaqMan RT-qPCR Reaction Conditions

The table below displays reaction conditions during the amplification of patient cDNA samples during RT-qPCR assays.

Cycle Step	Temperature and Length	Cycle Count
Initial Activation	Initial increase to temperature at a rate of 2.74°C/second and held at 95°C for 20 Seconds	1
Denature	Increase from 60°C anneal step at 2.74°C/second then 95°C for 1 Second	40
Anneal & Extend	Decrease from denature step at 2.12°C/second and held at 60°C for 20 Seconds	
Infinite Hold	Decrease from final Anneal & Extend step at 1°C/second and held at 10°C idenfinitely.	1

3.0 Reference Gene Selection and Validation of Primer Sets to be used in RT-qPCR assays for measurement of relative gene expression of HERV-K and HERV-W *env* transcripts in post-mortem brain tissue and to conform with MIQE guidelines.

3.1 Introduction

Amyotrophic Lateral Sclerosis (ALS) is a fatal disease involving the progressive degeneration of both upper (brain) and lower (spinal cord) motor neurons, starting with an initial focal paralysis and spreading to cover the majority of muscle groups (Valko and Ciesla, 2019). The upper motor neurons are grouped into a region of the brain known as the motor cortex, divided into the premotor & primary cortices and the supplementary motor area (James Knierim, 2018). The affected motor neuron cells include many gene expression changes which include Superoxide Dismutase 1 (SOD1), Tar DNA Binding Protein 43 (TDP-43) and Human Endogenous Retrovirus K (HERV-K) (Siddique and Ajroud-Driss, 2011; Li *et al.*, 2015; Tamaki *et al.*, 2018). These changes in gene expression can be monitored using Reverse Transcription Quantitative Polymerase Chain Reaction (RT-qPCR), which requires normalisation against known stably expressed reference genes for accurate quantification (Bustin *et al.*, 2009).

Reference genes are those related to the maintenance of cellular processes so are assumed to be expressed across both disease and non-disease state samples (Eisenberg and Levanon, 2013). A reference gene candidate for quantifying gene expression should ideally be stably expressed across all samples being studied in disease and non-disease state, regardless of tissue type or experimental conditions (Penna *et al.*, 2011). In practice the expression of reference genes varies between tissue types, with brain tissue showing differences between cortical regions, requiring reference genes to be validated to ensure their suitability (Dean, Udawela and Scarr, 2016). Glyceraldehyde 3-Phosphate Dehydrogenase (GAPDH), is a popular reference gene that has been reported in the literature due to its relatively high level of expression in almost every cell (de Jonge *et al.*, 2007), however, expression levels have been shown to vary between various tissue types such as the brain, skeletal muscle and breast cell lines (Barber *et al.*, 2005; Kozera and Rapacz, 2013). Reference genes can also vary in expression due to the quality of the samples, with factors such as post-mortem delay (PMD), RNA integrity (RIN) and prolonged

pre-mortem stress causing changes in tissue pH, having an observed effect (Harris, Reeves and Phillips, 2009; Koppelkamm *et al.*, 2011; Eisenberg and Levanon, 2013). Normalising data to reference genes is included in the minimum information required for publication of RT-qPCR data (MIQE) guidelines, which suggest using multiple reference genes (Bustin *et al.*, 2009), as using a single reference gene has been observed to cause significant errors in normalisation across samples, leading to incorrect reporting of fold expression changes up to 6-fold difference from true differential expression (Vandesompele *et al.*, 2002; de Kok *et al.*, 2005).

Selecting reference genes for use in RT-qPCR assays can be undertaken using mathematical models for identifying stability values from data generated from these molecular assays. Various algorithms exist for the generation of stability values, the most widely known of which are the geNorm (Vandesompele *et al.*, 2002), NormFinder (Andersen, Jensen and Ørntoft, 2004), BestKeeper (Pfaffl *et al.*, 2004) and ΔCt (Silver *et al.*, 2006) methods. With the exception of NormFinder these methods calculate pairwise comparisons to determine the most stable set of reference genes, generating a stability value for each gene and pair of genes, with geNorm providing an optimal number of reference genes for use (Vandesompele *et al.*, 2002; Pfaffl *et al.*, 2004; Silver *et al.*, 2006). NormFinder uses a complex mathematical model to determine stability values of RT-qPCR data, measuring the mean differences and providing an optimal pair of reference genes alongside its ranked stability values (Andersen, Jensen and Ørntoft, 2004). RefFinder is an additional online tool which uses versions of these applications to provide a basic comparison of all the methods, allowing users to add an additional level of validation to their analysis (Xie *et al.*, 2012).

Primer validation in a qPCR based expression assay requires careful analysis of candidate sequences to ensure efficient amplification of a target gene (Bustin *et al.*, 2009; Kuang *et al.*, 2018; Sreedharan, Kumar and Giridhar, 2018). While some primer sequences can be obtained from previous studies reported in the literature, such as Li *et.al.* (2015) who used different primer sets to measure HERV-K transcript levels in post-mortem brain tissue of ALS patients compared to controls, using *in-silico* methods for primer design allows for prediction of candidate primers and initial estimation of specificity to a gene target as well as their properties such as GC content and secondary structure for example. These can then be assessed by *in-vitro* methods; analysing efficiency of the primers to amplify the intended

target sequence and confirmation of specificity by sequencing (Dieffenbach, Lowe and Dveksler, 1993; Bustin *et al.*, 2009; Svec *et al.*, 2015).

Optimising primers depends on several factors, melting temperature, GC content, length and self-complementarity (Dieffenbach, Lowe and Dveksler, 1993). A primer length of between 18 and 24 bases tends to be highly sequence specific if the reaction conditions are validated to be close to their melting temperatures (Dieffenbach, Lowe and Dveksler, 1993). Primers should also have a GC content of around 50% as this ensures a thermal window for a melting temperature of 54-62^{o}C (Dieffenbach, Lowe and Dveksler, 1993). The melting temperatures of the primers should also be fairly close to one another, so the annealing step occurs simultaneously ensuring efficient PCR amplification reaction conditions (Dieffenbach, Lowe and Dveksler, 1993). Additionally, the 3' complementarity scores are important as these determine whether the primer is likely to form dimers, reducing the amount of available primer and negatively effecting target amplification (Dieffenbach, Lowe and Dveksler, 1993). Identifying primer sets within a targeted region of the gene is improved with the use of *in-silico* selection methods such as NCBI's Primer BLAST and Primer Quest from IDT due to the ability to factor in these criteria. After this *in-silico* selection step additional software tools such as OligoAnalyzer Tool (https://www.idtdna.com/pages/tools/oligoanalyzer) can aid in the evaluation of an optimal primer pair (Dieffenbach, Lowe and Dveksler, 1993; Ye *et al.*, 2012).

Multiple sequence alignment of targeted genomes utilises specialised search algorithms to look for similarities between multiple pairwise comparisons of nucleotide or amino acid sequences including accounting for any potential gaps (Higgins, 1997; Chatzou *et al.*, 2016). Many software packages exist for both aligning and analysing sequence data, of which Molecular Evolutionary Genetics Analysis software MEGA7 is a useful example (Kumar *et al.*, 2016). MEGA7 is packaged with 2 alignment algorithms, which have utility based on the length of and amount of sequences under investigation. One of these, Multiple Sequence Comparison by Log-Expectation (MUSCLE), was reported to have higher accuracy when dealing with high sequence numbers at the cost of computational power when compared to previous alignment algorithms (Edgar, 2004). Alternatively, online tools for multiple sequence alignment (Clustal Omega, T-Coffee) are available from various bioinformatics sites, such as those offered by the European Bioinformatics Institute (EMBL-EBI, Cambridge, UK).

Amplification efficiency is a measure of a primer sets ability to effectively double the concentration of target sequences in a sample in each subsequent PCR cycle (*Guide to Performing Relative Quantitation of Gene Expression Using Real-Time Quantitative PCR*, 2004). This is measured in a reaction by generating a slope value from a relative standard curve of dilutions and plotting the Ct values against a semi-log scale of their dilution factors (*Guide to Performing Relative Quantitation of Gene Expression Using Real-Time Quantitative PCR*, 2004; Svec *et al.*, 2015). Optimal efficiency ranges for the experiment are taken as falling between 90-110%, with values exceeding 100% efficiency due to inhibition of the reaction by enzyme or pipetting errors. In addition to the MIQE compliance a critical factor for using the 2^-ΔΔCt method for calculating differential gene expression is that all primer sets have approximately equal amplification efficiencies (Livak and Schmittgen, 2001; Taylor *et al.*, 2019). Amplification assays also provide a method for determining the dynamic range of the assay, including the limits of detection and quantification (Taylor *et al.*, 2010; Svec *et al.*, 2015).

The research work described in this chapter aims to identify candidate reference genes that are stably expressed for use in normalising RT-qPCR data generated from premotor cortex brain tissue samples derived from ALS and non-ALS specimens obtained at post-mortem, using the mathematical methods described above. A set of 9 reference genes were used to determine the optimal pair of reference genes for normalisation, with primer sets targeted to certain reference genes obtained from Primer Design (UK) and those mentioned in the paper by Li *et al.*, 2015. Primer Design's reference gene set included Ribosomal Protein L13A (**RPL13A**), Ubiquitin C (**UBC**), Polyubiquitin Precursor, Tyrosine-3-Monooxegenase/ Tryptophan-5-Monooxygenase Activation Protein Zeta (**YWHAZ**), Cytochrome C1 (**CYC1**), Subunit of Cytochrome Bc1 Complex, Eukaryote Translation Factor 4A2 (**EIF4A2**), Succinate Dehydrogenase Complex Flavoprotein Subunit A (**SDHA**) involved in the mitochondrial electron transport chain and β-Actin a cytoskeletal protein (**ACTB**). These reference genes have been tested in previous studies though not all have been included in subsequent experiments (Coulson *et al.*, 2008; Rydbirk *et al.*, 2016; Röhn *et al.*, 2018). Additionally, GAPDH and X-Prolyl Aminopeptidase (XPNPEP1) were included in the reference gene panel as they have been used in previous studies involving ALS derived brain tissue and shown to be stably expressed (Durrenberger *et al.*, 2012; Li *et al.*, 2015). Finally, novel and alternative

primers that target different regions of the HERV-K genome to those reported by Li *et al,* 2015 which might be able to capture a larger number of the HERV-K family members will be identified. Multiple sequence alignment tools were used to identify new primer sequences that are specific to HERV-K (HML-2) family members targeting different genomic regions within intact open reading frames. Subsequently *in-vitro* validation of HERV-K and validated reference gene primer sets by relative standard curve method, gel electrophoresis analysis and Sanger sequencing was performed.

3.2 Results

3.2.1 Assessing the presence of gDNA contamination in Patient derived total RNA following DNase I on-column treatment.

Total RNA that was isolated from n=5 ALS and n=5 non-ALS premotor cortex brain tissue samples underwent on column DNase I treatment to remove contaminating genomic DNA. The DNase I treated RNA was used in the first instance to spike the SYBR Green RT-qPCR for each of nine candidate reference genes to ensure that there was no contaminating genomic DNA in the RNA sample due to inefficient DNAse I treatment, as SYBR Green qPCR assays bind to double stranded DNA and also served as a no reverse transcription control. It was only when DNase I treatment of RNA was deemed successful from melt curve analysis and differences in Ct values was cDNA synthesis performed for use in the downstream real-time PCR assays that included the relevant primer sets targeting the 9 reference genes to measure the mean Ct values for each gene present in ALS and non-ALS brain tissue. Displayed in Figures 3.1 and 3.2 are the melt curve plots generated from both cDNA and RNA that was spiked into the qPCR reaction in which the DNase I treatment step of total RNA was efficient in removing the majority of contaminating gDNA from premotor cortex tissue derived total RNA when comparing the melt curves for each template added to the qPCR assay. However, there is also some evidence of low-level amplification for GAPDH (Figure 3.2.C), RPL13A (Figure 3.2.B) and CYC1 (Figure 3.2.E) as shown by the small peaks or shoulders of the melt curves for the RNA spiking qPCRs which appear at the same melting temperature as the cDNA derived qPCR amplicons. Overall, the Ct values generated from the cDNA, Supplementary Table S1, and RNA spike by RT-qPCR are separated by 10 Ct cycle difference representing a 1000-fold difference in detection of expression levels of the different genes by RT-qPCR. Practically this represents a negligible contamination of gDNA in the total RNA which will not negatively impact the assay.

To confirm the melt curve findings, qPCR products that were generated following cDNA amplification and RNA spiking of the qPCR were ran on a 2% agarose gel (Figure 3.3) along with the relevant no template controls (NTCs) in order to visualise if PCR amplicons were generated and if they were of the expected size. The gel electrophoresis data for both cDNA

and RNA derived qPCR products as shown in figure 3.3, are grouped according to the reference gene that was targeted using the relevant primer sets for the ten combined ALS and non-ALS samples. No PCR amplicons were generated for six of the reference genes when RNA was used as the template in the SYBR Green qPCR with the exception of GAPDH, RPL13A and β-Actin in which PCR products are evident for these three reference genes. For the reference gene RPL13A, two different DNA bands can be observed in the gel image, in which one of the bands is migrating at the expected size of a PCR amplicon using the RPL13A primer sets. For β-Actin, a single faint band is visible in the RNA spike gel image, the single band appearing at the expected size of a PCR amplicon using β-Actin primer mix. In the gel image for the GAPDH reference gene two bands can be observed with one appearing at the expected size for the GAPDH amplicon thus confirming small traces of gDNA in the extracted RNA sample or the presence of GAPDH pseudogenes which have similar sequences but are present in the genome at differing lengths (Sun *et al.*, 2012) In the UBC gel image (Figure 3.3F) a faint band matching the size region in the NTC lane can be seen in both the cDNA and RNA gel images, this is likely a primer dimer formed during the reaction process as it is present in the control lane.

The cDNA amplification of reference gene targets by RT-qPCR resulted in PCR amplicons of the expected size for each gene tested (Figure 3.3). This is confirmed in the melt curve plots shown in Figure 3.2 by a single peak appearing at the expected melting temperature. In all instances, agarose gel electrophoresis images showing NTC reactions (Figure 3.3, lanes 11) confirm that the no template controls did not produce an amplicon with only primer dimers evident in the relevant sample lanes.

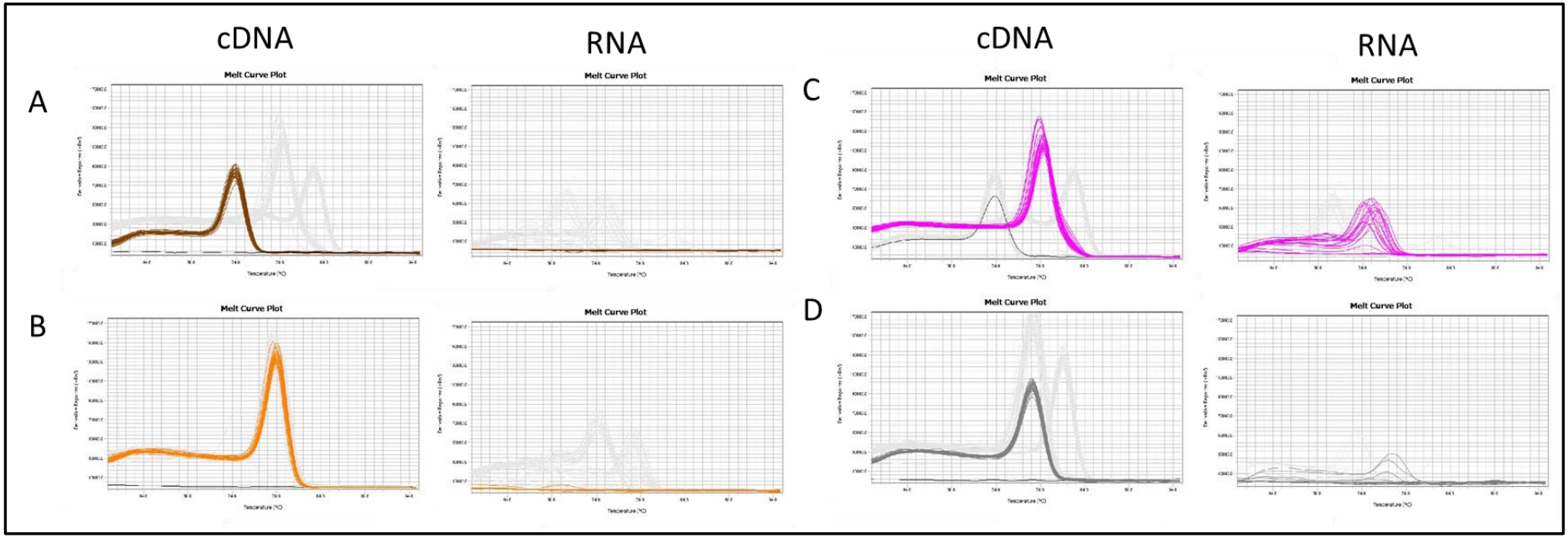


Figure 3.1. Comparison of Melt Curve Plots from cDNA amplification (Left) and RNA spiking amplification (Right) for YWHAZ, XPNPEP1, UBC, and EIF4A2 reference gene target for all 10 ALS and non-ALS combined samples tested.

The graphs displayed in the figure show Melt Curve plots from A) YWHAZ, B) XPNPEP1, C) UBC and D) EIF4A2. Displayed in the left plot for each gene is the melt curve produced following cDNA amplification and on the right is the melt curve plot when RNA is spiked in the PCR assay. Black lines in the reference gene selection plot indicate NTC reactions for those gene targets. The y axis shown in the image shows the Definitive Reporter (-RN) with readings running from 0-174000 with gradients every 20000 units, the x axis shows temperature (Tm) in °C from 60°C -95°C, no template controls amplified at a different temperature (tm) to cDNA generated PCR amplicons or no melt curve was generated.

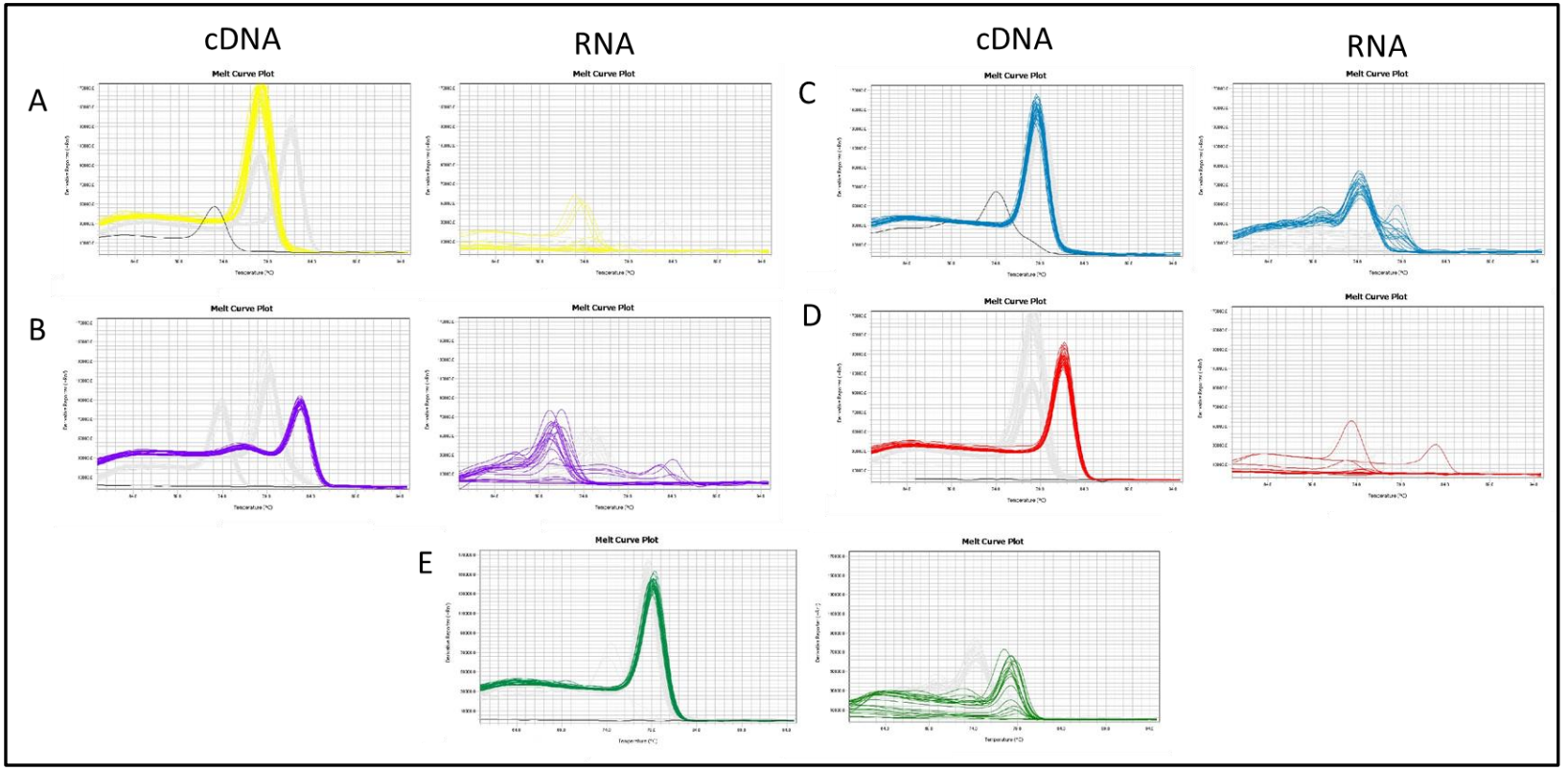


Figure 3.2. Comparison of Melt Curve Plots from cDNA amplification (Left) and RNA spiking amplification (Right) for SDHA, RPL13A, GAPDH, CYC1, and β-Actin reference gene target for all 10 ALS and non-ALS combined samples.

The graphs displayed in the figure show Melt Curve plots from A) SDHA, B) RPL13A, C) GAPDH, D) CYC1 and E) β-Actin. Displayed in the left plot for each gene is the melt curve produced following cDNA amplification and on the right is the melt curve plot when RNA is spiked in the PCR assay. Black lines in the reference gene selection plot indicate NTC reactions for those gene targets. The y axis shown in the image shows the Definitive Reporter (-RN) with readings running from 0-174000 with gradients every 20000 units, the x axis shows temperature (Tm) in ^{o}C from 60^{o}C -95^{o}C, no template controls amplified at a different temperature (tm) to cDNA generated PCR amplicons or no melt curve was generated.

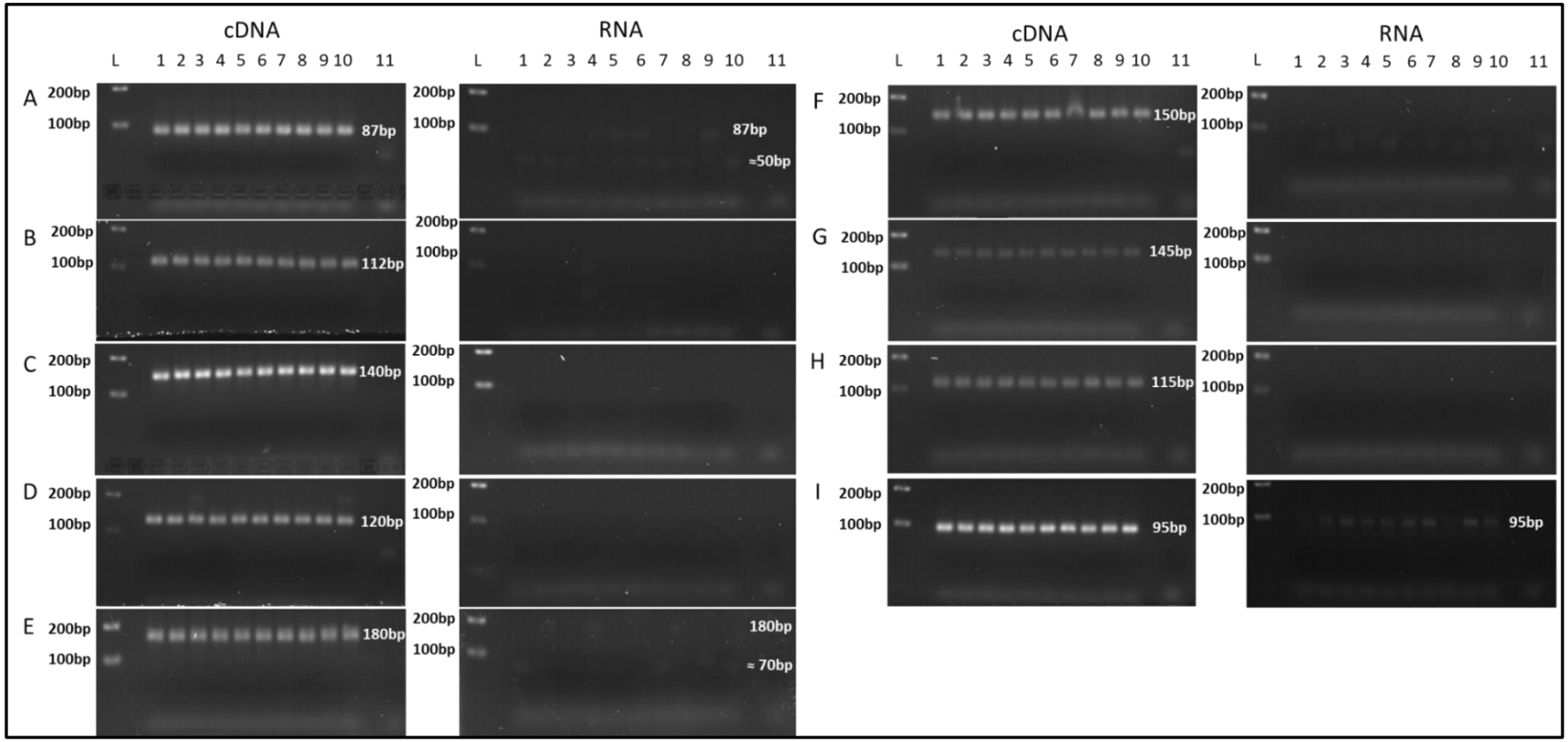


Figure 3.3. Comparison of 2% Agarose Gel Images from cDNA Amplification (Left) and RNA spiking of the qPCR reactions (Right) for each gene target and for each of the samples tested (n=5 ALS and n=5 non-ALS)

The figure above displays agarose gel images of qPCR products from cDNA Amplification (Left) and RNA amplification (Right) for each primer target. Primer targets for the above image are A) GAPDH, B) XPNPEP1, C) CYC1, D) SDHA, E) RPL13A, F) UBC, G) YWHAZ, H) EIF4A2 and I) β-Actin with sample wells corresponding to L) 100bp DNA molecular weight Ladder, bands 100bp and 200bp showing, 1) Sample A292/09 (Control), 2) Sample A151/10 (ALS), 3) Sample A265/08 (Control), 4) Sample A205/09 (ALS), 5) Sample A012/12 (Control), 6) Sample A203/11 (ALS), 7) Sample A346/10 (Control), 8) Sample A401/08 (ALS), 9) Sample A273/12 (Control), 10) Sample A115/08 (ALS) and 11) Primer Target NTC

3.2.2 Analysis of Ct Values Generated from n=5 ALS and n=5 Non-ALS Controls by SYBR Green RT-qPCR

In order to determine the most stably expressed reference genes for normalisation of gene expression, cDNA was synthesised from 1µg of total RNA that was obtained from n=5 ALS and n=5 non-ALS premotor cortex post-mortem brain tissue samples using Superscript III one step RT kit (ThermoFisher Scientific, UK) according to the manufacturer's instructions and incorporating the relevant no -RT controls. RT-qPCR was then performed (see section 2.2.6) using Fast SYBR Green qPCR chemistry supplied by ThermoFisher Scientific (UK) using 25ng cDNA per reaction and a set of 9 reference gene candidates. The RT-qPCR data generated was processed through a simple quality control procedure with any triplicate set of Ct values exceeding 0.5Ct difference removed from the dataset and a mean taken from the remaining duplicate readings. These mean Ct values were used as the basis for the reference gene selection algorithms described in section 2.2.10. The box plot graph displayed in Figure 3.4 shows the distribution of mean Ct values of the 9 candidate reference genes for the ALS and non-ALS derived sample set, providing an overview of the expression levels of these 9 reference genes. Low Ct values indicate a higher level of gene expression for a given sample, with the expression of GAPDH notably higher than the other candidate reference genes, as the mean Ct values range from Ct 17-20. Stability in reference gene expression can also be observed in the box plot with the Ct values of unstable genes occurring over a wider set of values. Both EIF4A2 and YWHAZ show unstable expression across samples with their upper and lower Ct values spread over a greater distance than other candidate genes. The median and interquartile ranges for these genes show a greater distribution of values than the other reference gene targets. The candidate genes show varying levels of Ct values across samples derived from the same anatomical region of the brain, in which the two outliers being YWHAZ and GAPDH are visible in the box plot and correspond to the lowest and highest expressed genes respectively.

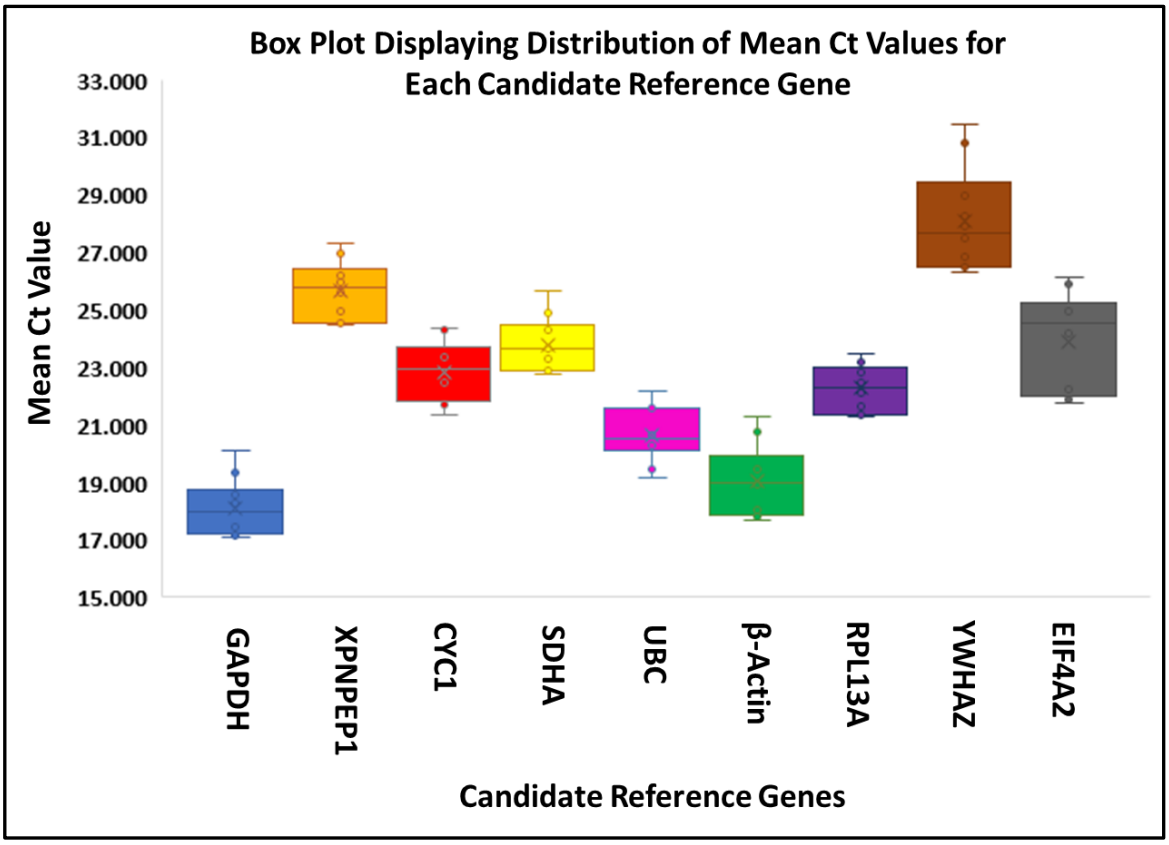


Figure 3.4. Box Plot Graph Showing Distribution of Mean Ct Values for Each ALS and Non-ALS Sample.

The middle line in the boxes indicate the median values for each primer data set that target a specific gene, with the "X" indicating the value of the mean. The Top and bottom whiskers show the maximum and minimum values for each data set respectively with the upper and lower margins of the box representing the interquartile ranges.

3.2.3 Analysis of the Stability of gene expression levels of a panel of reference genes in ALS and non-ALS premotor cortex brain tissue using the geNorm Algorithm.

The geNorm method utilised by the qBase+ software package uses pairwise correlation to calculate 2 separate stability measures for ranking the most stable gene and the optimal number of reference genes to use in a study. The internal control gene stability measure (M) is measured as the arithmetic mean of all pairwise correlations of a reference gene compared to other candidates and is used to generate the optimal number of reference genes for the tissue type to be analysed in the study. Stability is ranked with a separate calculation for the most stable candidate reference gene, given as value "V" measuring the pairwise variation of reference genes.

The mean Ct values obtained using RT-qPCR from n=5 ALS and n=5 non-ALS controls was reorganised into a format compatible for analysis by qBase+ software, version 3.2 (Biogazelle, Zwijnaarde, Belgium - www.qbaseplus.com) and processed using the geNorm algorithm. The initial results for gene stability are given in Figure 3.5, these data have been subjected to quality control as mentioned earlier, where triplicate values with high standard deviation are processed into duplicate values for higher accuracy of the results. Figure 3.5 shows the algorithm selecting XPNPEP1, GAPDH and SDHA as the 3 most stable reference genes with EIF4A2 and YWHAZ shown as the 2 least stably expressed. This data correlates well with the data shown above in Figure 3.4 that illustrates the Ct value distribution for each reference gene and both EIF4A2 and YWHAZ are detected at higher Ct values compared to the other reference genes.

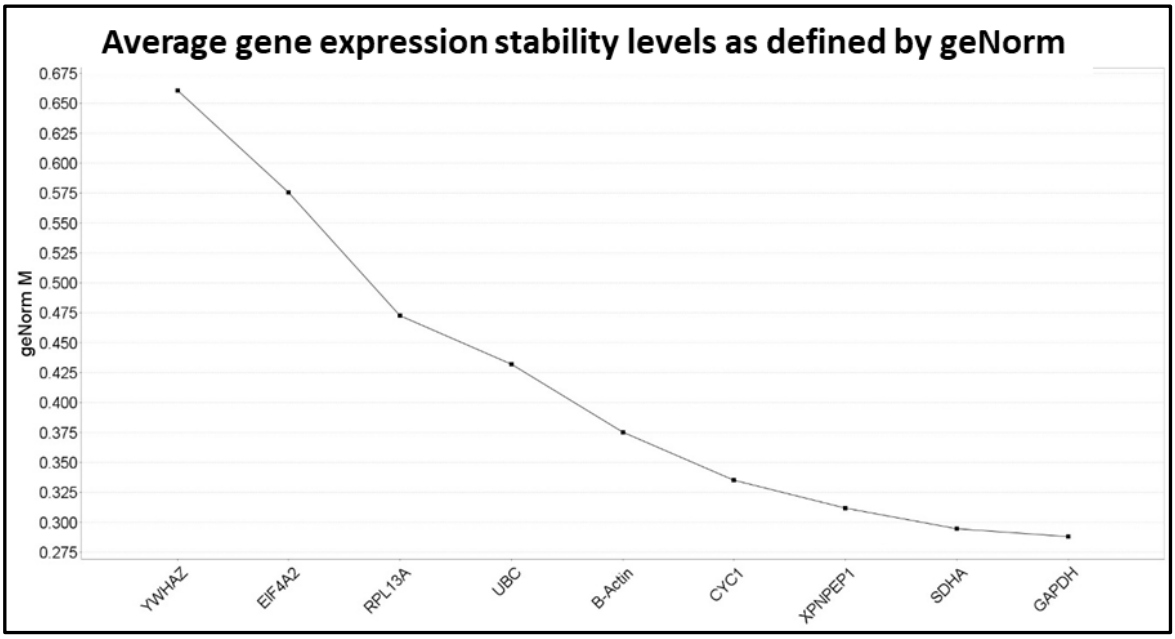


Figure 3.5. Average gene expression stability levels in premotor cortex of ALS and non-ALS cases as defined by geNorm.

The geNorm algorithm generated M Values shown in this graph relate to the pairwise stability of individual reference gene targets and are displayed left to right as the least stable (High M Value) to most stable (Low M Value) genes identified by the geNorm Algorithm.

3.2.4 Identifying the Optimal Number of Stable Reference Genes using the geNorm algorithm.

The geNorm V graph displayed in Figure 3.6 shows consecutive numbers of reference genes as candidate configurations for normalisation. As these appear under the default limit of the selection program, 0.15, qBase+ indicated that 2 reference genes would be an optimal number, being the lowest value of the available combinations. Using the information given in Figure 3.5 this would mean that GAPDH and SDHA would be the optimal reference genes.

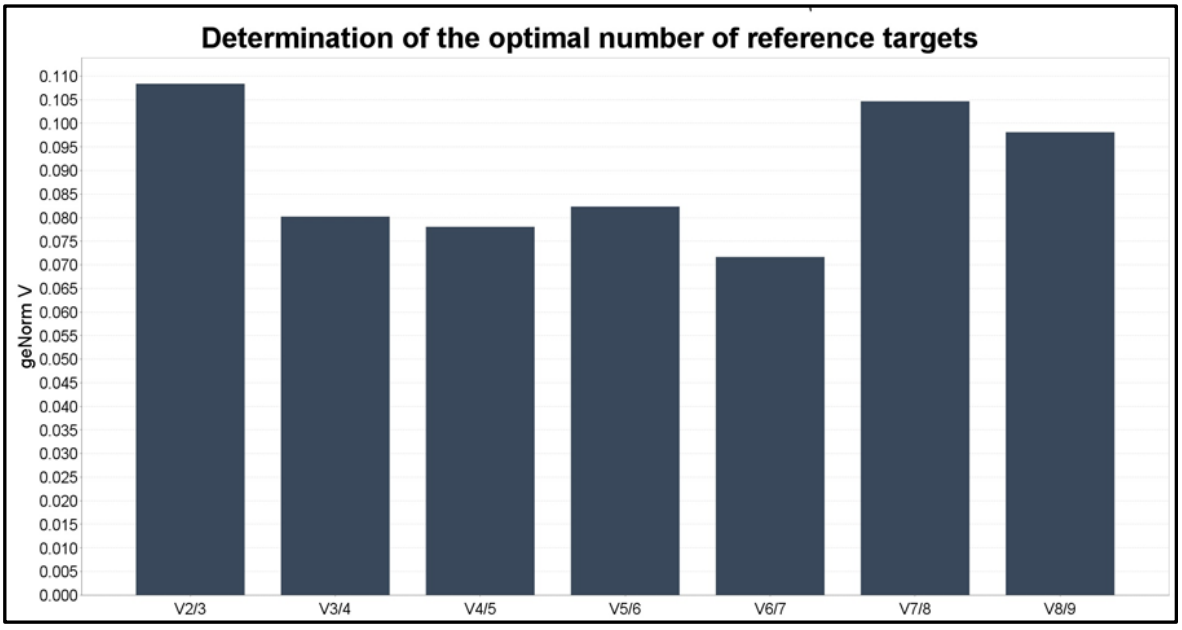


Figure 3.6. qBase+ generated geNorm V graph for selection of optimum number of reference genes to be used for normalisation of gene expression.

The data displayed in this graph gives a "V" value for the comparison of each consecutive number of reference genes to potentially include in a study. Lower V values indicate a benefit for using that number of reference genes as an optimal set, though the true benefit of adding more reference genes for comparison drops under the default cut-off value of 0.15.

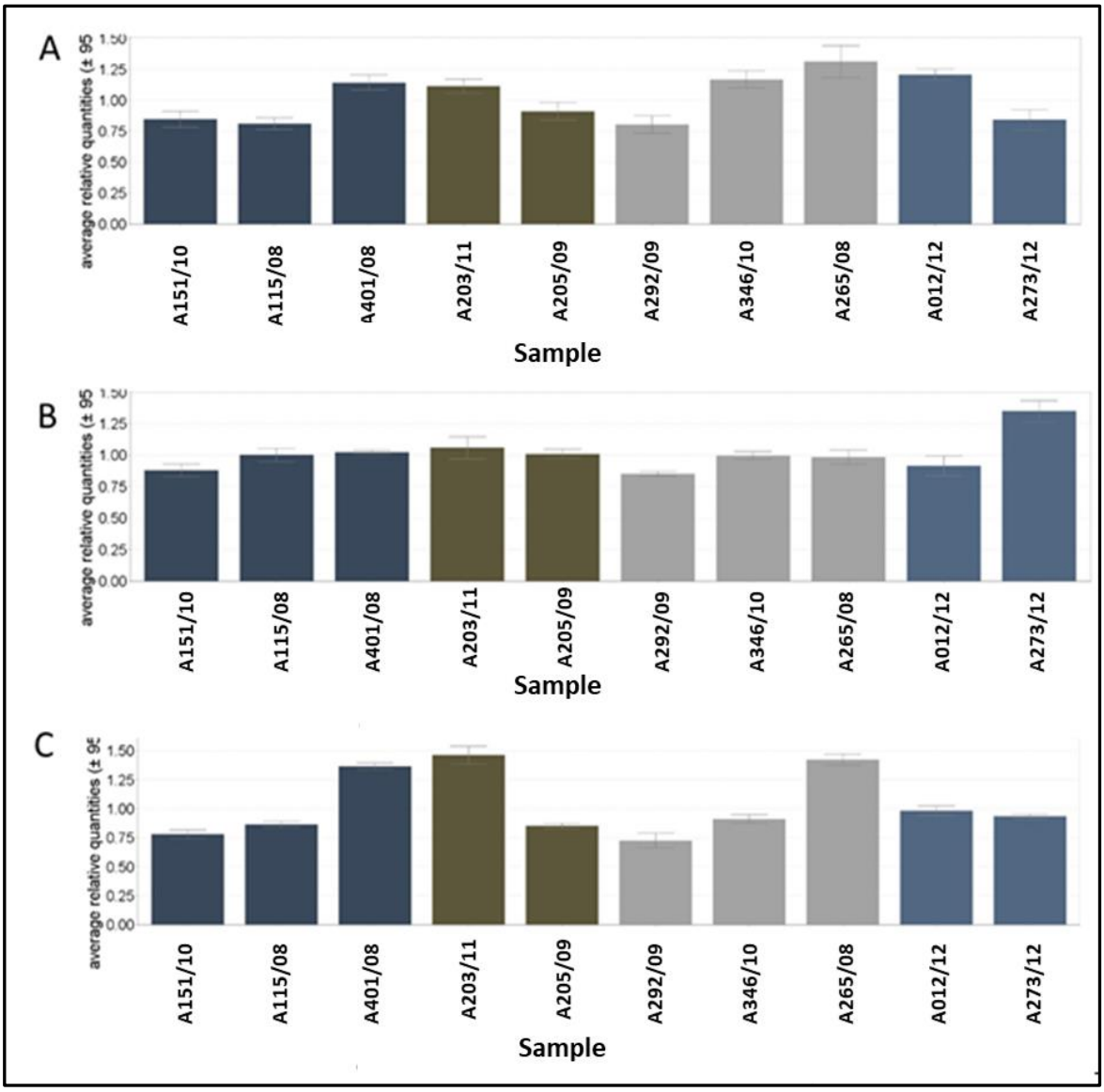


Figure 3.7. Relative quantities of GAPDH, XPNPEP1 and SDHA genes in the premotor cortex from n=5 ALS and n=5 non-ALS cases.

Bars in the above graphs are shown with error bars indicating a 95% confidence interval for each patient sample data set. Primer targets displayed in the figure above are: A) GAPDH B) XPNPEP1 C) SDHA. The sample data on the graph is split into ALS and non-ALS controls with the 5 ALS samples A151/10, A115/08, A401/08, A203/11 and A205/09 to the left of the graphs, and the 5 Control samples A292/09, A346/10, A265/08, A012/12 and A273/12 to the right of the graphs. The data is further split into male and female samples with the colours separating out the 3 female and 2 male samples in the ALS and non-ALS control data.

In Figure 3.7 we can see the relative average quantities of the 3 most stable reference genes as was defined previously in Figure 3.5. Qualitative analysis of this data indicates that SDHA is less stably expressed than GAPDH and XPNPEP1, shown in the variation of SDHA's relative quantities in the different premotor cortex brain tissue samples derived from ALS and non-ALS cases. This can be observed in Figure 3.7C with the average relative quantities of SDHA for each sample showing a wider distribution as the error bars are not as close together as those observed for GAPDH and XPNPEP1 genes. Based on the expression data given in Figure 3.7, XPNPEP1 and GAPDH exhibit higher stability values across ALS and non-ALS Control samples and are more ideal candidates for normalisation of RT-qPCR data in a future expression studies measuring gene expression levels in premotor cortex brain tissue. Additionally, we can observe that gender has no effect on the average relative quantities of each of these genes.

3.2.5 RefFinder Comparison of geNorm, NormFinder, BestKeeper and ΔCt Reference Gene Selection Algorithms using RT-qPCR data from n=5 ALS and n=5 non-ALS control Samples.

RefFinder is an online tool for comparing the most commonly used reference gene selection algorithms, using website-based versions of the geNorm, NormFinder, BestKeeper and ΔCt reference gene selection methods. Table 3.1 shows the combined output of this ranking method, which displays each gene from the most stable to the least stable gene. The RefFinder version of geNorm shows SDHA and GAPDH as the 2 most stable reference genes and suggesting them as a pair to be used, appearing in the same position in the table ranking. GAPDH and XPNPEP1 appear in the top 3 reference genes for the ΔCt and NormFinder selection methods, with the principle difference being that CYC1 takes second place and SDHA is moved down to 5^{th}. BestKeeper shows a discrepancy in rankings when compared to the other methods, with only YWHAZ and EIF4A2 being in similar positions on the table. The Recommended Comprehensive Ranking is provided by the RefFinder web service and is worked out based on the geometric mean of the stability values from all 4 methods, this ranking lists XPNPEP1 and GAPDH as the 2 most stable genes, this compensates slightly for the lower rankings given in the BestKeeper table.

Table 3.1. RefFinder prediction table of most stably expressed reference genes in premotor cortex brain tissue derived from ALS and non-ALS cases.

Data used by the online RefFinder tool are the mean Ct Values for each primer gene target for all samples tested. These mean Ct values were imported into Excel and arranged according to each reference gene and pasted into the web-based program for analysis. The RefFinder program then analyses the mean Ct data using each of geNorm, NormFinder, BestKeeper and ΔCt reference gene selection methods. Also included is a geometric mean based comprehensive ranking based on the rank of each gene for the 4 selection methods.

	Ranking Order (Better--Good--Average)								
Method	1	2	3	4	5	6	7	8	9
Delta CT	XPNPEP1	CYC1	GAPDH	β-Actin	SDHA	UBC	RPL13A	EIF4A2	YWHAZ
BestKeeper	UBC	SDHA	RPL13A	GAPDH	XPNPEP1	CYC1	β-Actin	YWHAZ	EIF4A2
NormFinder	XPNPEP1	CYC1	GAPDH	β-Actin	SDHA	UBC	RPL13A	EIF4A2	YWHAZ
GeNorm	SDHA | GAPDH		XPNPEP1	CYC1	β-Actin	UBC	RPL13A	EIF4A2	YWHAZ
Recommended comprehensive ranking	XPNPEP1	GAPDH	SDHA	CYC1	UBC	β-Actin	RPL13A	EIF4A2	YWHAZ

3.2.6 Validating RefFinder using Original Software Tools for NormFinder and BestKeeper

RefFinder provides an easy access tool for the comparison of reference gene selection methods though it does not provide the full functionality of these programs in their intended format. As with the qBase+ software package the original software for each of the reference gene selection methods was used to confirm the RefFinder predictions listed in table 3.1.

3.2.6.1 Analysis of Ct Values generated by RT-qPCR for different genes using NormFinder Software

NormFinder works on a complex mathematical model for its selection of a stable reference gene, comparing inter and intragroup variation for the calculation of its stability value. One of the main differences with the input methods between RefFinder and the NormFinder Visual Basic Excel Add-on is the ability to assign these groups to the data, along with factoring in primer efficiency values. For NormFinder the groups of Ct means were assigned to non-ALS control and ALS samples for the inter and intra group variation.

While the top 3 stably expressed genes remain the same as RefFinder predicted XPNPEP1 has switched its position with CYC1, with NormFinder stating that the latter is the best gene for the proposed normalisation of gene expression levels measured in the premotor cortex region of the brain with the stability value given as 0.007 (Table 3.2). This is similar to the next most stable reference gene identified, XPNPEP1 which has the same rounded stability value and standard error as CYC1. GAPDH also appears in the top 3 results of the NormFinder algorithm but has a slightly lower (0.008) stability value compared to the to 2 reference gene candidates. As the NormFinder program was also able to analyse the variation between groups it was able to generate a stability value for a pair of reference genes which had a higher combined stability than using one reference gene alone, as shown in Table 3.3. The recommended geometric mean of GAPDH and XPNPEP1 has a higher stability value (0.005) than any single reference gene from the full NormFinder ranking.

Table 3.2. NormFinder stability value for a panel of candidate reference genes.
Stability values of each candidate reference gene tested on premotor cortex brain tissue samples as generated by the NormFinder algorithm are displayed in the table from most (top) to least (bottom) stable with the standard error for each provided.

Gene name	Stability value	Standard error
CYC1	0.007	0.004
XPNPEP1	0.007	0.004
GAPDH	0.008	0.004
SDHA	0.011	0.005
β-Actin	0.012	0.005
YWHAZ	0.014	0.006
UBC	0.015	0.006
RPL13A	0.017	0.006
EIF4A2	0.018	0.008

Table 3.3. NormFinder prediction of the most stable reference gene and the most stable pair of reference genes.
Included in the NormFinder algorithm is the identification of the most stable reference gene and a stability value for the most stable pair of genes, in this case GAPDH & XPNPEP1.

Best gene	CYC1
Stability value	0.007
Best combination of two genes	GAPDH and XPNPEP1
Stability value for best combination of two genes	0.005

3.2.6.2 Analysis of Ct values generated by RT-qPCR for different reference genes using BestKeeper Software.

The BestKeeper selection algorithm uses pairwise comparisons of reference gene stabilities followed by a comparison of these pairs against its own generated stability index. The BestKeeper index value is based on a combined root value of a selection of the most stable genes with the final stability ranking based on the standard deviation of the comparison to this index. In Table 3.4 the BestKeeper rankings are shown along with related data for the samples crossing point values. As shown in the results the BestKeeper program agrees with its RefFinder ranking, with the same standard deviation values. SDHA, UBC and RPL13A, the 3 most stably expressed reference genes according to BestKeeper, with SD values 0.71, 0.71 and 0.72 respectively, appearing in the same ranks as the RefFinder prediction.

Table 3.4. Table of BestKeeper values for reference gene selection.

The standard deviation [± Crossing Point (CP)] values are taken as the stability measures (Highlighted) and arranged most to least stable (SDHA-EIF4A2). Other data generated by BestKeeper is used for generating the highlighted stability values.

	SDHA	UBC	RPL13A	GAPDH	XPNPEP1	CYC1	B-Actin	YWAHAZ	EIF4A2
geo Mean [CP]	23.76	20.63	22.24	18.08	25.64	22.79	19	28.04	23.84
ar Mean [CP]	23.77	20.65	22.26	18.1	25.66	22.82	19.04	28.09	23.9
min [CP]	22.78	19.18	21.26	17.06	24.46	21.35	17.65	26.28	21.74
max [CP]	25.62	22.14	23.48	20.1	27.32	24.37	21.27	31.43	26.11
std dev [± CP]	0.71	0.71	0.72	0.8	0.84	0.98	0.99	1.42	1.53
CV [% CP]	3	3.45	3.21	4.43	3.27	4.3	5.2	5.04	6.42
min [x-fold]	-1.96	-2.73	-1.97	-2.02	-2.27	-2.72	-2.56	-3.39	-4.3
max [x-fold]	3.64	2.85	2.36	4.07	3.19	2.99	4.82	10.49	4.83
std dev [± x-fold]	1.64	1.64	1.64	1.74	1.79	1.97	1.99	2.67	2.9

With the stability rankings from the excel based NormFinder software showing differences to RefFinder it was then necessary to provide an updated geometric mean-based ranking to see if the change in CYC1 stability metric influenced the comprehensive ranking. Table 3.5 shows the relative positions of each reference gene according to the individual selection algorithm. These rankings are expressed in number format and a geometric mean of the values derived for the final "comprehensive" ranking similar to the method employed by RefFinder's comprehensive ranking system. This was calculated in Excel using the geometric mean of the individual ranks from each of the reference gene selection methods, (Table 3.5). As we can see in Table 3.5 the top 4 most stably expressed genes have remained but with their rankings changed. The new ranking system gives GAPDH and XPNPEP1 as the two most stably expressed reference genes, confirming the NormFinder predicted geometric mean pair.

Table 3.5. Revised Geometric Mean ranking of the most stable reference genes based on the NormFinder program ranking.

The ranking position for each reference gene according to the different algorithms are shown along with the geometric mean of these ranks. These values are taken as the comprehensive geometric mean-based ranking.

Gene	NormFinder	BestKeeper	ΔCt	geNorm	GeoMean	GeoMean Ranking
CYC1	1	6	2	4	2.632148	GAPDH
GAPDH	2	4	3	1	2.213364	XPNPEP1
XPNPEP1	3	5	1	3	2.59002	CYC1
SDHA	4	2	5	2	2.990698	SDHA
β-Actin	5	7	4	5	5.143687	UBC
UBC	6	1	6	6	3.833659	B-Actin
YWHAZ	7	8	9	9	8.206694	RPL13A
EIF4A2	8	9	8	8	8.239069	YWHAZ
RPL13A	9	3	7	7	6.031009	EIF4A2

3.2.7 *In silco* Identification of Alternative HERV-K specific Primer sets by Multiple Sequence Alignment of available full-length and partial HERV-K nucleotide sequences.

HERV-K (HML-2) sequences (n=116) obtained from GenBank (NCBI, USA) and additional sequences reported in the paper by Subramanian *et al.*, 2011, were aligned using the MUSCLE algorithm (Edgar, 2004) in the Molecular Evolutionary Genetics Analysis Software v7.0 (MEGA7, Kumar *et al.*, 2016) and the European Molecular Biology Lab (EMBL, Cambridge, UK) CLUSTAL-O (Sievers and Higgins, 2018) service. Additional sequence fragments for each HERV-K genomic region (i.e.: *gag, pol* and *env*) were also obtained from GenBank (NCBI, USA) to aid in the evaluation of primer targets. The alignment of multiple HML-2 family members was performed using the HERV-K *gag*, *pol* and *env* primer sets taken from the Li *et al.* (2015) paper and assessed for suitability by quantifying their matches to available HERV-K partial and full-length nucleotide sequences (Table 3.6). New candidate primer pairs were chosen from multiple HERV-K genomic regions to increase the confidence in selecting primers that are specific to the target region. These primer sequences were assessed for their specificity to HERV-K sequences at various chromosome sites and compared relative to primer sets utilised by Li et al (2015) and the information is displayed in Table 3.6. The data displayed in Table 3.6 indicates that the redesigned primer sets have a higher sequence similarity across the HERV-K family members than the original Li *et.al.* primers reported in 2015. The Li *et al.* (2015) HERV-K gag and pol primers reported a lower number of matches to the HERV-K nucleotide sequences obtained from GenBank and Subramanian *et al*. (2011) resulting in a lower overall percentage match (Table 3.6). This is also displayed in supplementary figures S3-S40 which show the alignment of the primer sequences to the full length ERV and genomic regions. Also highlighted in the supplementary information is the base pairs which differ from the primer sequence in the mismatched full length HERV-K (HML-2) sequences.

Table 3.6. Primer Sequence Matches from Multiple Alignment of Known HERV-K nucleotide sequences.

Information provided in this table gives a comparison between Li *et.al.* (2015) primer sets and the proposed redesigned primers for the different HERV-K genomic regions. HERV-K *gag-pol-env* sequences (n=116) were obtained from Subramanian *et.al.* (2011) and GenBank (NCBI) including short sequences aligning to genomic regions within the HERV-K *gag, pol,* or *env* region (n=31 for *pol*, n=308 for *env* and n=30 for *gag*). Information on HERV-K primer sequences can be found in Materials Table 2.4.

	Primer	HERV-K *gag-pol-env* Sequence Matches	Alignment of primers to either HERV-K *gag, pol* or *env* genomic regions	Overall % match to full length *gag-pol* and *env* seq
New Primers	RT Forward	57/116	13/31	48%
	RT Reverse	55/116	14/31	46%
	GagED Forward	49/116	21/30	48%
	GagED Reverse	41/116	18/30	40%
	Env Forward	54/116	258/308	73%
	Env Reverse	54/116	289/308	81%
Li *et.al.* Primers	Pol Forward	60/116	0/31	41%
	Pol Reverse	35/116	0/31	23%
	Gag Forward	15/116	4/30	13%
	Gag Reverse	36/116	16/30	36%
	Env Forward	48/116	0/308	11%
	Env Reverse	37/116	0/308	9%

Identifying primer targets within the HERV-K genomic region was accomplished by converting the primer sequence to the amino acid sequence and mapping to the amino acid sequence of HERV-K identified open reading frames. Both Li *et.al.* (2015) primers and the new primer sets obtained from multiple sequence alignment amplify a region within the *Pro* (Proline) rich area within the HERV-K *gag* genomic region, however, they have significantly different overall matches with the Li *et.al* primers. The HERV-K *gag* forward primer derived from Li et al (2015), only aligns to 13% of available HERV-K *gag* sequences (Supplementary Images S3-S5) and the reverse primer aligning to 36% of HERV-K *gag* sequences (Supplementary Images S9-S11). By contrast the new HERV-K *gagED reverse and forward* primer sets that I designed, both have 40% or greater matches to the aligned HERV-K sequences (Table 3.6). This is due to the Li *et.al.* primers aligning to particular regions of the HERV-K sequence alignments with single nucleotide mismatches between the primer and HERV-K genome (Supplementary Images S3-S40). Other primer targets such as HERV-K *env* do not appear to have much difference between overall matches to full-length HERV-

K sequences, with overall percentage weight favouring the redesigned primers only due to their location within the short HERV-K *env* sequences included for the multiple sequence alignments (Table 3.6). Both the Li *et.al.* (2015) HERV-K primer sets and the new primers that I designed were then evaluated for optimal primer criteria using the NCBI (USA) primer BLAST tool, the information for which is displayed in the materials section in Table 2.4, in which all primers have a GC content of between 40%-55%.

In order to view and compare all the primer locations within a representative full length HERV-K genome, the HERV-K family member 115 was selected (GenBank Accession AY037929.1) as this provirus contains intact Open Reading Frames (ORFs) for gag, pol and env proteins and intact LTR Regions (Turner *et al.*, 2001). The genome was translated to the amino acid sequence using ExPASy's (SIB, Switzerland) translate tool and protein coding regions identified using Uniprot (Bateman *et al.*, 2017). Figure 3.8 displays the location of these ORFs along with identified protein coding regions. The location of both Li *et.al.* (Figure 3.8, iii, A, B & C) and new primers (Figure 3.8, iii, D, E & F) are shown. Figure 3.8 shows the Li *et.al. pol* primer targets the Integrase (*int*) protein region (Figure 3.8, iii, B); with the new primer set for *pol* designed to target the Reverse Transcriptase (*RT*) protein coding region (Figure 3.8, iii, E) which was a prerequisite since all retroviruses possess RT activity, which is used as a generic marker for retroviruses.

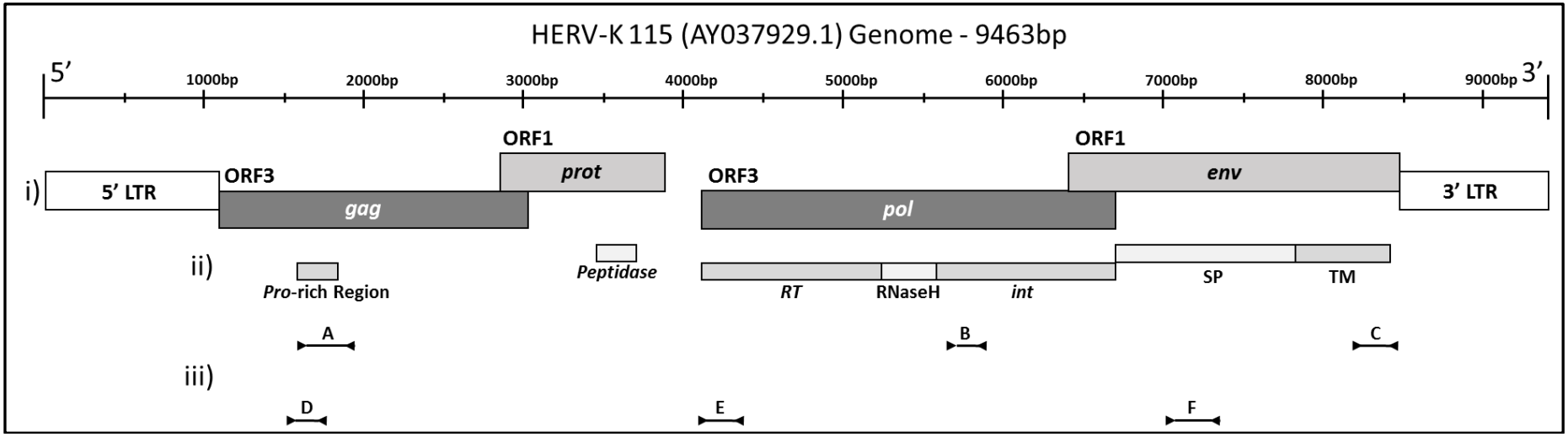


Figure 3.8. Features of HERV-K 115 Representative Genome with Identified Protein Domains and Primer Target Positions.
Displayed in the figure above is an annotated representation of the HERV-K 115 genome (Accession: AY036929.1). Annotations were achieved by identifying each of the protein coding regions. The 3 layers of information in the figure are (i) Information about primary regions including *gag* (group specific antigen), *prot* (protease), *pol* (polymerase) and *env* (envelope) within the genome along with Long Terminal Repeat (LTR) and the relevant Open Reading Frames (ORF) identified using BioEdit (v7.0.5, Ibis Therapeutics, USA) and ExPASy (SIB, Switzerland) translation, (ii) Additional features and protein regions of polyproteins identified in "i", with abbreviations *RT* (Reverse Transcriptase), *int* (Integrase), SP (Surface Protein) and TM (Transmembrane Protein) identified by Uniprot (Bateman *et al.*, 2017) and (iii) the amplicon locations for Li *et.al.* (2015) ((A) HERV-K *gag*, (B) HERV-K *pol* & (C) HERV-K *env*) and new primer sets ((D) HERV-K *gagED*, (E) HERV-K *RT* & (F) HERV-K *env*) for the genomic regions.

3.2.8 Amplification Efficiency of HERV-K, HERV-W *env*, XPNPEP1 and GAPDH primer sets by Standard Curve as well as primers targeting TDP-43 and BCL11b transcriptional regulators.

In order to determine if the new HERV-K primers that I designed and those identified in the Li *et al*. (2015) paper were suitable for use in the SYBR Green-based qPCR assay, a standard curve method was used to determine amplification efficiency for each of the primer sets. A primer set for HERV-W *env* (Levet *et al.*, 2017) was selected from the literature as HERV-W expression has been associated with multiple sclerosis (MS) (Morandi *et al.*, 2017). Additionally, primer sets for BCL11b and TDP-43 were tested for amplification efficiency, this is due to their observed transcriptional modification of Retroviruses such as HIV and HERV-K in the central nervous system. Based on the reference gene selection work GAPDH and XPNPEP1 were also included in this stage of the validation procedure having been identified as stably expressed in both ALS and non-ALS postmortem premotor cortex tissue samples. Initially, cDNA was generated from 1ug total RNA from n=1 ALS (A151/10) and 1= non-ALS control sample (A292/09) using ThermoFisher Scientific (UK) Superscript III one - step RT assay kit. To generate a standard curve undiluted cDNA was diluted 1:4 over a 6 step dilution series for qPCR assays as detailed in the methods section 2.2.11. Primer specificity for a single amplicon was determined by performing melt curve analysis on RT-qPCR products that were generated using the different primer sets as shown in Figure 3.9 and 3.10.

The HERV-K *gagED* primer sets that I designed were not included in downstream RT-qPCR assays as they were non-specific to the *HERV-K encoding gag* regions as well as the HERV-K *env* primers that I also designed based upon the total number of HERV-K env sequences they aligned against, which was not larger than the number of sequences that aligned to HERV-K env primer sets used by Li *et.al.* (2015). The melt curves shown in Figure 3.9 and 3.10 show a similar pattern for the ALS and non-ALS sample for each of the HERV-K primer sets used in the RT-qPCR step, along with the primer sets for TDP-43, BCL11b and the reference genes GAPDH & XPNPEP1, and indicate that a single amplicon was generated using these primer sets and the size of the amplicon was of the expected size as determined by agarose gel electrophoresis as shown in Figure 3.11 & 3.12. For some higher dilutions

such as 1:1024 dilution of the cDNA for the control sample, both the HERV-K *gag* and HERV-K *RT* primer sets did not show a visible band in the gel image and can be attributed to the effect of diluting the sample too much as well as the efficiency of the primers to amplify the target region. As predicted, a similar trend can be seen in other gel images that are related to DNA bands appearing more faintly with increasing cDNA dilution. The HERV-W env primer set failed to produce a single peak at the expected Tm and two DNA bands migrating at different distances in the agarose gel are evident (Figure 3.11G & 3.12G). Only primer dimers are visible in the no template controls (NTC).

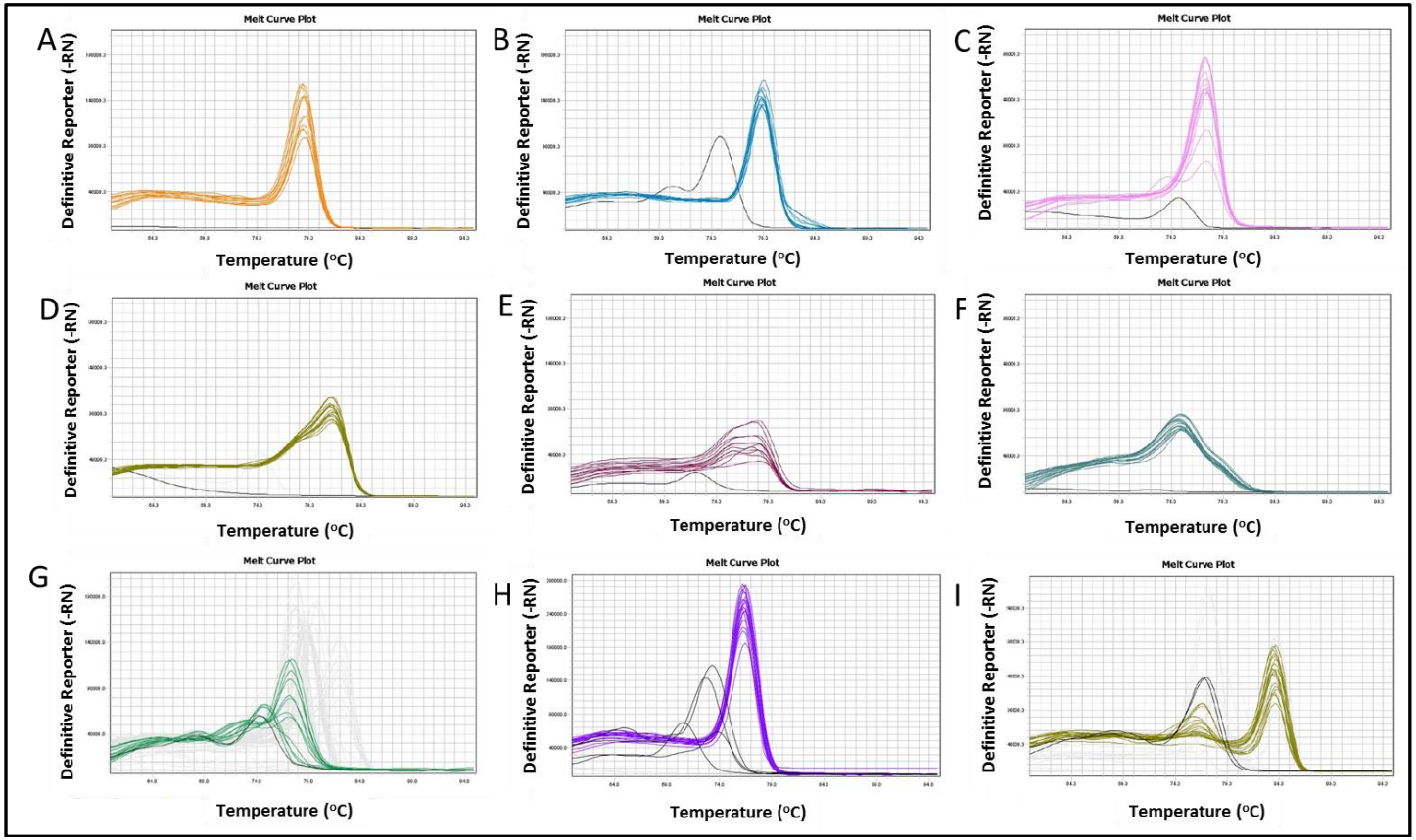


Figure 3.9. cDNA Melt Curve Plots for HERV-K, HERV-W *env*, TDP-43 and BCL11b Primer Targets in an ALS Sample.

The graphs displayed in the figure show Melt Curve plots from A) XPNPEP1, B) GAPDH, C) HERV-K *env*, D) HERV-K *gag*, E) HERV-K *pol*, F) HERV-K *RT*, G) HERV-W env, H) TDP-43 and I) BCL11b. Black lines in the reference gene selection plot indicate NTC reactions for those gene targets. The y axis shown in the image shows the Definitive Reporter (-RN) with readings running from 0-216000 with gradients every 20000 units, the x axis shows temperature (Tm) in °C from 60°C -95°C, no template control amplified at different temperature (tm) to the target transcript or showed no melt curve.

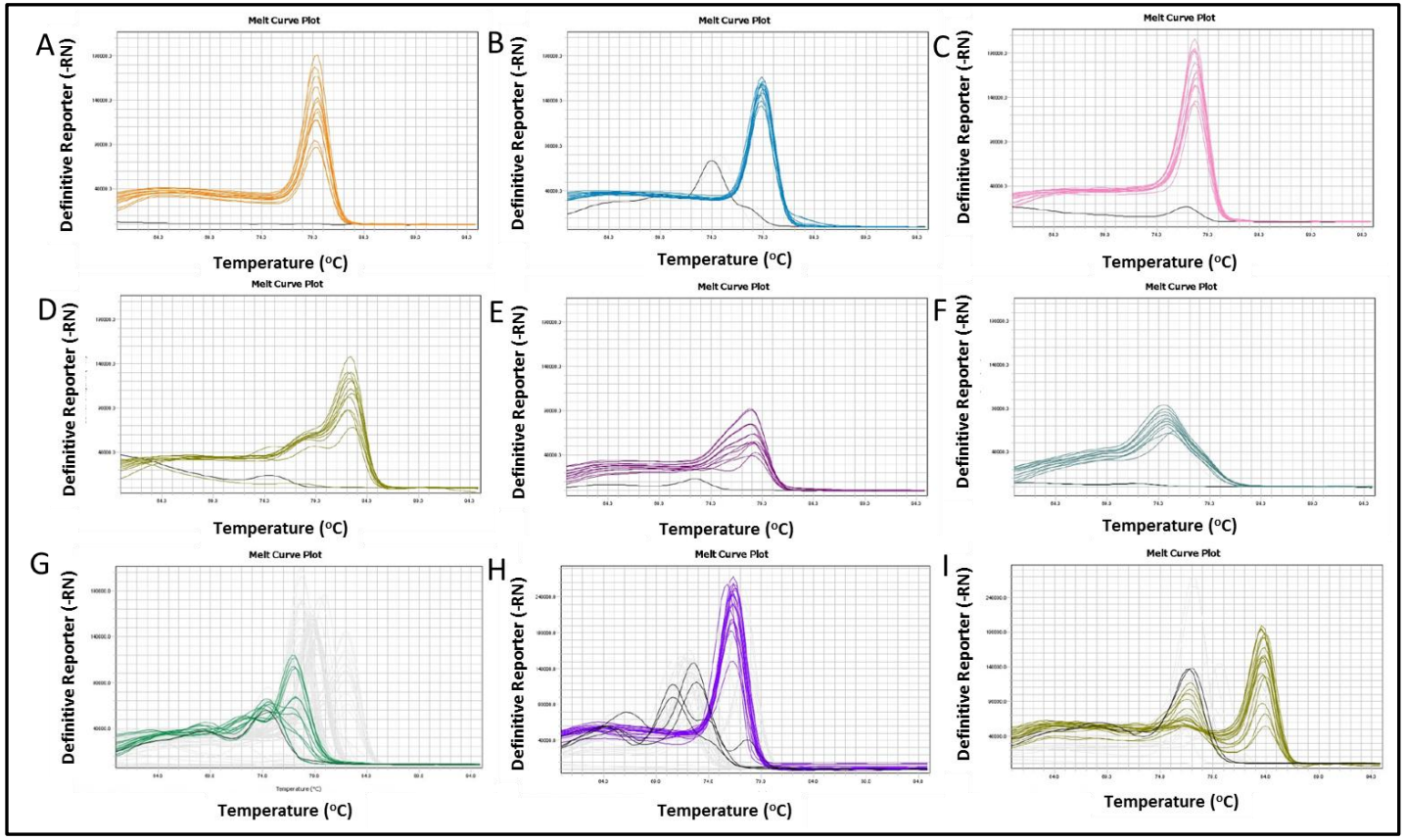


Figure 3.10. cDNA Melt Curve Plots for HERV-K, HERV-W, TDP-43 and BCL11b Primer Targets in a non-ALS Control Sample.

The graphs displayed in the figure show Melt Curve plots from A) XPNPEP1, B) GAPDH, C) HERV-K *env*, D) HERV-K *gag*, E) HERV-K *pol*, F) HERV-K F) HERV-K *RT*, G) HERV-W env, H) TDP-43 and I) BCL11b. Black lines in the reference gene selection plot indicate NTC reactions for those gene targets. The y axis shown in the image shows the Definitive Reporter (-RN) with readings running from 0-216000 with gradients every 20000 units, the x axis shows temperature (Tm) in °C from 60°C -95°C, no template control amplified at different temperature (tm) to the target transcript or showed no melt curve.

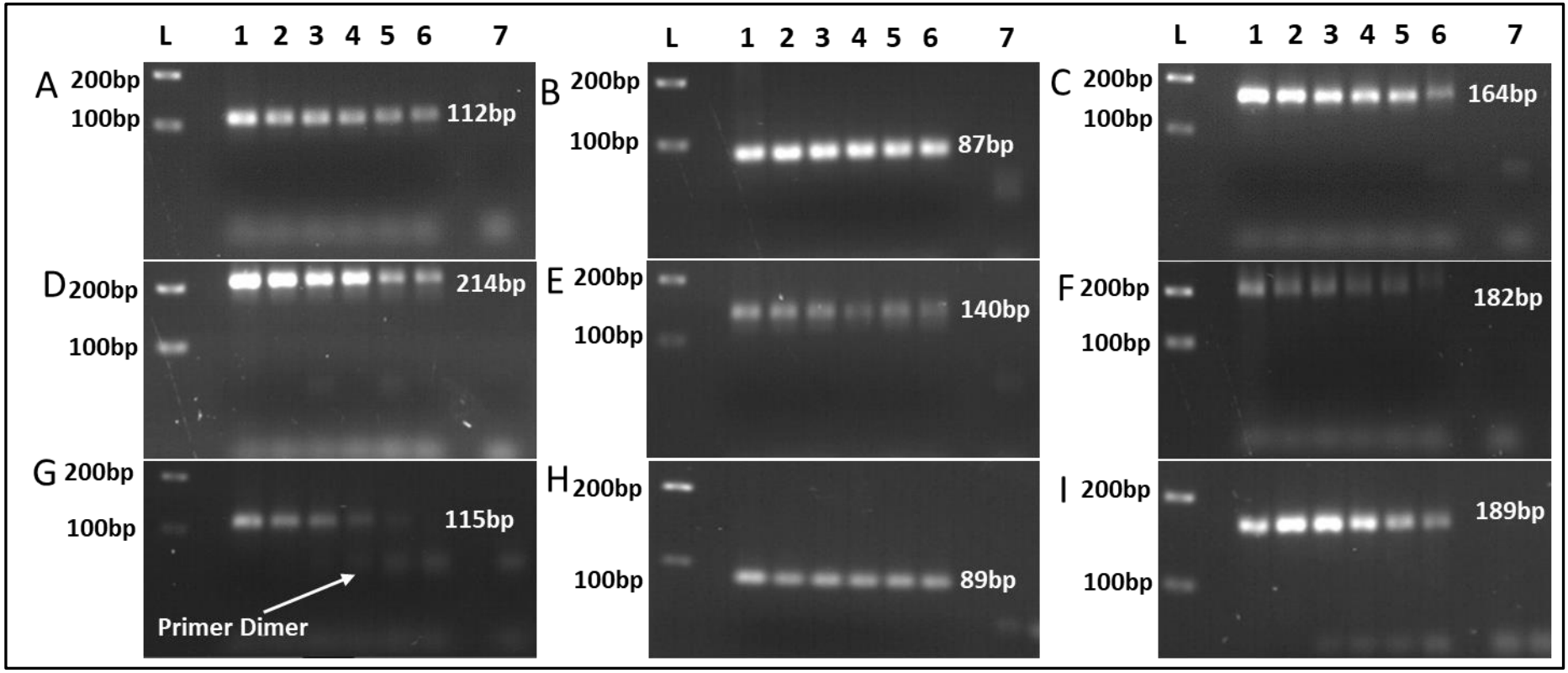


Figure 3.11. Agarose Gel Electrophoresis Results for Amplicons Generated from cDNA Amplification of Gene Target Transcripts in ALS cDNA Derived Samples

The figure above displays agarose gel images of RT-qPCR amplicons that were generated following cDNA Amplification of ALS Patient Sample A151/10 for each primer target. Primer targets for the above image are A) XPNPEP1, B) GAPDH, C) HERV-K *env*, D) HERV-K *gag*, E) HERV-K *pol*, F) HERV-K *RT*, G) HERV-W env, H) TDP-43 and I) BCL11b with sample wells corresponding to L) 100bp Ladder, DNA bands corresponding to 100bp and 200bp are shown, 1) Undiluted cDNA, 2) cDNA Diluted to 1:4, 3) cDNA Diluted to 1:16, 4) cDNA Diluted to 1:64, 5) cDNA Diluted to 1:256, 6) cDNA Diluted to 1:1024 and 7) Water Control.

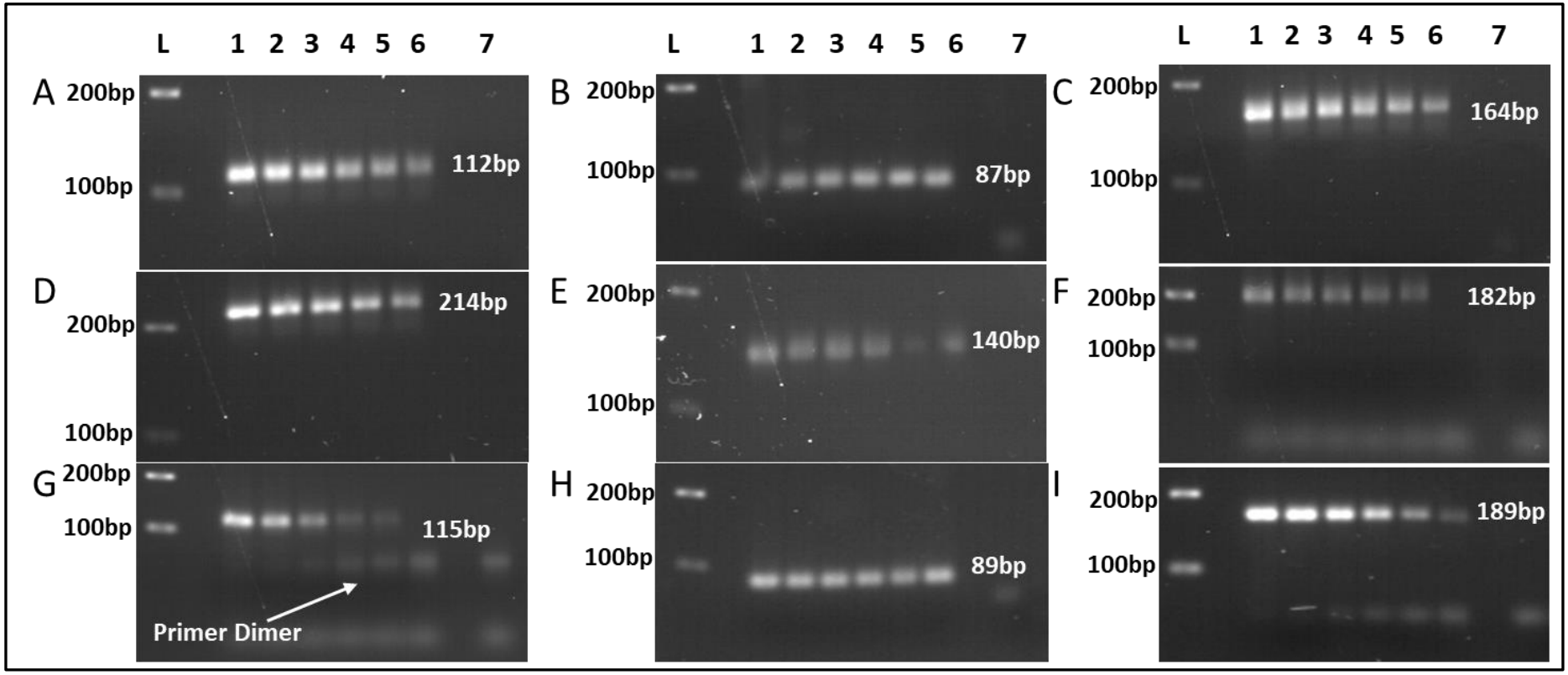


Figure 3.12. Agarose Gel Electrophoresis Results for Amplicons Generated from cDNA Amplification of Gene Target Transcripts in Non-ALS Control cDNA Derived Samples

The figure above displays agarose gel images of RT-qPCR amplicons that were generated following cDNA Amplification of Control non-ALS Sample A292/09 for each primer target. Primer targets for the above image are A) XPNPEP1, B) GAPDH, C) HERV-K *env*, D) HERV-K *gag*, E) HERV-K *pol*, F) HERV-K *RT*, G) HERV-W env, H) TDP-43 and I) BCL11b with sample wells corresponding to L) 100bp Ladder, DNA bands corresponding to 100bp and 200bp are shown, 1) Undiluted cDNA, 2) cDNA Diluted to 1:4, 3) cDNA Diluted to 1:16, 4) cDNA Diluted to 1:64, 5) cDNA Diluted to 1:256, 6) cDNA Diluted to 1:1024 and 7) Water Control.

Straight line graphs were plotted using the Ct values obtained from the relevant HERV-K and HERV-W *env* qPCR assays, along with the assays for reference genes XPNPEP1 & GAPDH and the transcriptional modifiers TDP-43 and BCL11b, versus the log dilution of cDNA that was performed (Figures 3.13 and 3.14). The slope of the line and R^2 values are summarised in Table 3.7, alongside calculated efficiency values displayed as a percentage. Efficiency percentages were derived from the equation $E = 10^{(-1/slope)}*100$ where the slope on the graphs (example from Figure 3.13A y= -3.3263x) determines the eventual efficiency. Only the primer sets with efficiency that were within the 90-110% recommended criteria for primer efficiency scores were selected which included all the primer pairs under evaluation. R^2 scores were > 0.99 or close, with HERV-W *env* being the exception with R^2 values at 0.96 for ALS and 0.94 for non-ALS control samples. This indicates high confidence in the trend of values within the slope. Only in data points with mean standard deviation <0.3 were used and resulted with graphs having ether 4, 5 or 6 data points. This was deemed acceptable for data analysis as the remaining points were at sufficient and consecutive dilutions to accurately map the slope with a sufficient R^2 value. These values were only deselected from inclusion in the graph if they were at the higher dilution values due to diminishing starting product.

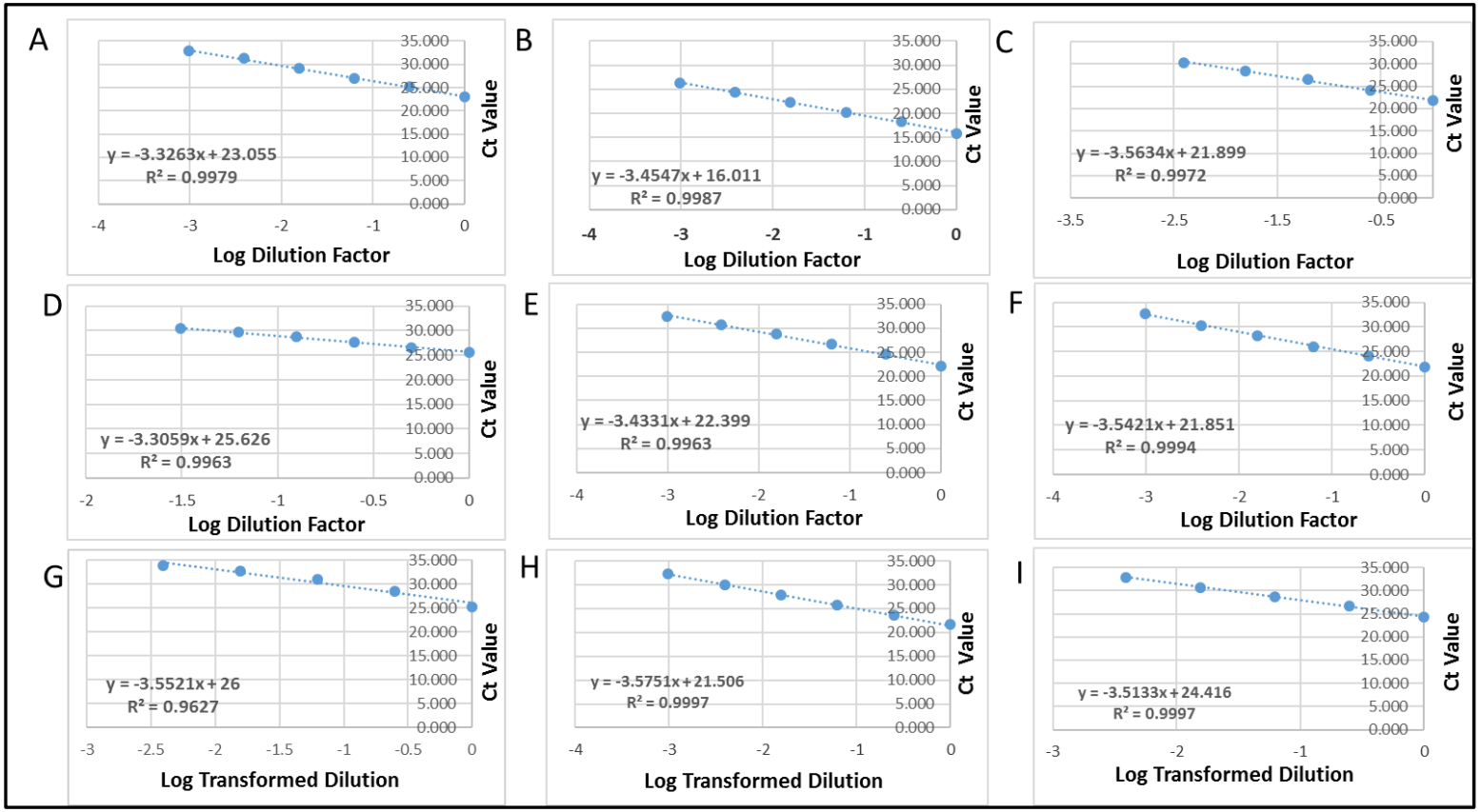


Figure 3.13. Amplification Efficiency Graphs for Reference Genes, HERV-K, HERV-W *env*, TDP-43 and BCL11b Primer sets performed on ALS derived cDNA.

The figure above displays standard curve graphs for ALS Patient Sample ID: A151/10 for each primer target. Primer targets for the above image are A) XPNPEP1, B) GAPDH, C) HERV-K *env*, D) HERV-K *gag*, E) HERV-K *pol*, F) HERV-K *RT*, G) HERV-W *env*, H) TDP-43 and I) BCL11b from cDNA amplification data. The axes for the graphs display Ct values on the y-axis plotted against log transformed dilution factors performed on the cDNA on the x-axis.

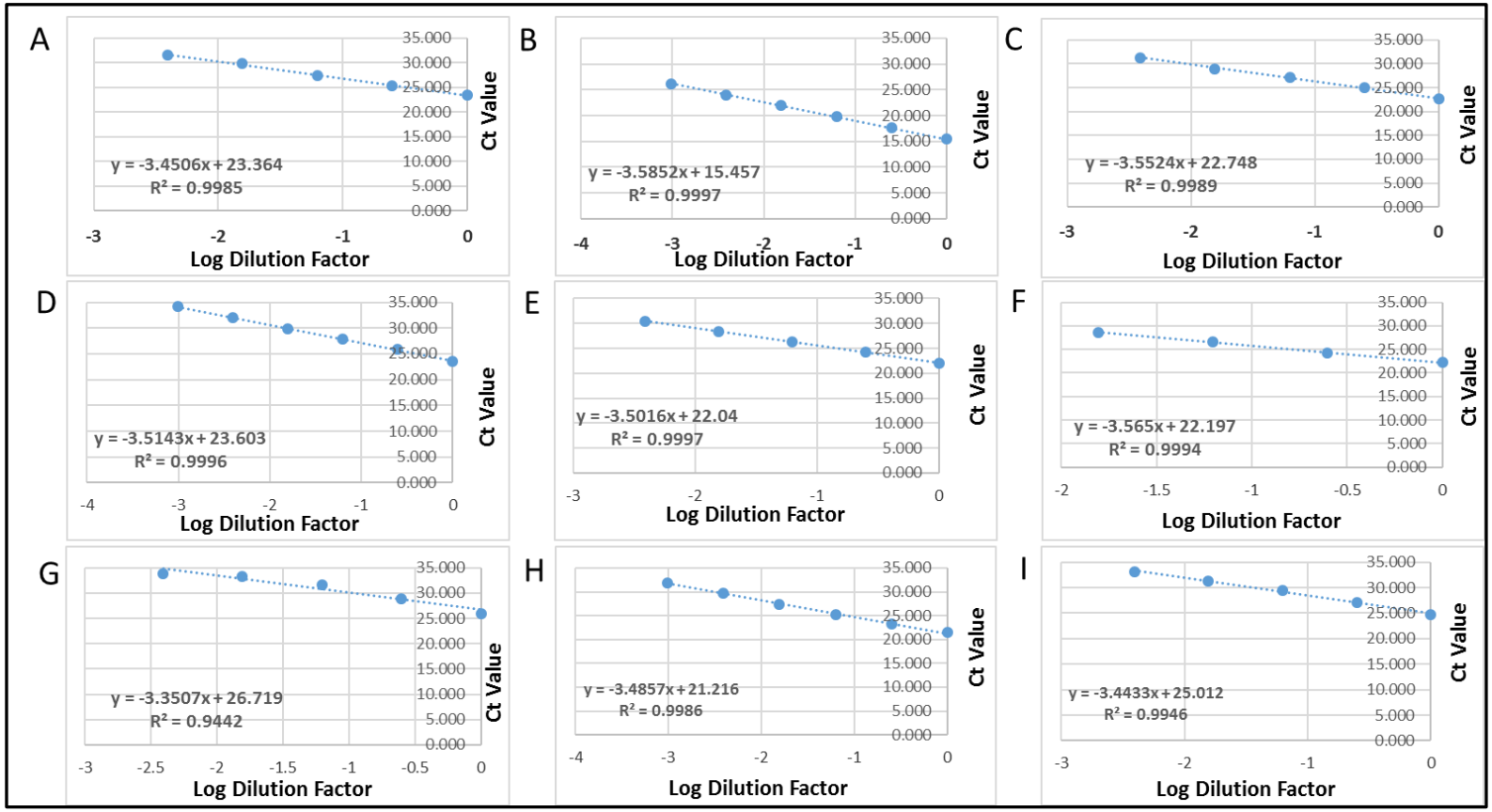


Figure 3.14. Amplification Efficiency Graphs for Reference Genes, HERV-K, HERV-W *env*, TDP-43 and BCL11b Primer sets performed on non-ALS derived cDNA.

The figure above displays standard curve graphs for Control Sample ID: A292/09 for each primer target. Primer targets for the above image are A) XPNPEP1, B) GAPDH, C) HERV-K *env*, D) HERV-K *gag*, E) HERV-K *pol*, F) HERV-K *RT*, G) HERV-W *env*, H) TDP-43 and I) BCL11b from cDNA amplification data. The axes for the graphs display Ct values on the y-axis plotted against log transformed dilution factors of cDNA on the x-axis.

Table 3.7. Summary of Amplification Efficiency Data for HERV-K, HERV-W *env*, TDP-43 and BCL11b primers tested on ALS and non-ALS Patient Sample.

The table below shows primer efficiency data obtained from Standard curves generated from cDNA amplification efficiency graphs shown in Figure 3.13 and Figure 3.14. Efficiency Percentages were generated from the equation $E = 10^{(-1/slope)}*100$.

Primer Target/ALS or Control	Primer Efficiency
XPNPEP1 (ALS)	99.82%
XPNPEP1 (Control)	94.89%
GAPDH (ALS)	97.74%
GAPDH (Control)	90.07%
HERV-K *gag* (ALS)	100.67%
HERV-K *gag* (Control)	92.20%
HERV-K *pol* (ALS)	95.56%
HERV-K *pol* (Control)	93.01%
HERV-K *env* (ALS)	90.79%
HERV-K *env* (Control)	91.21%
HERV-K *RT* (ALS)	91.57%
HERV-K *RT* (Control)	90.77%
HERV-W *env* (ALS)	97.48%
HERV-W *env* (Control)	91.22%
TDP-43 (ALS)	90.42%
TDP-43 (Control)	93.59%
BCL11b (ALS)	92.60%
BCL11b (Control)	95.18%

3.2.9 Confirmation of the Specificity of each Primer set by Sanger Sequencing of PCR amplicons

HERV-K *gag, pol and env* and RT amplicons as well as GAPDH, XPNPEP1, TDP-43 and BCL11b amplicons generated by RT-qPCR for the Control non-ALS and ALS sample were purified and the concentration of DNA was quantified using the Nanodrop (see section 2.2.4) before sending externally for Sanger sequencing by Eurofins (Germany) GATC Sanger Sequencing Service to confirm primer specificity. As the primer mix for XPNPEP1 was provided as a pooled sample from the company Qiagen (Germany) with the primer sequences being proprietary information, the purified XPNPEP1 PCR amplicons from the ALS and control sample were cloned into pGEM-T easy vector plasmid and transfected into JM109 high efficiency *E.coli* bacteria (see section 2.2.9). After Blue/White Colony selection on X-Gal IPTG ampicillin agar, culturing of putative clones in liquid broth media was performed,

followed by plasmid DNA isolation and quantification (see section 2.2.9), and the purified DNA was sent for sequencing using either of the M13 forward and reverse primer sets.

Sequences obtained for ALS and non-ALS control specimens for each of the HERV-K amplicons, and HERV-W *env* amplicons and for GAPDH, XPNPEP1, BCL11b and TDP-43 are provided in Table 3.8 and Table 3.10 respectively. The sequenced amplicons were then analysed by NCBI's nucleotide BLAST service which searches for areas of similarity in sequences using an algorithm to determine statistical significance of matched results obtained from 3 sources, NCBI's GenBank, Uniprot's Swiss Prot and TrEMBL. The BLAST search tool confirmed sequences for HERV-K *gag, pol, env* & *RT*, HERV-W *env*, GAPDH, XPNPEP1, BCL11b and TDP-43 gene targets (Tables 3.9 & 3.11, Supplementary images S1-S2). The BLAST searches show that the sequences of the PCR amplicons generated were highly similar to human genes with mostly single nucleotides differing from their closest matching reference sequence. The lowest of these for ALS and Controls was the HERV-K *env* primer amplicon which had 3 mis-matched nucleotides to its closest Genbank sequence in ALS and 4 in controls. Given the variability in repeat sequences this was deemed to be non-significant as the majority of hits were to human genes.

Table 3.8. Sequencing Information for PCR amplicons generated by PCR using HERV-K, HERV-W *env*, TDP-43, BCL11b primer sets and primers for XPNPEP1 and GAPDH reference genes obtained from ALS Samples.

The information given in the table below is the sequencing data obtained for the RT-qPCR amplicons generated using the different primer targets. The exception to this is the sequence data that was obtained for XPNPEP1 which was obtained by cloning PCR amplicons into pGEM-T easy vector and sequencing using M13 primer sets. Most sequences shown in the table are obtained following amplification of cDNA from one ALS sample (A151/10) by RT-qPCR.

Primer Gene Target (Fwd/Rev Primer)	5'-3' nucleotide sequence obtained by Sanger Sequencing of PCR amplicons.
XPNPEP1 (M13 Forward Primer)	CTACAGCCTACGAGAAGGAATGCTTCACATATGTCCTCAAGGGCCACATAGC TGTGAGTGCAGCCGTTTTCCCGACTGGAACCAAAGGTCACCTTCTTGACTCCT TTGCCCG
GAPDH (Forward Primer)	CATGACACTTTGGTATCGTGGAAGGACTCATGACCACAGTCCATAGCAC
HERV-K *gag* (Forward Primer)	TCAATACTGGCCGCCGGCTGAACTTCAGTATCGGCCACCCCCAGAAAGTCAG TATGGATATCCAGGAATGCCCCCAGCACCACAGGGCAGGGCGCCATACCCT CAGCCGCCCACTAGGAGACTTAATCCTACGGCAACAAA
HERV-K *pol* (Forward Primer)	GTCACTCAAGAGGCAGGAGTTAATCCCAGAGGTCTGTGTCCTAATGCATTAT GGCAAATGGATGTCACGCATGTACCTC
HERV-K *env* (Forward Primer)	CCTGTCACTTGGGTTAGACCATCGGAAGTACTATGATTATAAATCTCATATTA ATCCTTGTGTGCCTGTTTTGTCTGTTGTTAGTCTGCAGGTGTACCCAACAGCT CCGAAGAGACAGCCAA
HERV-K *RT* (Forward Primer)	TGCTTTTTTACCATCCCTCTGGCAAAGCAGGATTTTGAAAAATTTGCCTTTAC TATACCAGCCATACTATT
HERV-W *env* (Forward Primer)	AGGGTACATGAGCACCTCTAGCCCCTACAAAGGACTAGAGGTCT
TDP-43 (Forward Primer)	AGACTTTGCCTTTGTTACATTTGCAGATGATCAGATTGCAGCAGACCGTAC
BCL11b (Forward Primer)	GCCTGGGGCTGATGGTGGGTGGCCCCCACCCTGACCTGCTCACCTGTGGCCGG

Table 3.9. NCBI BLAST Results for HERV-K, HERV-W *env*, TDP-43, BCL11b and Reference Genes Sequences obtained from one ALS Sample (A151/10).

The information displayed in the table below shows BLAST search results for sequencing information displayed in Table 3.8. Search results given for each primer gene target are shown and the Accession number of the closest match and percentage identity of the closest match for each PCR amplicon are shown.

Primer Gene Target	NCBI Reference Sequence and Accession Number of Closest Match	Sequence Coverage
XPNPEP1 (M13 Forward Primer)	NM_020383.4: Homo sapiens X-prolyl aminopeptidase 1 (XPNPEP1), transcript variant 1, mRNA	112/112**(100%)**
GAPDH (Forward Primer)	NM_002046.7: Homo sapiens glyceraldehyde-3-phosphate dehydrogenase (GAPDH), transcript variant 1, mRNA	44/45**(98%)**
HERV-K *gag* (Forward Primer)	XM_017007620.2: PREDICTED: Homo sapiens endogenous retrovirus group K member 7 Gag polyprotein (LOC107986113), mRNA	135/136**(99%)**
HERV-K *pol* (Forward Primer)	AF298588.1: Homo sapiens clone 2a HERV-K polymerase (pol) gene, pol-HML-2.HOM allele, partial cds	76/76**(100%)**
HERV-K *env* (Forward Primer)	JN202404.1: Human endogenous retrovirus K envelope protein (Env) gene, partial cds	117/120**(98%)**
HERV-K *RT* (Forward Primer)	DQ821442.1: Homo sapiens endogenous virus HERV-K reverse transcriptase (pol) mRNA, partial cds	64/66**(97%)**
HERV-W *env* (Forward Primer)	LT744319.1: Human ORFeome Gateway entry vector pENTR223-ERVW-1, complete sequence.	37/38**(97%)**
TDP-43 (Forward Primer)	HQ628636.1: Homo sapiens TDP43 isoform I (TDP43) mRNA, complete cds, alternatively spliced	40/41**(98%)**
BCL11b (Forward Primer)	NM_001282237.2: Homo sapiens BAF chromatin remodelling complex subunit BCL11B (BCL11B), transcript variant 3, mRNA	50/51**(98%)**

Table 3.10. Sequencing Information for PCR amplicons generated by PCR using HERV-K, HERV-W *env*, TDP-43, BC11b primer sets and primers for XPNPEP1 and GAPDH reference genes obtained from a non-ALS control and Reference Gene Primer Targets obtained from non-ALS Controls.

The information given in the table below is the sequencing data obtained for the RT-qPCR amplicons generated using the different primer targets. The exception to this is the sequence data that was obtained for XPNPEP1 which was obtained by cloning PCR amplicons into pGEM-T easy vector and sequencing using M13 primer sets. Most sequences shown in the table are obtained following amplification of cDNA from one non-ALS sample (A292/09) by RT-qPCR.

Primer Gene Target (Fwd/Rev Primer)	5'-3' nucleotide sequence obtained by Sanger Sequencing of PCR amplicons.
XPNPEP1 (M13 Forward Primer)	CTACAGCCTACGAGAAGGAATGCTTCACATATGTCCTCAAGGGCCACATAGC TGTGAGTGCAGCCGTTTTCCCGACTGGAACCAAAGGTCACCTTCTTGACTCCT TTGCCCG
GAPDH (Reverse Primer)	GTCATGGATGACCTTGGCCAGGGGTGCTAAGCAGTTGGTGGTGCAACGGTTG
HERV-K *gag* (Forward Primer)	ATCAATACTGGCCGCCGGCTGAACTTCAGTATCGGCCACCCCCAGAAAGTCA GTATGGATATCCAGGAATGCCCCCAGCACCACAGGGCAGGGCGCCATACCC TCAGCTGCCCACTAGGAGACTTAATCCTACGGCACAAA
HERV-K *pol* (Forward Primer)	CAGTGTCAGTCTTACACCTGTCCACTCAAGAGGCAGGAGTTAATCCCAGAG GTCTGTGTCCTAATGCGTTATGGCAAATGGATGTCACGCATGTACCTCG
HERV-K *env* (Forward Primer)	GACCATCGGAGTACTATGATTATAAATCTCATATTAATCCTTGTGTGCCTGTT TTGTCTGTTGTTAGTCTGCAGGTGTACCCAACAGCTCCGAAGAAACAGC
HERV-K *RT* (Forward Primer)	ATGATCCCAAAGATTGGCCTTTATTTATAATTGATCTAAAGGATTGCTTTTTT ACCATCCCTCTGGCGGAGCAGGATTGTGAAAAATTTGCCTTTACTATACCA GCCATAAATAATAAAGAACCAGCCACCAGGTTCA
HERV-W *env* (Reverse Primer)	TTACTTCTTTTACATGTTTTTCTCTTGCCTGATCTTGAACTCCACCCCCATCCGAC
TDP-43 (Forward Primer)	TTTGGGGATGAGACATCCATCACATCCCCGTAT
BCL11b (Forward Primer)	GCCTGGGGCTGATGGTGGGTGGCCCCCACCCTGACCTGCTCACCTGT

Table 3.11. NCBI BLAST Results for HERV-K, HERV-W *env*, TDP-43, BC11b and Reference Genes Sequences obtained from one non-ALS sample (A292/09).

The information displayed in the table below shows BLAST search results for sequencing information displayed in Table 3.10. Search results given for each primer gene target are shown and the Accession number of the closest match and percentage identity of the closest match for each PCR amplicon are shown.

Primer Gene Target	NCBI Reference Sequence and Accession Number of Closest Match	Sequence Coverage
XPNPEP1 (M13 Forward Primer)	NM_020383.4: Homo sapiens X-prolyl aminopeptidase 1 (XPNPEP1), transcript variant 1, mRNA	112/112**(100%)**
GAPDH (Reverse Primer)	NM_002046.7: Homo sapiens glyceraldehyde-3-phosphate dehydrogenase (GAPDH), transcript variant 1, mRNA	45/45**(100%)**
HERV-K *gag* (Forward Primer)	DQ157723.1: Human endogenous retrovirus K clone 4.4 non-functional gag protein (gag) gene, partial sequence	137/138**(99%)**
HERV-K *pol* (Forward Primer)	KF254365.1: Homo sapiens endogenous virus HERV-K clone 11A31.Lm reverse transcriptase (pol) mRNA, partial cds	96/99**(97%)**
HERV-K *env* (Forward Primer)	JN202404.1: Human endogenous retrovirus K envelope protein (Env) gene, partial cds	99/103**(96%)**
HERV-K *RT* (Forward Primer)	DQ841442.1: Homo sapiens endogenous virus HERV-K reverse transcriptase (pol) mRNA, partial cds	134/137 **(98%)**
HERV-W *env* (Reverse Primer)	LT744319.1: Human ORFeome Gateway entry vector pENTR223-ERVW-1, complete sequence.	55/56**(98%)**
TDP-43 (Forward Primer)	HQ628636.1: Homo sapiens TDP43 isoform I (TDP43) mRNA, complete cds, alternatively spliced	31/32**(97%)**
BCL11b (Forward Primer)	NM_001282237.2: Homo sapiens BAF chromatin remodelling complex subunit BCL11B (BCL11B), transcript variant 3, mRNA	46/47**(98%)**

3.3 Discussion

As highlighted in the MIQE guidelines, RT-qPCR is an immensely popular tool for determining gene expression levels in any given biological sample, although scientific publications based upon the data obtained from RT-qPCR assays still lack consistency when it comes to reporting the procedures involved in generating and analysing the data obtained (Bustin *et al.*, 2009). When it came to looking for suitable reference genes to be used in this research study, it was important to determine the reference gene that was most stably expressed in the premotor cortex brain tissue from disease and non-ALS disease state which is an important factor that has been reported in the literature (Penna *et al.*, 2011; Eisenberg and Levanon, 2013). For instance, if only one reference gene is used for normalisation purposes a more in-depth analysis of the qPCR data is warranted as well as justifying the selection of a particular reference gene, with multiple reference genes being strongly recommended for more accurate quantification of mRNA expression levels in test samples (Bustin *et al.*, 2009).

Whilst the sample source is important in determining the reference gene to select, other additional information regarding sample quality is needed to gauge their validity for use in this study, with information such as RNA integrity, donor source (including cause of death), tissue region, post mortem interval, sex and age all giving useful data for analysis of the suitability of a given sample (Koppelkamm *et al.*, 2011; Dean, Udawela and Scarr, 2016). RNA integrity is possibly the most important metric as it is a measure of the quality of RNA in a given sample. Post-mortem interval can have a profound effect on RNA integrity as RNA degrades over time if not processed or stored correctly, as well as the conditions at the time of death can have a profound effect (Koppelkamm *et al.*, 2011). The most significant factor affecting RNA integrity in post-mortem brain tissue is prolonged stress prior to death as this creates an acidic environment in the brain contributing to the accelerated degradation of RNA in the tissue (Durrenberger *et al.*, 2010; Koppelkamm *et al.*, 2011). Unfortunately, in this analysis the cause of death for the non-ALS cases was not provided in detail for all samples (Materials Table 2.5) though the range of RNA integrity values for samples tested in the assays ranged between 4.0 and 7.0. Additional information that we were provided by the MRC Neurodegenerative Disease Brain Bank (London, UK), relating primarily to the variation in sex and age, are important as they will provide

important information on how well distributed the stability of expression levels of certain reference genes are across genders with increasing age and this is an important factor to consider as highlighted in the literature (Touchberry *et al.*, 2006; Naumova *et al.*, 2013). For this reason, the ALS and non-ALS brain tissue samples used in the RT-qPCR assay were selected for these particular variables in order to select the ideal set of reference genes that are stably expressed in both disease and non-disease state so that they can be utilised for normalisation of expression levels of certain genes.

As recommended by the geNorm protocol 10 postmortem premotor cortex samples were selected consisting of n=5 ALS and n=5 non-ALS controls, which were matched as ideally as possible according to age and sex. This number of samples was selected as it provided a subset of tissue samples that was representative of the whole cohort. These samples were tested in triplicate in the RT-qPCR assay to minimise variations due to sampling or human error, with duplicate or triplicate results taken that had favourable Standard Deviation values, falling below 0.5 SD in variance. All primers that targeted specific reference genes, were provided either by the company Primer Design or obtained from research findings published by Durrenberger *et al.*, (2012) and Li *et al.*, (2015). All of the RT-qPCR experiments that were performed for each of the panel of 9 candidate reference genes, produced a single PCR amplicon of the expected size as confirmed by agarose gel electrophoresis and melt curve analysis (Figures 3.1- 3.3). This provided strong evidence as to the specificity of the primers to the targeted reference gene. The differences in Ct values obtained by RT-qPCR resulting from input cDNA or RNA into the reaction was 10 Ct cycles representing a 1000-fold difference in concentration of the respective genes, indicating that overall, the DNaseI treatment of total RNA was successful in removing residual genomic DNA (gDNA) from the extracted samples. However, additional bands were observed in the gel images for GAPDH, RPL13A and β-Actin when RNA was spiked into the RT-qPCR assays and are likely a result of a high copy number of the original gene i.e. GAPDH or the result of primer dimer bands or incomplete DNaseI digestion. Another potential factor in the appearance of additional bands on the 2% agarose gel for β-Actin and GAPDH is the presence of numerous pseudogenes, non-functional homologues of functional genes in the chromosome which have become inactive due to numerous mutations (Sun *et al.*, 2012). Both of these genes have over 50 pseudogenes associated with them which could

potentially result in numerous bands on an agarose gel if primer sequences also match regions of the pseudogene (Sun *et al.*, 2012).

In this study, geNorm, which is the one of the most popular algorithms for measuring stability values of reference gene candidates (Curis *et al.*, 2019) was the primary focus for analysis of our experimental data, which is generated using qBase+ (Biogazelle, Zwijnaarde, Belgium), which is an excellent analytical tool and provides information about the effects of variability of samples on gene expression levels. When looking at gene stability within patient samples any identified high instability of reference genes should be analysed for any potential correlation with disease status in particular. The reference gene candidates that were selected in this study were supplied by Primer Design (UK) as they have been used in other studies involving measuring gene expression in brain tissue. Some of the reference genes that were selected were: RPL13A, which is involved in viral mRNA translation, Ubiquitin C is involved in innate immunity and cellular stress responses and removal of toxic proteins and YWHAZ which is associated with Schizophrenia, which has been shown to have links with HERV expression in brain tissue (Slokar and Hasler, 2016; Küry *et al.*, 2018). The 3 least stably expressed genes, YWHAZ, EIF4A2 and RPL13A that was reported in this study according to the geometric mean ranking, could have some significance when related to disease state. Each of these genes has the potential to be upregulated in response to increased HERV-K expression, with 2 of the genes, EIF4A2 and RPL13A being involved in the translation of mRNA. EIF4A2 and YWHAZ have been identified by RT-qPCR as highly unstable genes in this study (Jia *et al.*, 2004; Douville *et al.*, 2011; Suntsova *et al.*, 2013; Slokar and Hasler, 2016).

Other reference gene selection methods have their own programs, with NormFinder and BestKeeper having free excel based software available while the ΔCt method requires users set up calculations manually. An easy method of comparing the geNorm, NormFinder, BestKeeper and ΔCt methodologies is RefFinder, an online tool with these algorithms available to analyse RT-qPCR data. As we can see in Table 3.1, RefFinder also provides their own comprehensive ranking system which calculates the geometric mean of rank positions in the other methods analysed to give a ranking based on all of the different methodologies provided. While this is a useful method of validating your selection method it is useful to corroborate the information given to the original programs where available. BestKeeper is

the most disparate in its ranking of reference genes, which had been confirmed using the original software (Table 3.4), showing an almost completely different stability rating compared to the other software methods. BestKeeper has been shown in other studies, such as the reference gene selection work done by Petriccione *et al.*, (2015), which consistently showed disparity between BestKeeper rankings compared to geNorm, Normfinder and ΔCt reference gene selection methods. However, as BestKeeper has been compared to geNorm, NormFinder and the ΔCt algorithms in other studies and provided similar rankings and this may be down to differences in the variability of genes in the tissue under investigation.

Differences in RefFinder analysis can be seen in the data provided by the NormFinder excel program, with a slight change in the rankings and a marked change in the stability values generated (Table 3.1 & 3.2). As the NormFinder mathematical model requires groups to be assigned to the data to calculate intergroup variation, such as ALS and Control brain tissue samples in this instance, as well as generates confidence intervals for the data, this might explain the slight disparity in the results (Andersen, Jensen and Ørntoft, 2004). RefFinder does not have any option for separating out sample groups and while this can still produce a stability metric it does not have the same functionality as the complete mathematical model of geNorm, which can provide a stability value for 2 combined genes (Figure 3.6), and this difference was enough to alter the geometric mean rankings of the candidate reference genes (Table 3.5), although the top two reference genes, GAPDH and XPNPEP1, in the revised geometric mean ranking still coincided with the pair provided by NormFinder (Table 3.3). Using this combined ranking of methods helps to identify stably expressed genes across all mathematical processes and allows researchers to add an additional layer of validation to their selections of candidate reference genes.

Selecting the correct set of reference genes for a study is an important part of the validation process for ensuring MIQE compliance for future publication (Bustin *et al.*, 2009; Bustin and Wittwer, 2017). The information provided by the reference gene selection process in this study has provided GAPDH and XPNPEP1 as the 2 candidate reference genes to be used for normalisation of RT-qPCR gene expression data obtained from the premotor cortex region of the brain. The ideal number of reference genes to be used was confirmed in geNorm and NormFinder, with the former selecting 2 reference genes based on the lowest number of

reference genes provided by the geNorm V values below the cut-off value of 0.15. Based upon these findings both GAPDH and XPNEP1 will be used in all downstream RT-qPCR assays for normalisation of HERV gene expression levels in ALS derived premotor cortex post-mortem brain tissue.

Designing primers for HERV-K is challenging, as HERV-K is present in multiple copies throughout the human genome, with sequences present in the form of both full length *gag-pol-env* and partial sequences having undergone multiple cycles of silencing mutations (Subramanian *et al.*, 2011; Garcia-Montojo *et al.*, 2018). To ensure that as many of the different HERV-K sequences that are available would be targeted by our HERV-K primer sets, full-length (5'LTR-*gag-pol-env*-3'LTR) sequences from NCBI's GenBank service were gathered. Together with the sequences obtained from the paper by Subramanian *et al.*, (2011), this allowed sufficient complexity of possible sequence variations to be considered for primer design. New sets of primers were selected for each target genomic region of HERV-K, with outputs for both the NCBI primer BLAST and the UCSC *in-silico* searches indicating that the primer sets were able to capture a wide range of HERV-K family members. The HERV-K alignment dataset was then used to validate a set of HERV-K primers from the literature (Li *et al.*, 2015) in order to verify that they would amplify a wide range of HERV-K sequences.

While the Li *et.al.* primer sets targeting the *env* and *pol* regions align well, the *gag* primers potentially align to a more variable section of the HERV-K genome. This is evidenced by the relatively low number of matched sequences obtained (Table 3.6, Supplementary Figures S3-S14). However, this observation may not be fully representative of the entire HERV-K family for this region as there may be many more sequences not annotated within the genome (Tokuyama *et al.*, 2018a). Since the newly designed HERV-K *gagED* primers that I designed did not meet the criteria for RT-qPCR primer efficiency amplification (data not shown) and showed multiple bands on 2% Agarose gel electrophoresis it was decided to stick with the original Li *et.al.* HERV-K primers for this genomic region. Additionally, the newly designed HERV-K *env* primer that I designed, while aligning to more full-length *gag-pol-env* HERV-K sequences than the Li *et.al.* HERV-K env primer set; they did not offer a significant improvement in matches to full length sequences, and therefore the Li *et.al.* primers for HERV-K *env* were therefore carried forward in downstream RT-qPCR assays.

The *pol* region of HERV-K encodes reverse transcriptase, RNaseH and Integrase and the Li *et.al.* HERV-K *pol* primers amplify within the integrase region (Figure 3.8). However previous studies have reported increased reverse transcriptase activity in ALS patient serum samples compared to non-ALS control serum samples (Steele *et al.*, 2005; McCormick *et al.*, 2008). Therefore, we designed primers within the reverse transcriptase region as this is a biomarker of retroviral activity. The newly designed primer sets (HERV-K *RT*) passed the primer efficiency criteria 90-110%, otherwise stated as a slope value between -3.3 and -3.6, (Table 3.7) and produced a single amplicon of the expected size, following gel electrophoresis (Figure 3.11) and was taken forward into RT-qPCR expression assays.

HERV-W *env* primers were obtained from the literature (Levet *et al.*, 2017) and resulted in primer efficiency values of 91% and 97% in terms of being efficient in amplification of the target region (Table 3.7), and produced an amplicon of the expected size on a 2% agarose gel. A second DNA band appearing at a higher template dilution appeared at the same position on the gel as the band present in the water control reaction, indicating that the DNA band is likely to be the result of primer dimers. These primers were also assessed for specificity to target sequence by Sanger sequencing (Tables 3.8-3.11) and were only taken forward if they produced an amplicon that could be sequenced and confirmed as being specific to human HERV-W *env* sequences. This was true for the HERV-K *gag, pol* & *env* primers along with GAPDH, XPNPEP1, BCL11b and TDP-43 primer pairs as well. Primer sequences for TDP-43 and BCL11b were also obtained from the literature (Douville *et al.*, 2011; Bartram *et al.*, 2014) though needed to be validated for postmortem premotor cortex tissue by primer efficiency and specificity for human genes. Both primer sets proved to be within the 90%-110% efficiency range, producing a single amplicon on 2% gel electrophoresis and showed specificity to known human gene sequences for their proteins. This provided evidence of their suitability for use in the RT-qPCR expression assays.

Assessing amplification efficiencies of potential candidate primer sets is included under the MIQE guidelines as an essential step in validating RT-qPCR assays. Therefore, only those primer sets with efficiency values within the appropriate range, 90%-110% (Bustin *et al.*, 2009) were selected for downstream RT-qPCR expression assays. Many potential factors can influence the estimation of efficiency values, including; inhibition in cDNA synthesis,

pipetting errors when preparing dilutions of cDNA or when adding cDNA template to the qPCR reaction wells (Svec *et al.*, 2015). To ensure accuracy, RT-qPCR assays were conducted in duplicate and standard deviation (SD) of less than 0.3 was maintained for each set of duplicate samples tested.

In conclusion, findings from these experiments have produced a suitable primer set for targeting within the *pol* region of HERV-K, nominally named HERV-K *RT*, for the use in quantifying HERV-K expression in post-mortem premotor cortex tissue samples and verified that the Li et al 2015 primer sets and HERV-W-*env*, BCL11b and TDP-43 primer sets were specific for the target gene and could be used in downstream RT-qPCR assays to measure relative gene expression levels in frozen post-mortem brain tissue of sporadic ALS patients compared to controls.

4.0 Quantification of HERV-K, HERV-W, TDP-43 and BCL11b gene expression in ALS and non-ALS post-mortem premotor cortex tissue samples by RT-qPCR

4.1 Introduction

HERV-K expression has been associated in various neurodegenerative diseases, including Multiple Sclerosis (MS), Schizophrenia and Amyotrophic Lateral Sclerosis (ALS) (Brudek *et al.*, 2007; Li *et al.*, 2015; Slokar and Hasler, 2016; Mayer *et al.*, 2018; Savage *et al.*, 2018). The principle research confirming a link with HERV-K and ALS was reported by Li et al., 2015 in an American cohort, in which they observed a 2/3-fold increase in expression of HERV-K transcripts in ALS post-mortem frozen tissue compared to controls. This work built on previous studies showing actively replicating loci of endogenous retroviruses in human premotor cortex brain tissue as detected by HERV-K *pol* transcripts (Douville *et al.*, 2011), evidence of reverse transcriptase activity (a retroviral marker) in ALS patients serum (Steele *et al.*, 2005; McCormick *et al.*, 2008), and the observation of ALS-like symptoms in HIV and HTLV retroviral infections (Alfahad and Nath, 2013; Bowen *et al.*, 2016; Douville and Nath, 2017). These observations have however been called into question recently as other research groups have not been able to confirm these recent findings by Li et al, 2015 (Mayer *et al.*, 2018; Garson *et al.*, 2019). In the paper by Garson *et al.*, 2019 the methodology of Li *et al.*, 2015 was followed closely, modifying the process to conform to the MIQE guidelines (Bustin *et al.*, 2009) with the inclusion of a second reference gene, XPNPEP1 which was identified by the validation of reference genes using different algorithms as documented in section 3.0 of this thesis.

Alongside HERV-K another endogenous retroviral family, HERV-W, has been implicated in neurodegenerative conditions and has been reported to have intact full-length ORFs encoding for *gag*, *pro*, *pol* and *env* viral proteins and the potential to form virus like particles. HERV-W has been shown to be upregulated in certain ALS cases as well as being linked to familial ALS through the expression of Superoxide Dismutase 1 (SOD-1), which can damage cellular processes when linked with HERV-W expression (Ajroud-Driss and Siddique, 2015; Li *et al.*, 2015; Küry *et al.*, 2018). This research suggests a potential

relationship between these two HERV families though a definitive association has yet to be proven.

TDP-43 and BCL11b are both known modifiers of viral transcription in HIV, the former involved in the binding of viral nucleic acids prior to packaging and the latter involved in the suppression of HIV transcription during vial latency in the central nervous system (Cismasiu *et al.*, 2008; Desplats *et al.*, 2013; Douville and Nath, 2017). The in depth study of HERV-K expression by Li *et al.*, 2015 showed that overexpression of TDP-43 in HeLa cells *in vitro*, resulted in increased expression of HERV-K transcripts providing a potential link between the expression of TDP-43 and its effects on HERV-K expression. More recently Douville and Nath (2017) provided a link between HIV and HERV-K pathology through neurotoxic TDP-43 accumulation in nerve cells. Research into a link between BCL11b and HERV-K has been more scarce than TDP-43, though Lennon *et al.*, (2016 & 2017) reported that BCL11b can suppress proviral HIV expression by binding to the LTR regions of HIV-1, and it is thought that BCL11b can perform a similar function in HERV-K pathology (Desplats *et al.*, 2013; Lennon *et al.*, 2016).

The premotor region of the motor cortex is of interest in this study as progression of ALS in the brain and the involvement of the premotor cortex region corresponds to stage 2 of neurodegeneration, the first stage presenting as abnormalities in the betz cells in layer V of the primary motor cortex (Eisen et al., 2017). During stage 2 lesions develop in the premotor areas marking the beginning of cerebellar dysfunction in the disease (Eisen et al., 2017). Additionally, this has been the area of interest in previous studies looking at HERV-K expression in ALS patient postmortem brain tissue by Douville et.al. (2011) and Li et.al. (2015) which showed differential expression of HERV-K transcripts in the motor cortex and morphological changes to the premotor region in transgenic mice.

RT-qPCR is a popular tool for estimating both absolute and relative changes in gene expression of transcripts in biological material obtained from disease and non-disease states, as well as monitoring the effect of treatment regimens on gene expression (Pfaffl, Vandesompele and Kubista, 2009; Jia, 2012; Bhetariya, Kriesel and Fischer, 2017). Relative quantification of gene expression by RT-qPCR is used to determine changes in expression levels between patient sample sets (Livak and Schmittgen, 2001; Schmittgen and Livak, 2008). The principle method for calculating relative quantification in expression studies is

by using a mathematical model referred to as the 2^-ΔΔCt method, which calculates the relative change in gene expression RT-qPCR data, normalised against one or more stably expressed reference gene (Livak and Schmittgen, 2001). The 2^-ΔΔCt calculates the difference in Ct values between the Ct of the gene of interest and the Ct of the reference gene mean for a sample (ΔCt) which is then normalised to a calibrator such as the mean of all control ΔCt's (Livak and Schmittgen, 2001). Expression is therefore determined relative to the mean expression of the control group. An alternate mathematical model to the 2^-ΔΔCt method exists in the form of the Pfaffl method (M. W. Pfaffl, 2001). The Pfaffl model for relative gene quantification, unlike the 2^-ΔΔCt method, accounts for the amplification efficiency of the primer sets in the experiment (M. W. Pfaffl, 2001). This normalises the RT-qPCR expression data from different gene targets with separate amplification efficiency values allowing for a more accurate estimation of relative expression across all gene targets. In this study we will be using both the 2^-ΔΔCt and Pfaffl approaches to analyse gene expression data generated by RT-qPCR to determine if there is a difference in relative gene expression of HERV-K/W, TDP-43 and BCL11b between ALS and no-ALS cases.

The research work described here aims to provide an independent analysis of HERV-K and HERV-W expression in the premotor cortex of ALS and non-ALS cases for a UK cohort, which have been matched as close as possible for age, sex, and post-mortem delay (PMD). The same set of HERV-K *gag*, *pol* and *env* primer sets and GAPDH specific primers were used in this study as were adopted by Li *et al.*, 2015 in their HERV-K RT-qPCR expression studies. We incorporated an additional reference gene, XPNPEP1 to be used alongside GAPDH for normalisation of HERV-K expression in brain tissue of ALS and non-ALS case controls. HERV-W *env* expression was also analysed in brain tissue from the same UK cohort of ALS and control samples, as HERV-W has been liked to other neurological conditions such as MS disease (Levet *et al.*, 2017). In addition, regulators of retroviral transcription BCL11b and TDP-43 (Desplats *et al.*, 2013; Li *et al.*, 2015; Lennon *et al.*, 2016) were tested alongside HERV transcripts in order to analyse their potential effects on HERV –K transcription in ALS derived premotor cortex tissue.

4.2 Results

4.2.1 Extraction and quantification of total RNA isolated from ALS and non-ALS post-mortem premotor cortex brain tissue.

Total RNA was extracted from 50-75 mg of post-mortem premotor cortex brain tissue; n=20 ALS and n-20 non-ALS, (see materials table 2.5). Only samples with RNA concentration above 125 ng/µl and RIN values above 4.0 were selected for HERV-K and HERV-W specific RT-qPCR assays. Table 4.1 below shows RIN values as measured by the Agilent Bioanalyser 2100 with RNA concentrations measured by Qubit BR (Broad Range) assay.

Table 4.1. Summary of the Quantification of Total RNA Extracted from n=20 ALS and n=20 Non-ALS Premotor Cortex brain tissue Samples obtained at post-mortem.
In the table below summary information is given on RIN values obtained from Agilent Bioanalyser 2100 along with range and medians of RNA yield as measured using the Qubit BR (Broad Range) assay. Qubit means were derived from duplicate/triplicate values that were within 40ng/µl of each other, and only those values used for the mean quantification of RNA yield are given in the table. Full data is shown in Supplementary Table 14.

Variable	Summary Statistic	Value
RIN	Median	6.2
	Range	4.1-7.8
QuBit Derived Conc. ng/µl	Median	530
	Range	131-841

Out of a total of forty premotor cortex brain tissue samples that were processed, only one ALS derived brain tissue specimen was rejected (A331/09) due to the low concentration of RNA that was extracted as shown in Table 4.1, which prevented RIN from being derived for the sample. Therefore, the 39 remaining samples (n=19 ALS and n=20 non-ALS controls) were used for downstream RT-qPCR assays to measure HERV-K and HERV-W expression in these 2 groups with normalisation against GAPDH and XPNPEP1 reference genes. These were identified in chapter 3.0 as the most stable reference genes for use in post-mortem premotor cortex brain tissue samples.

4.2.2 HERV-W *env*, HERV-K *gag, pol, env* and RT Expression in ALS and Non-ALS derived premotor cortex brain tissue obtained at post-mortem.

Prior to measurement of HERV-K transcript expression the individual assays were assessed using quality control methods initially described in chapter 3. To summarise the results of the quality control steps briefly, Ct values for all gene targets appear within a 5Ct range, indicating a 32-fold difference in expression across ALS & Non-ALS control samples. Each patient sample was tested in duplicate with nearly all duplicate Ct values reported as being below 0.3 Standard Deviations (SD) from one-another (Supplementary Material Table S3, Figure S195). Additionally Supplementary Figure S196 shows a single amplicon produced for all datasets with residual genomic DNA presence measured using an RNA spike showing a difference in Ct values of 5-10 cycles indicating minimal impact on measured expression (data not shown).

The mean relative expression levels of all 4 HERV-K transcripts (*gag, pol, env* & *RT*) were slightly lower in ALS than controls when normalised against the XPNPEP1 reference gene compared to GAPDH (Table 4.2). In each case there was minimal difference in the geometric mean expression levels of each of the HERV-K transcripts. HERV-K *RT* expression levels are of interest due to being slightly higher than HERV-K *gag, pol* & *env* expression levels (although expression was not significantly different compared to controls).

Table 4.2. Geometric Mean of HERV-K *gag, pol, env* & *RT* Relative Expression in ALS and non-ALS control cases, Normalised to GAPDH or XPNPEP1 Reference Genes.
The table displays the geometric means of the **2^-ΔΔCt differential expression values** n=19 ALS and n=20 non-ALS control samples for each of the HERV-K gene targets used in the RT-qPCR expression assay. These were normalised to 2 separate reference genes, GAPDH (left) and XPNPEP1 (right).

	GAPDH				XPNPEP1			
	HERV-K gag	HERV-K pol	HERV-K env	HERV-K RT	HERV-K gag	HERV-K pol	HERV-K env	HERV-K RT
ALS	1.139	1.180	1.134	1.235	0.955	0.989	0.950	1.035
Control	1	1	1	1	1	1	1	1
p value	0.681	0.281	0.850	0.192	0.347	0.641	0.509	0.893
Statistical Significance	NS	NS	NS	NS	NS	NS	NS	NS

When HERV-K RT-qPCR data was normalised to GAPDH using the 2^-ΔΔCt method there was no statistically significant difference in expression of transcripts between ALS and non-ALS control cases (Figure 4.1). Additionally, when the data was normalised to XPNPEP1, there was also no statistically significant difference in expression of HERV-K gene transcripts between ALS and non-ALS control cases. This data was initially generated using the ΔΔCt differential expression method in Microsoft Excel and p-Values determined using GraphPad PRISM 6.0. Whether normalised against GAPDH or XPNPEP1 the p-values for all ALS/Control comparisons was greater than p=0.1.

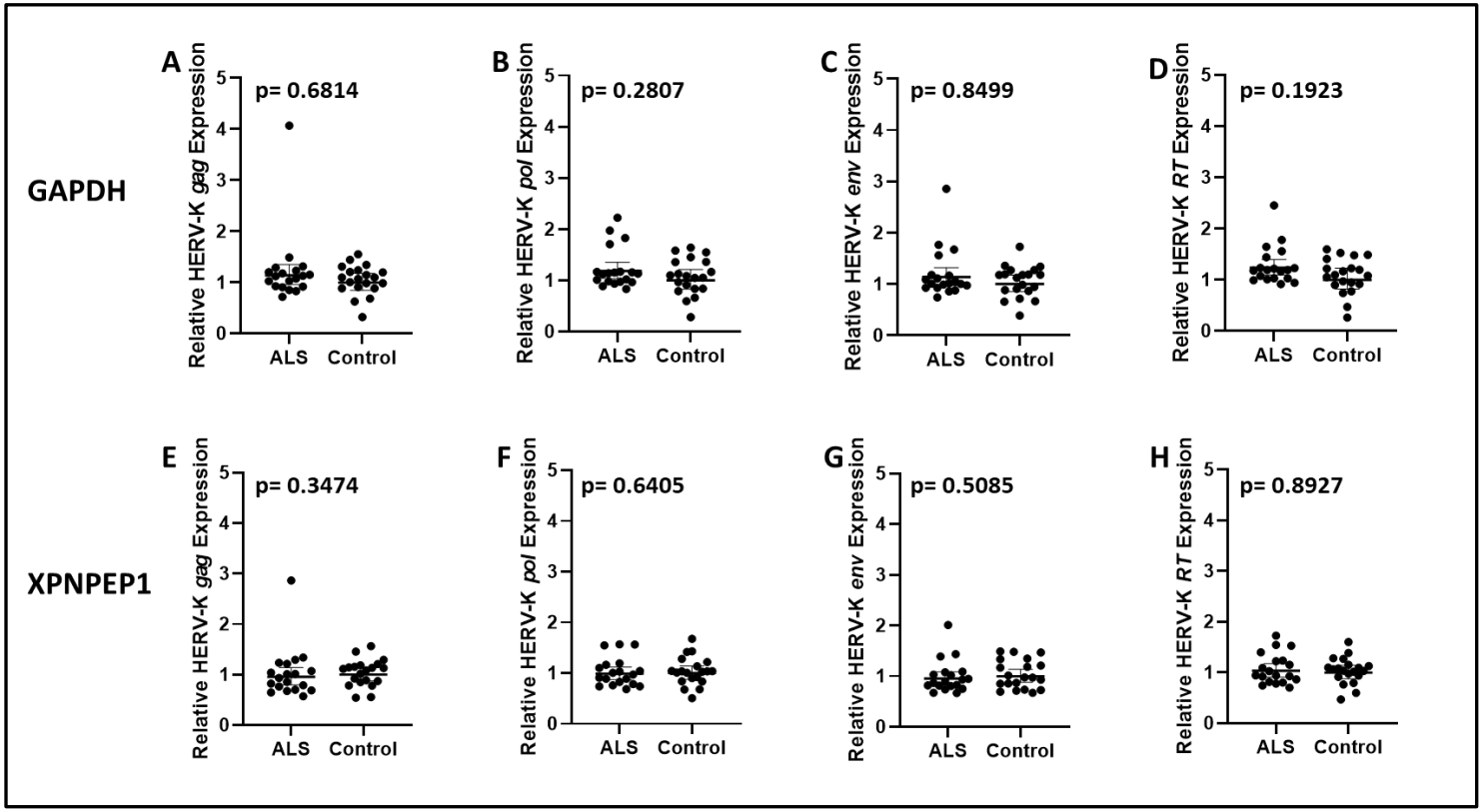


Figure 4.1. 2^-ΔΔCt Differential Expression levels for HERV-K *gag, pol, env* and *RT* gene transcripts in n=19 ALS and n=20 non-ALS Control Cases

The graphs displayed in the figure above show 2^-ΔΔCt Differential Expression levels of HERV-K *gag, pol, env* and *RT* transcripts in ALS and non-ALS control cases. The data is normalised either GAPDH (A-D) or XPNPEP1 (E-H), the horizontal lines and error bars represent the geometric mean for the data set and its 95% confidence interval. *p*-values for all gene transcripts are >0.05 indicating a lack of statistical significance. The outlier reading seen in ALS HERV-K *gag* (A & E) is sample A381/11 (derived from an ALS patient), the same as the highest relative expression seen in the HERV-K *env* gene target (normalised against GAPDH).

Expression levels of HERV-K *gag, pol and env* correlate well to one another in ALS samples, with comparisons showing similarity in expression across all ALS samples (Figure 4.2). In each case the *p*-values were less than p=0.05 whether the data was normalised using GAPDH or XPNPEP1 reference genes. The expression data from non-ALS control samples (Figure 4.3) shows similar data when normalised against GAPDH, with HERV-K transcripts showing significant p-values when compared to one-another. When non-ALS control samples were normalised against XPNPEP1 the comparisons between HERV-K *env* & HERV-K *gag* (p=0.0871) and HERV-K *pol* & HERV-K *gag* (p=0.0685) failed to show any significant result.

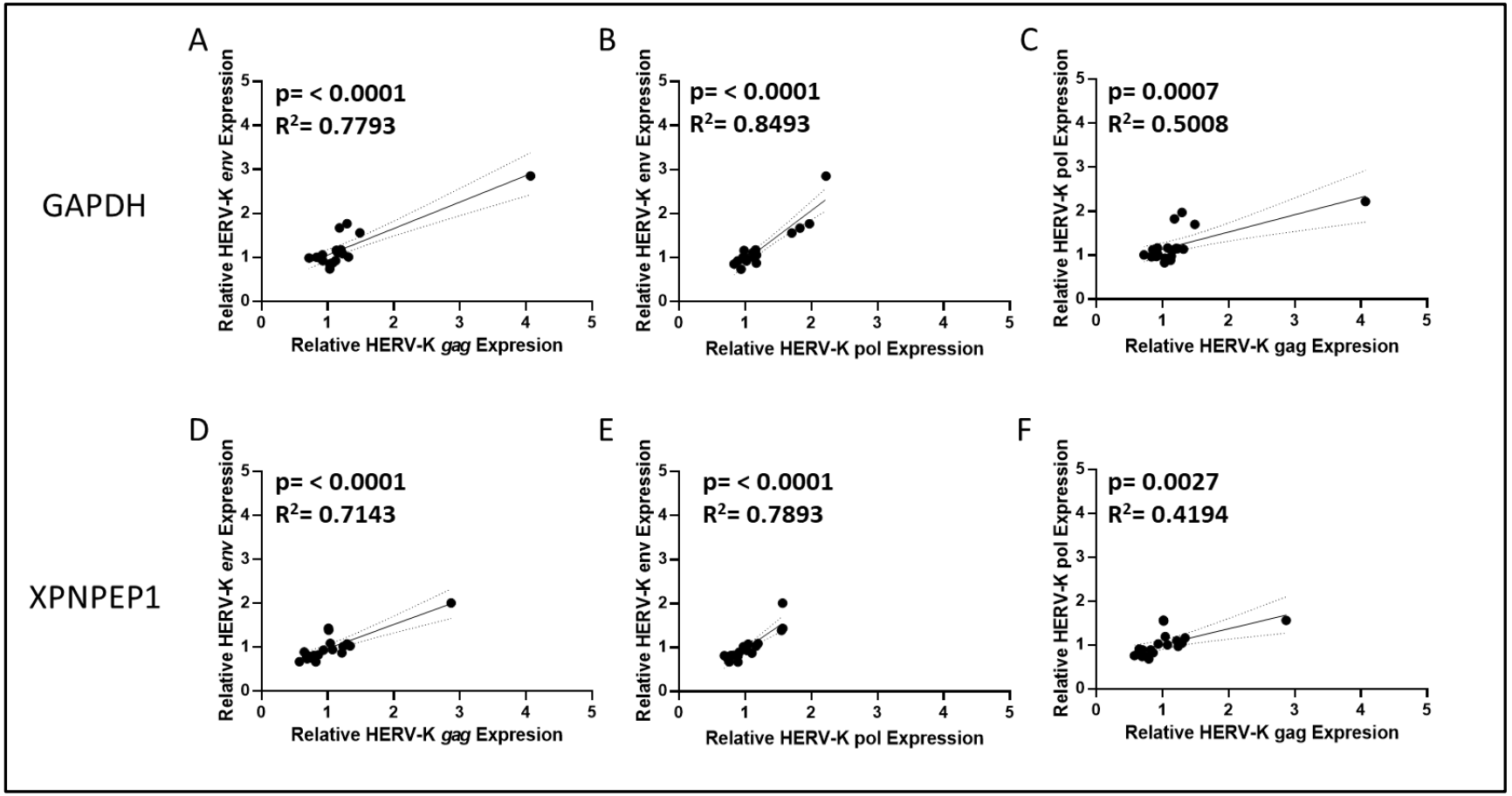


Figure 4.2. Graphs Displaying Correlations between HERV-K *gag, pol* and *env* transcripts differential expression in n=19 ALS Samples.
In the figure above HERV-K transcripts are compared for correlation between their relative expression levels. The HERV-K *gag, pol, env* and *RT* are normalised against GAPDH (A-C) or XPNPEP1 (D-F) with the R^2 and *p*-values calculated in GraphPad using its linear regression analysis. The outlier reading seen in HERV-K *gag* comparisons (B, C, E F) is sample A381/11, the same as the highest relative expression seen in the HERV-K *env* gene target.

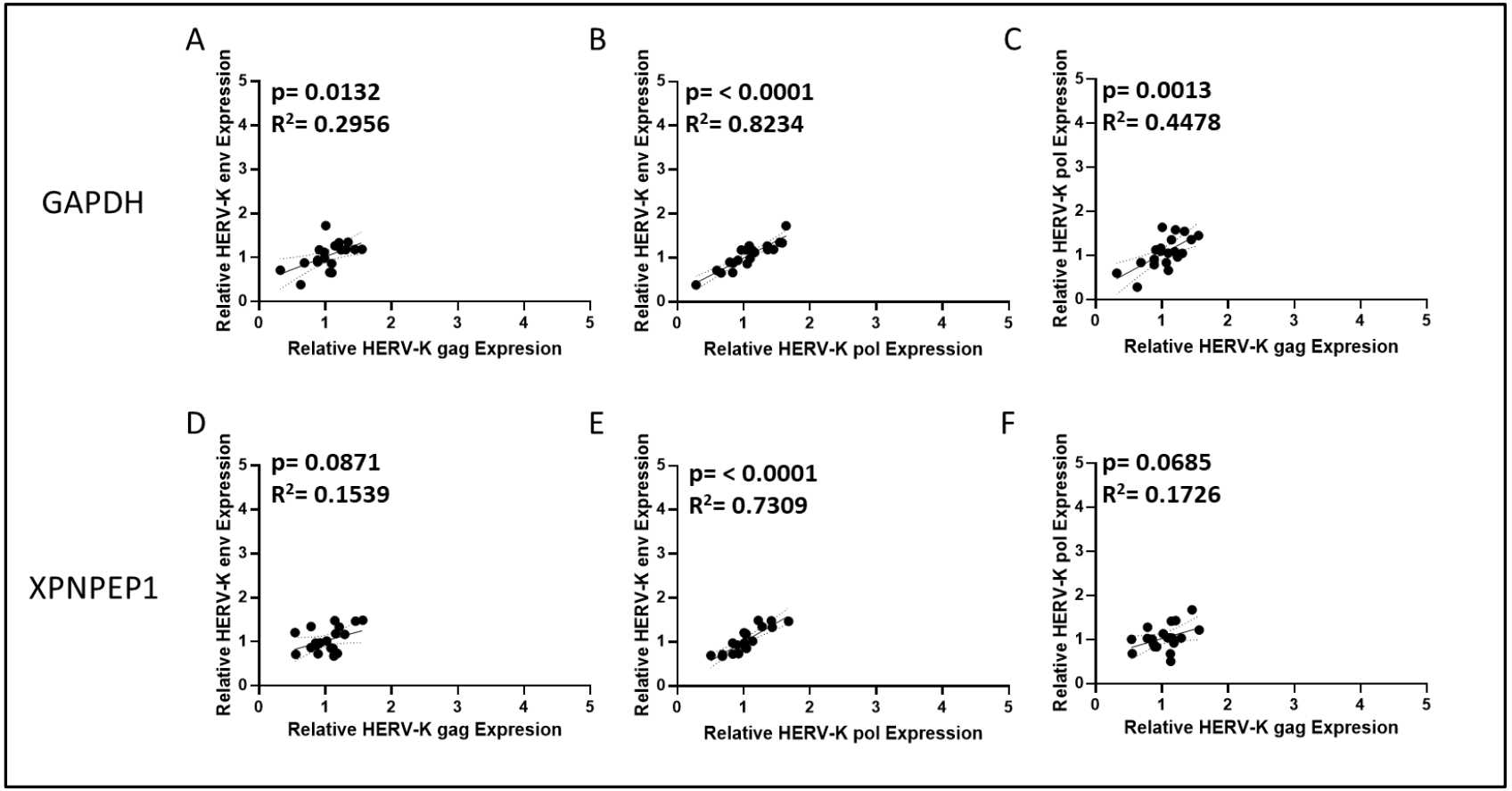


Figure 4.3. Graphs Displaying Correlations between HERV-K *gag, pol* and *env* transcripts differential expression in n=20 non-ALS Control Samples.

In the figure above HERV-K transcripts are compared for correlation between their relative expression levels. The HERV-K *gag, pol, env* and *RT* are normalised against GAPDH (A-C) or XPNPEP1 (D-F) with the R^2 and *p*-values calculated in GraphPad using its linear regression analysis.

The ALS and control samples were matched as closely as possible for age, gender and PMD however due to sample availability there were more female samples than males, with 14 females and 6 males for control cases and 13 females and 6 males for ALS cases. In addition, the mean age for the ALS group was 72 years and the mean age for the control group was 71 years. Post-mortem Delay (PMD) ranged from 3.5 hours to 71 hours with a mean PMD of 39.5 hours for the ALS group and a mean PMD of 42.6 hours for the control group.

The expression data for ALS and non-ALS controls was analysed separately to see if there was any statistically significant difference when the data is compared to PMD, Age at time of death, RIN and gender. When normalised to GAPDH there was a significant negative correlation with HERV-K *pol* to PMD in control tissue (Supplementary Figure S46). Similar results were seen when the data was normalised against XPNPEP1 which showed a significant negative correlation in both HERV-K *pol* & HERV-K *RT* gene targets (Supplementary Figure S46). The only statistically significant result in ALS patient tissue for PMD was a positive correlation between HERV-K *gag* gene transcripts when normalised against XPNPEP1. Other notable significant differences in the expression data when normalised against GAPDH was a significant positive correlation between HERV-K *pol* & *env* transcripts and increasing age of patients at time of death in ALS tissue and a positive correlation between increasing RIN and HERV-K *pol* transcripts in non-ALS control tissue. When normalised against XPNPEP1 there was a significant negative correlation between HERV-K *pol* & *RT* transcripts and RIN in ALS tissue and no significant correlation in the comparison between HERV-K *pol* & *gag* transcripts in non-ALS control tissue.

4.2.3. HERV-W env RNA expression in ALS and non-ALS derived premotor cortex brain tissue obtained at post-mortem.

Utilising the same n=19 ALS and n=20 non-ALS controls as the HERV-K assays HERV-W *env* transcript expression between ALS and non-ALS controls was measured and normalised to GAPDH or XPNPEP1 reference genes (Figure 4.4). The geometric mean for ALS when normalised to GAPDH was 1.135 (p=0.989) and was slightly lower when normalised to XPNPEP1 at 0.952 (p=0.142). The p-values for the comparisons between ALS and non-ALS control expression data are given in Figure 4.3 (p=0.9813) when normalised against GAPDH and p=0.1404 when normalised against XPNPEP1). This represents no statistically significant difference in relative gene expression of HERV-W *env* between ALS and non-ALS controls.

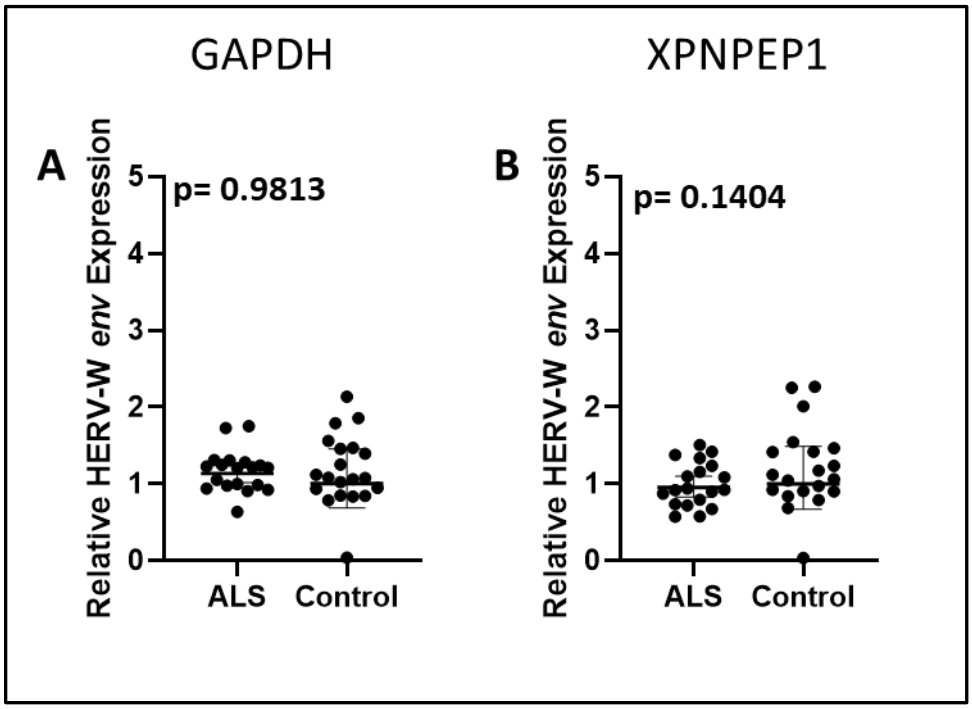


Figure 4.4. 2^-ΔΔCt Differential Expression levels for HERV-W *env* transcript in premotor cortical brain tissue derived from n=19 ALS and n=20 non-ALS Control Cases.
Displayed in the figure above are relative expression data for HERV-W *env* transcript normalised to A) GAPDH and B) XPNPEP1. The *p*-values for differences between the groups when normalised against GAPDH were *p*=0.9813 and when normalised against XPNPEP1 *p*=0.1404. Horizontal black lines in the centre of the sample distribution represent the geometric mean of the sample set along with error bars representing a 95% confidence in the group mean.

Both HERV-K and HERV-W expression assays mentioned above, have been repeated showing no significant difference in relative gene expression of HERV-K and HERV-W transcripts between ALS and non-ALS control tissue (data not shown).

When measuring the effect of PMD, Age and RIN on the expression data in ALS and non-ALS control samples there were significant differences in both ALS and non-ALS control samples. When normalised against GAPDH there was a significant negative correlation between PMD and expression in control tissue (p= 0.0089), a significant positive correlation between HERV-W expression and increasing age in ALS samples (p= 0.0263) and a negative correlation in control samples (p= 0.0378) (Supplementary Figures S49-S52). When normalised against XPNPEP1 there was a similar negative correlation to GAPDH between expression levels and increasing age in control samples (p= 0.0137) and a significant negative correlation to RIN in ALS samples (p= 0.0001) (Supplementary Figures S49-S52).

4.2.4 Utilising the Pfaffl Method for Analysis of HERV-W *env*, HERV-K *gag, pol, env* & *RT* Expression Data from ALS and Non-ALS Premotor Cortex Tissue Samples.

The Pfaffl method for analysing expression data uses a novel mathematical model to derive relative quantification values from expression data. This method is a considerable improvement on the 2^-ΔΔCt model as it considers the relative amplification efficiencies of the primers for each gene target in its equation. This method also utilises a geometric mean of multiple reference genes Relative Quantification (RQ) values to generate its relative expression values.

Table 4.3. Geometric Mean of HERV-W, HERV-K *gag, pol, env* & *RT* Relative Expression in ALS and non-ALS control cases, using the Pfaffl gene normalisation method to a geometric mean of derived GAPDH or XPNPEP1 expression data.

The table displays the geometric means of the Pfaffl derived differential expression values for n=19 ALS and n=20 non-ALS control premotor cortex brain tissue samples against each of the HERV-K gene targets used in the RT-qPCR expression assay, which were normalised against GAPDH and XPNPEP1 and taking into account the amplification efficiency of primer pairs.

	HERV-K *gag*	HERV-K *pol*	HERV-K *env*	HERV-K RT	HERV-W *env*
ALS	1.131	1.095	1.039	1.116	0.890
Control	1.030	1.040	1.029	1.039	1.039
p-value	0.8491	0.8729	0.3399	0.9062	0.2018
Statistical Significance	NS	NS	NS	NS	NS

A slight difference between the two methods can be observed in the geometric mean of the non-ALS control expression data which does not equal 1 (Table 4.2 & Table 4.3), this value occurs in the ΔΔCt method as all primers are assumed to have the same amplification efficiency across disease and control tissues. Also similar to the 2^-ΔΔCt method there is a slight negative correlation between relative HERV-W expression and increasing RNA Integrity (RIN) values (p=0.0004, Supplementary Figure S61).

When both the HERV-K and HERV-W RT-qPCR data was analysed using the Pfaffl method for relative gene expression this showed no statistically significant differences in expression for all gene targets between ALS and non-ALS controls with all p-values above 0.3 (Figure 4.5). However, there are significant results when ALS and non-ALS control sample expression data is analysed for differences in PMD, Age and correlations between transcripts. There was a significant positive correlation in PMD to HERV-K *gag* transcripts in ALS tissue (p= 0.0385) and a negative correlation to PMD with HERV-K *gag* (p= 0.0288) *pol* (p= 0.0154) and *RT* (p= 0.0246) in non-ALS control samples (Supplementary Figure S54). Also observed in the expression data was a significant positive correlation between increasing Age and expression of HERV-K *env* (p= 0.0378) and *pol* (p= 0.0303) transcripts in ALS samples and no significant correlation between HERV-K *env* and HERV-K *gag* expression data in non-ALS controls (p= 0.1235) (Supplementary Figure S55).

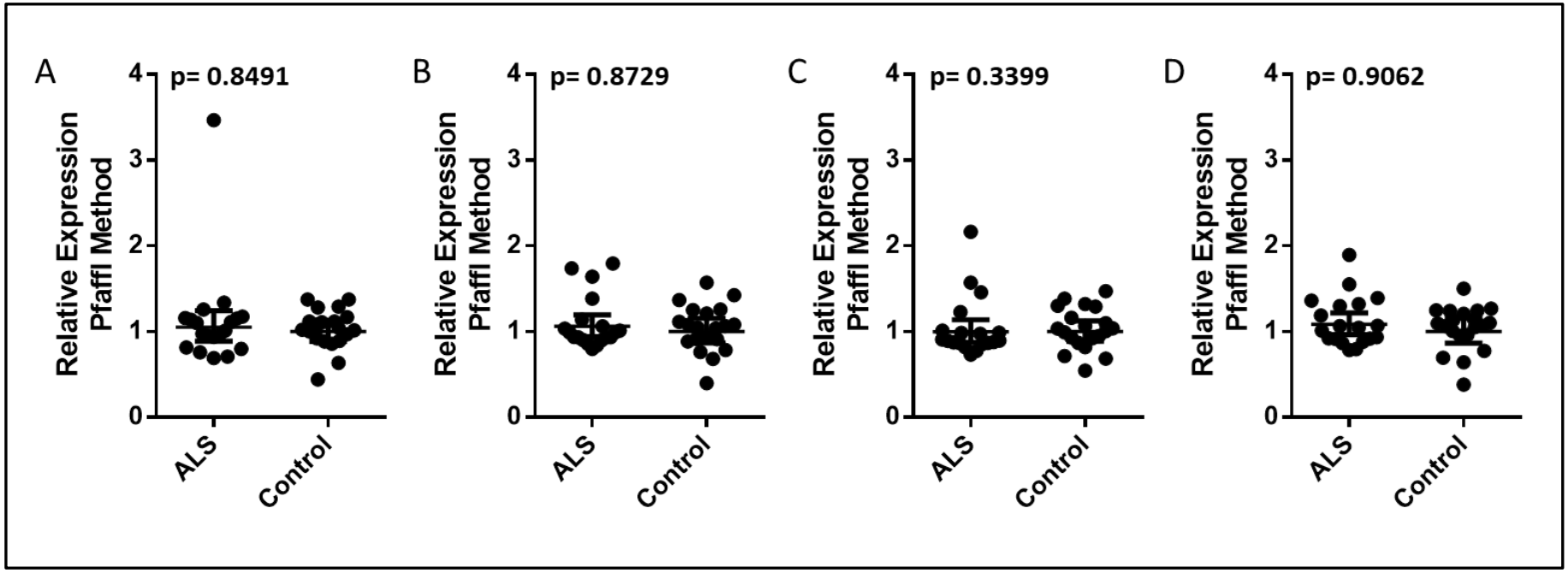


Figure 4.5. Differential Expression Calculated by Pfaffl Method for HERV-K *gag, pol, env* and *RT* gene transcripts in n=19 ALS and n=20 non-ALS Control Cases

The graphs displayed in the figure above show Pfaffl Differential Expression of A) HERV-K *gag*, B) HERV-K *pol*, C) HERV-K *env* and D) HERV-K *RT* and transcripts in ALS and non-ALS control cases. The data is normalised to a geometric mean of GAPDH and XPNPEP1 expression values with the horizontal lines and error bars representing the geometric mean for the data set and its 95% confidence interval. *p*-values for all gene transcripts are >0.05 indicating a lack of statistical significance. Outlier value seen in HERV-K *gag* ALS is sample A381/11. Data for HERV-W is located in Supplementary Information Figure S57.

4.2.5 Relative Expression of HERV-W *env,* HERV-K *gag, pol, env* and RT in ALS and No-Cancer Controls

Numerous conditions such as cancer, neurodegenerative and autoimmune diseases have long been associated with the differential regulation of HERV families. These conditions have the potential to distort differential expression calculations due to the background upregulation of HERVs. As several of the samples in the HERV-K and HERV-W assays had conditions related to these associated conditions the relevant control sample data was removed from the analysis and relative expression levels recalculated using both ΔΔCt and Pfaffl methods.

The samples removed from the analysis were A261/12, A308/09, A346/10, A012/12, A407/13, A153/06, A273/12, A308/14, A319/14, A103/17, A132/14 and A346/10 (relevant conditions listed in Table 2.5 in the Materials Section) due to having cancer, both metastatic and primary tumour and one patient sample presenting with grade 2 Alzheimer's. The table below (Table 4.4) lists the differences in geometric mean for the gene targets and their respective p-values. As shown in the table the respective p-values are all above the p=0.05 cut-off for statistically significant differences in expression between ALS and no-Cancer control samples.

Table 4.4. Geometric Mean of HERV-K *gag, pol, env* & *RT* Relative Expression in ALS and no-cancer controls, Normalised to GAPDH or XPNPEP1 Reference Genes.
The table displays the geometric means of the 2^-ΔΔCt differential expression value**s** n=19 ALS and n=10 non-ALS control samples for each of the HERV-K gene targets used in the RT-qPCR expression assay. These were normalised to 2 separate reference genes, GAPDH (top) and XPNPEP1 (Bottom).

	GAPDH				
	HERV-K *gag*	HERV-K *pol*	HERV-K *env*	HERV-K RT	HERV-W *env*
ALS	1.041	1.098	1.035	1.121	0.944
Control	1	1	1	1	1
p value	0.9252	0.7174	0.9054	0.3212	0.8524
Statistical Significance	NS	NS	NS	NS	NS
	XPNPEP1				
	HERV-K *gag*	HERV-K *pol*	HERV-K *env*	HERV-K RT	HERV-W *env*
ALS	0.863	0.910	0.857	0.929	0.782
Control	1	1	1	1	1
p value	0.1040	0.2856	0.1405	0.4011	0.1794
Statistical Significance	NS	NS	NS	NS	NS

Figure 4.6 displays graphs for the comparison of relative expression data between ALS and Controls along with their p-values for statistical significance of differences between the geometric means of the two groups. As shown in the figure there is no statistically significant difference in the ALS vs no-cancer control samples. An additional test performed to see if cancer controls had any effect on the measured differential expression of transcripts was performed by comparing the differential expression of cancer controls to non-cancer controls (Supplementary Figure S273). This test showed that while the relative HERV transcript expression of cancer controls was lower than normal controls the difference was not statistically significant. When comparing other clinically relevant data for the samples such as differences in expression between Male & Female, Postmortem Delay, Age and RNA integrity values between ALS and no-cancer controls there were no statistically significant differences or correlations between the expression data (Supplementary Figures S62-S69). Additionally, aside from the comparison of HERV-K *env* and *pol* transcripts in the n=8 non-ALS control samples there were only significant correlations between HERV-K transcripts in ALS patient tissue (Supplementary Figures S70 & S71).

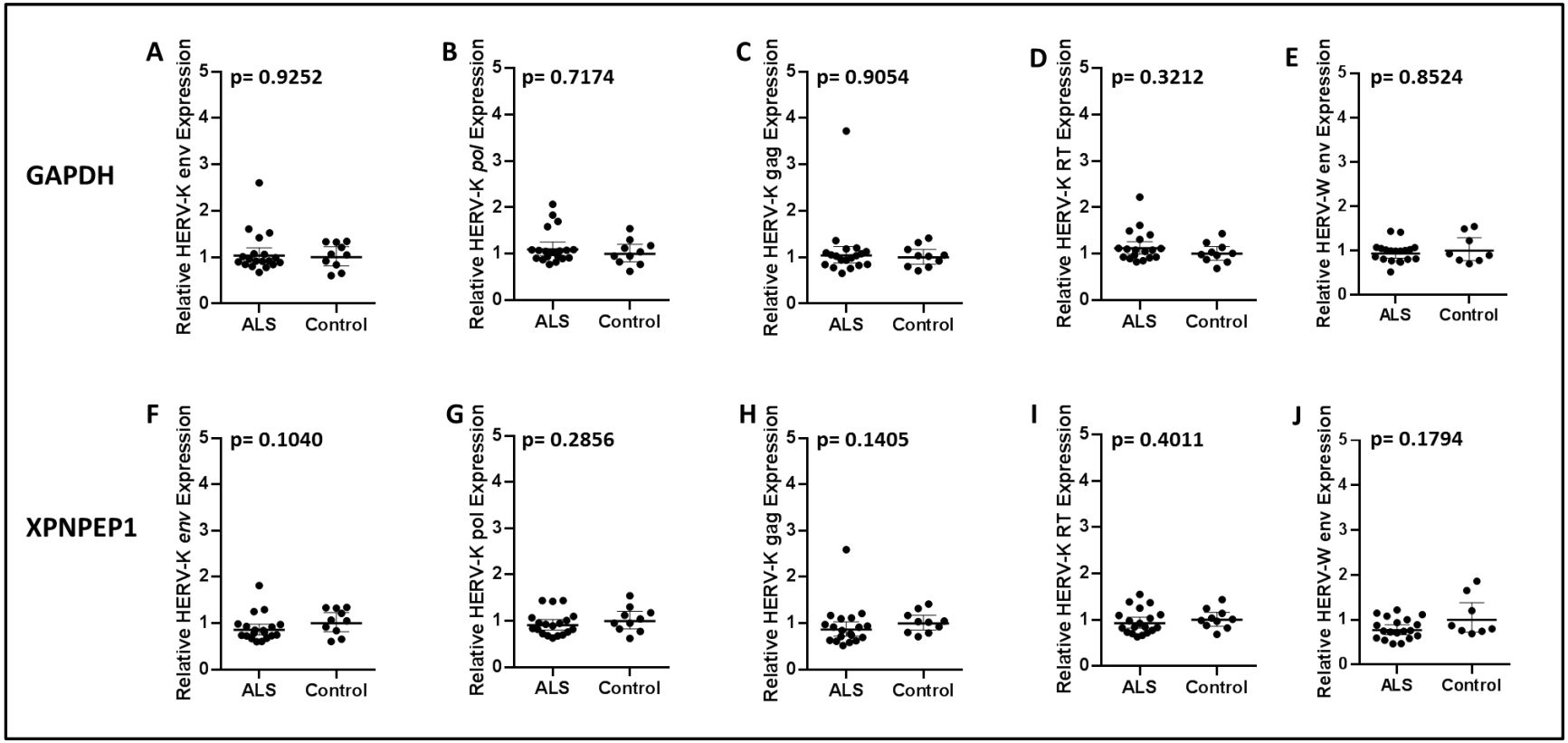


Figure 4.6. 2^-ΔΔCt Differential Expression levels for HERV-W *env* HERV-K *gag, pol, env* and *RT* gene transcripts in n=19 ALS and n=10 no-Cancer Control Cases

The graphs displayed in the figure above show 2^-ΔΔCt Differential Expression levels of A) & F) HERV-K *gag,* B) & G) HERV-K *pol,* C) & H) HERV-K *env* D) & I) HERV-K *RT* transcripts and E) & J) HERV-W *env* in ALS and no-cancer control cases. The data is normalised either against GAPDH (A-E) or XPNPEP1 (F-J), the horizontal lines and error bars represent the geometric mean for the data set and its 95% confidence interval. *p*-values for all gene transcripts are >0.05 indicating a lack of statistical significance. The outlier reading seen in ALS HERV-K *gag* (A & F) is sample A381/11, the same as the highest relative expression seen in the HERV-K *env* gene target.

The data was then analysed to a geometric mean of GAPDH and XPNPEP1 reference genes using the Pfaffl method for relative expression. The summary table of the differences between ALS and no-cancer controls using the Pfaffl method is given in Table 4.5 which displays no significant differences ($p > 0.05$) between ALS and no-cancer control samples.

Table 4.5. Geometric Mean of HERV-K *gag, pol, env* & *RT* Relative Expression in ALS and No-Cancer control cases, using the Pfaffl gene normalisation method to a geometric mean of derived GAPDH or XPNPEP1 expression data.
The table displays the geometric means of the Pfaffl derived differential expression values for n=19 ALS and n=10 no-cancer control premotor cortex brain tissue samples against each of the HERV-K gene targets used in the RT-qPCR expression assay, which were normalised against GAPDH and XPNPEP1 and taking into account the amplification efficiency of primer pairs.

	HERV-K gag	HERV-K pol	HERV-K env	HERV-K RT	HERV-W *env*
ALS	1.026	1.020	1.039	1.022	0.867
Control	1.012	1.025	1.024	1.014	1.041
p-value	0.3353	0.4769	0.1232	0.6651	0.4176
Statistical Significance	NS	NS	NS	NS	NS

This data is also displayed in Figure 4.9 which shows the GraphPad PRISM graphs for the data displayed in Table 4.5. This identifies a single sample as an outlier in HERV-K *gag* (Figure 4.7A) and the highest expressed in HERV-K *env* (Figure 4.7C) which corresponds to ALS sample A381/11. When observing the data for comparisons to Male & Female, Postmortem Delay (PMD), Age and RNA integrity values between ALS and no-cancer controls for the Pfaffl method some significant data was revealed. The no-cancer controls for HERV-K *pol* (p=0.0437), HERV-K *env* (p=0.0436) and HERV-W *env* (p=0.0107) gene targets showed a significant negative correlation between increasing PMD and relative expression of gene transcripts. Similar to the comparison of Age and relative expression in the original assay there were significant positive correlation in the HERV-K *pol* (p=0.0304) and HERV-K *env* (p=0.0377) gene targets. Finally, there was a similar significant p-value for the negative correlation between increasing RNA quality and HERV-W *env* (p=0.0004) relative expression to the original assay (Supplementary Figures S72-S80).

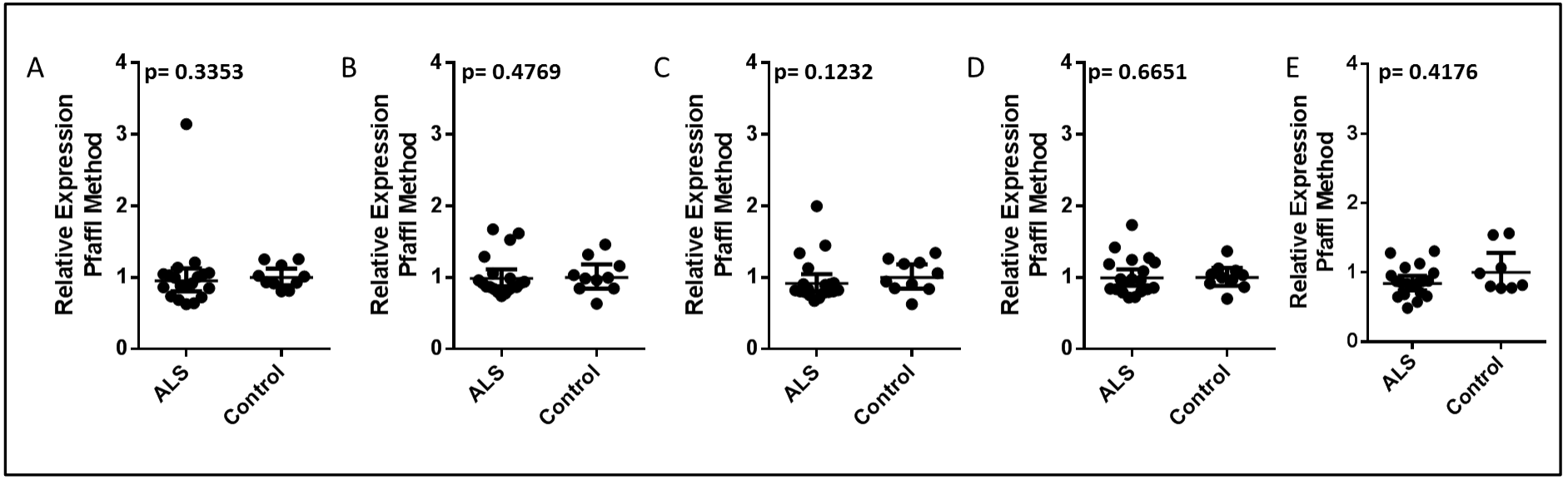


Figure 4.7. Pfaffl Relative Expression levels for HERV-W *env* HERV-K *gag, pol, env* and *RT* gene transcripts in n=19 ALS and n=10 no-Cancer Control Cases using Pfaffl.

The graphs displayed in the figure above show relative expression levels of A) HERV-K *gag,* B) HERV-K *pol,* C) HERV-K *env* D) HERV-K *RT* transcripts and E) HERV-W *env* in ALS and no-cancer control cases using Pfaffl. The data is normalised to a geometric mean of XPNPEP1 and GAPDH reference genes, the horizontal lines and error bars represent the geometric mean for the data set and its 95% confidence interval. *p*-values for all gene transcripts are >0.05 indicating a lack of statistical significance. The outlier reading seen in ALS HERV-K *gag* is sample A381/11, the same as the highest relative expression seen in the HERV-K *env* gene target.

4.2.6 Relative Expression of BCL11b, TDP-43, HERV-K env & RT Using Post-Mortem Premotor Cortex Brain Tissue from ALS and No-Cancer Controls.

As several of the Non-ALS controls used in the previous experiment came from patients suffering from clinical conditions such as cancer, in which HERV expression has been shown by other researchers to be more highly expressed compared to non-cancer controls, we therefore removed all cancer controls from further analysis and made up the number of non-cancer controls from additional premotor cortex brain tissue we had obtained from the MRC neurodegenerative disease brain bank for the Garson *et.al* (2019) paper (see Table 2.6 in the methods section).

These samples were assessed for the presence of residual gDNA by RNA spike of RT-qPCR assay similar to the assays detailed in section 4.2.2. This assay used n=19 ALS and n=17 no cancer controls from which 1 ALS and 3 control samples were removed due to the presence of gDNA not eliminated during the DNase I step of RNA extraction. The remaining samples were used on RT-qPCR assays to assess the expression of HERV-K *env*, HERV-K *RT*, BCL11b and TDP-43.

Tables 4.6 and 4.7 show the results of the 2^-ΔΔCt differential expression calculation for HERV-K *env*, HERV-K *RT*, BCL11b and TDP-43 gene targets. When normalised to either GAPDH or XPNPEP1 reference genes there was no significant difference between the geometric means of ALS and non-ALS controls. As shown in the tables, the differences between ALS and non-ALS controls were similar whether normalised against GAPDH or XPNPEP1.

Table 4.6. Geometric Mean of BCL11b, TDP-43, HERV-K *env* & *RT* Relative Expression in ALS and No-Cancer control cases, Normalised to GAPDH Reference Gene.
The table displays the geometric means of the 2^-ΔΔCt differential expression values n=18 ALS and n=14 no-cancer, control samples for each of the HERV-K gene targets used in the RT-qPCR expression assay. These were normalised to the GAPDH reference gene.

	HERV-K *env*	HERV-K *RT*	TDP-43	BCL11b
ALS	1.018	1.204	1.082	0.994
Control	1.000	1.000	1.000	1.000
p value	0.7011	0.0894	0.6062	0.9754
Statistical Significance	NS	NS	NS	NS

Table 4.7. Geometric Mean of BCL11b, TDP-43, HERV-K *env* & *RT* Relative Expression in ALS and non-ALS control cases, Normalised to XPNPEP1 Reference Gene.
The table displays the geometric means of the 2^-ΔΔCt differential expression values n=18 ALS and n=14 no-cancer control samples for each of the HERV-K gene targets used in the RT-qPCR expression assay. These were normalised to the XPNPEP1 reference gene.

	HERV-K *env*	HERV-K *RT*	TDP-43	BCL11b
ALS	1.016	1.202	1.080	1.000
Control	1.000	1.000	1.000	1.000
p value	0.9178	0.0537	0.5526	0.8214
Statistical Significance	NS	NS	NS	NS

The graphs in Figure 4.8 show the comparison of expression data between n=18 ALS and n=14 non-ALS controls for HERV-K, TDP-43 and BCL11b gene targets. The data shows that there is no statistically significant difference in the expression data between ALS and non-ALS control samples, with all p-values in excess of 0.05. The closest result to this cut-off is HERV-K *RT* when normalised against GAPDH with a p-value of 0.0894 and when normalised against XPNPEP1 with a p-value of 0.0537.

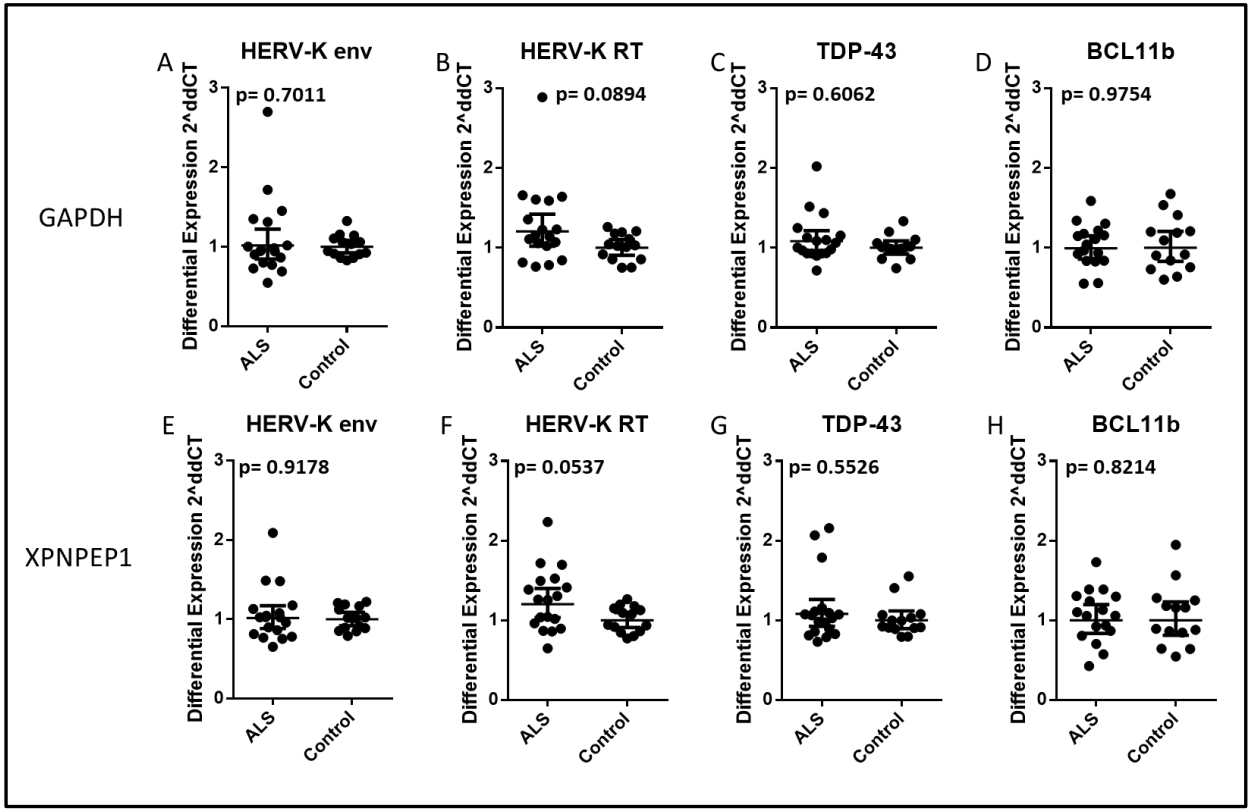


Figure 4.8. 2^-ΔΔCt Differential Expression levels for HERV-K *env* and *RT,* TDP-43 and BCL11b gene transcripts in n=18 ALS and n=14 No-Cancer Controls.

The graphs displayed in the figure above show 2^-ΔΔCt Differential Expression levels of HERV-K env and RT, TDP-43 and BCL11b transcripts in ALS and non-ALS control cases. The data is normalised either GAPDH (A-D) or XPNPEP1 (E-H), the horizontal lines and error bars represent the geometric mean for the data set and its 95% confidence interval. *p*-values for all gene transcripts are >0.05 indicating a lack of statistical significance. The outlier reading seen in ALS HERV-K *env* (A & E) & HERV-K *RT* (B & F) is sample A381/11, the same as the highest relative expression seen in the previous assays (Sections 4.2.4 & 4.2.5).

When correlating expression data between HERV-K *env* and HERV-K RT transcripts they were well correlated in ALS patient samples (Figure 4.9), showing a p-value of below 0.0001 whether the data was normalised to GAPDH or XPNPEP1. In control samples HERV-K transcripts showed no significant correlation to each other with the comparison between HERV-K *env* and HERV-K *RT* showing a p-value of 0.5728 when normalised against GAPDH and 0.5262 when normalised against XPNPEP1 (Figure 4.10). The proportion of Male to Female subjects were similar to the previous assay as well as the mean ages for the ALS and control groups due to sample availability. Additionally, the correlation between HERV-K *RT* and BCL11b transcripts was analysed to see if there was any similarity in correlation to the HERV-K *env* comparisons. Unlike the HERV-K *env* comparisons HERV-K *RT* reported significant positive correlation to TDP-43 in ALS samples across both GAPDH and XPNPEP1 reference genes. In control tissue when normalised against XPNPEP1 HERV-K *RT* was significantly positively correlated to both TDP-43 and BCL11b but no significant results were seen when normalised against GAPDH in no-cancer controls.

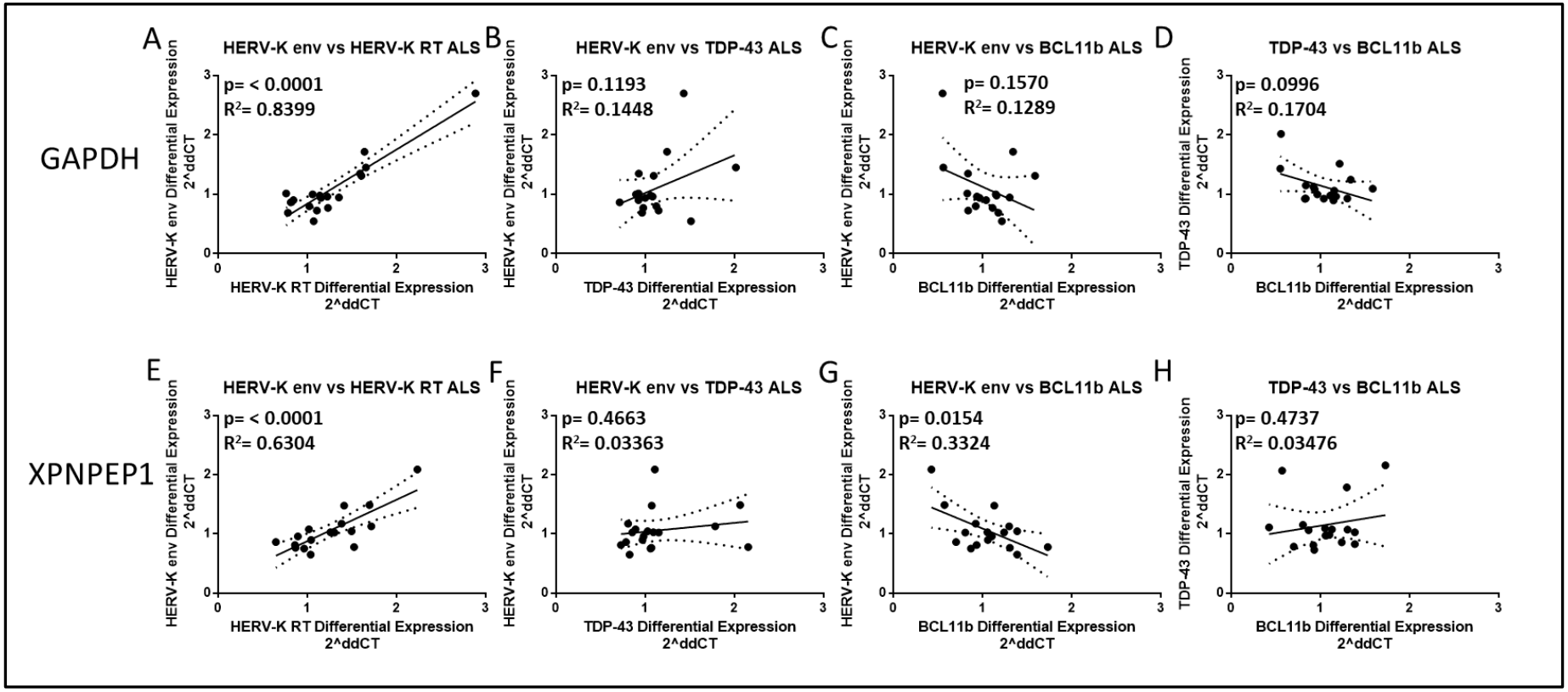


Figure 4.9. Graphs Displaying Correlations between HERV-K *env* and *RT* transcripts differential expression in n=18 ALS Samples.

In the figure above HERV-K transcripts are compared for correlation between their relative expression levels. The HERV-K *env* and *RT* are normalised against GAPDH (A-D) or XPNPEP1 (E-F) with the R^2 and *p*-values calculated in GraphPad using its linear regression analysis. Additionally, comparisons between HERV-K *env* and transcriptional regulators BCL11b and TDP-43 are also displayed. The outlier reading seen in ALS HERV-K comparisons (A, B, C, E, F & G) is sample A381/11.

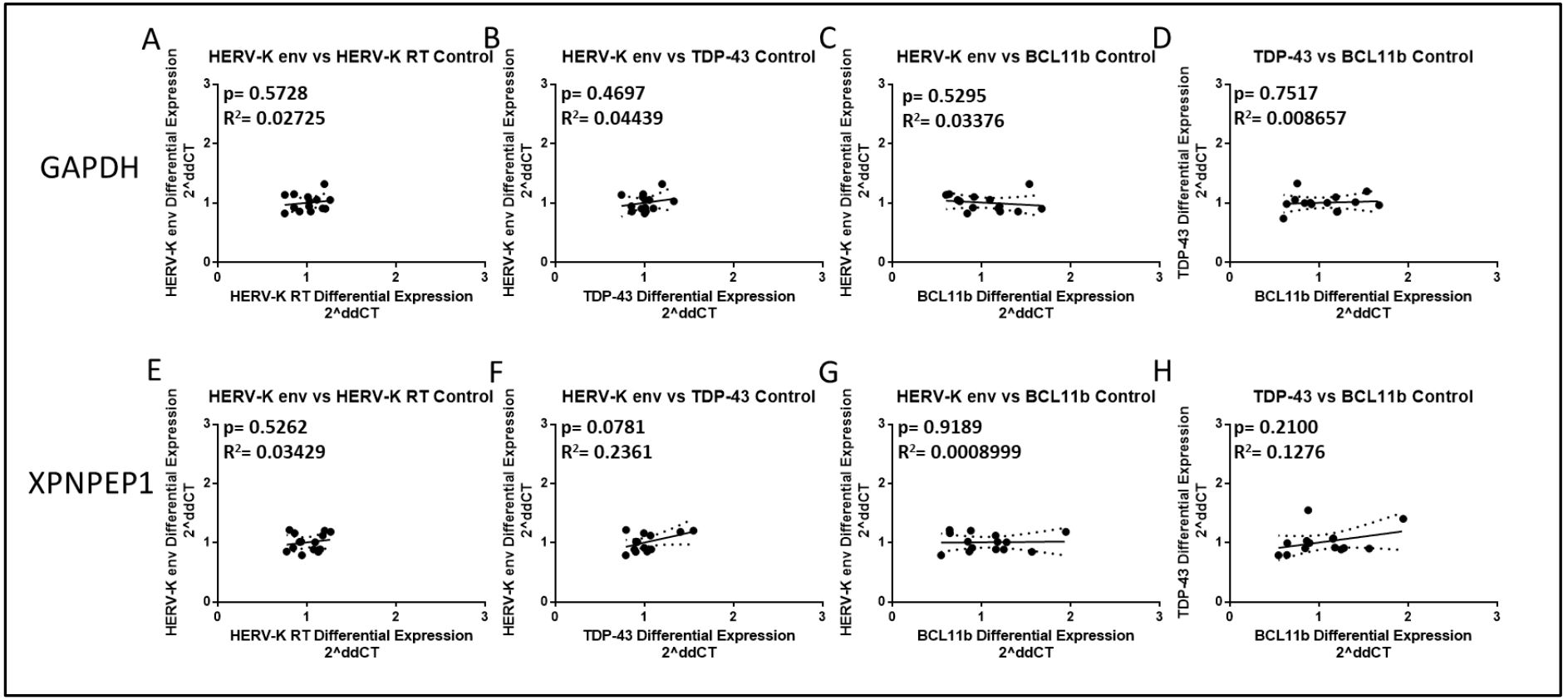


Figure 4.10. Graphs Displaying Correlations between HERV-K *env* and *RT* transcripts differential expression in n=14 No-Cancer Control Samples.

In the figure above HERV-K transcripts are compared for correlation between their relative expression levels. The HERV-K *env* and *RT* are normalised against GAPDH (A-D) or XPNPEP1 (E-F) with the R^2 and *p*-values calculated in GraphPad using its linear regression analysis. Additionally, comparisons between HERV-K *env* and transcriptional regulators BCL11b and TDP-43 are also displayed.

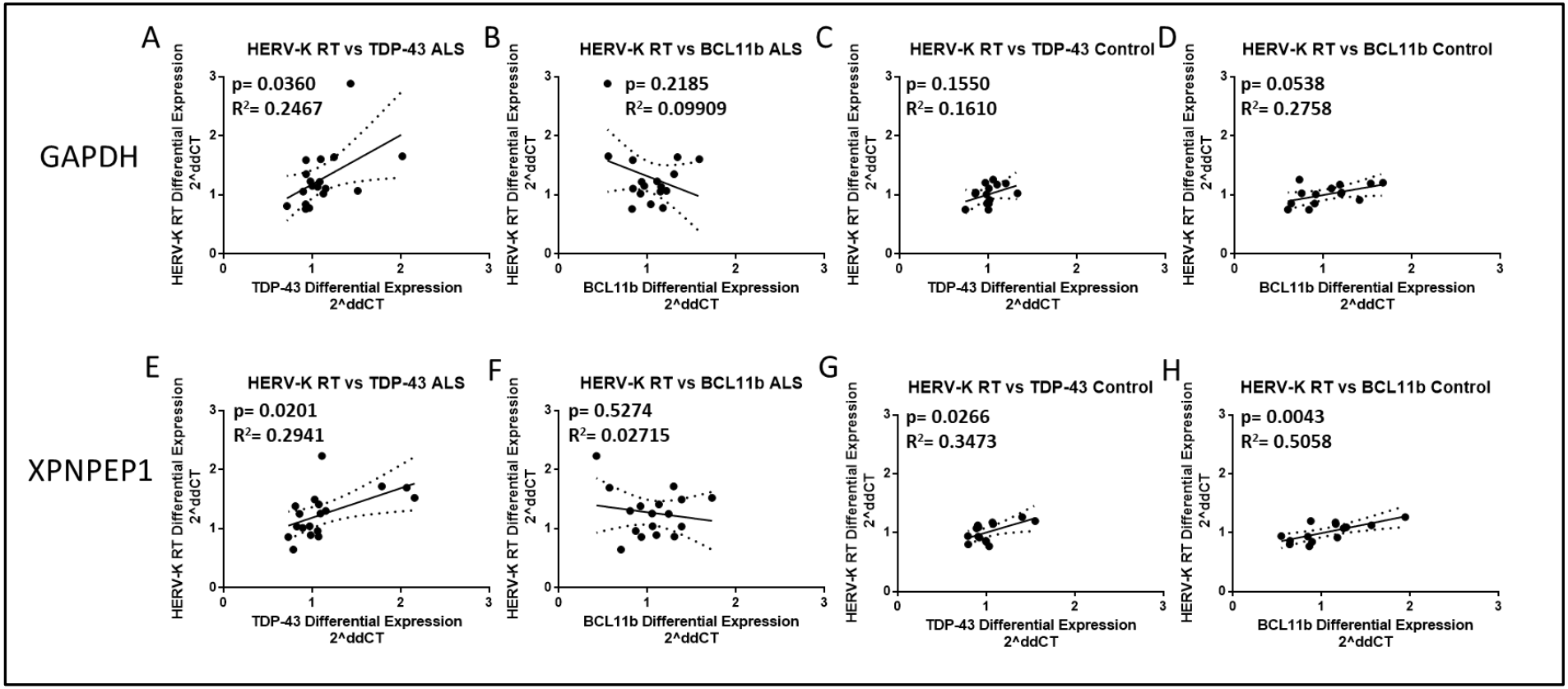


Figure 4.11. Graphs Displaying Correlations between HERV-K *RT,* BCL11b & TDP-43 differential expression in n=18 ALS and n=14 No-Cancer Control Samples.

In the figure above HERV-K *RT* transcripts and the transcriptional modifiers BCL11b and TDP-43 are compared for correlation between their relative expression levels. The HERV-K *RT,* BCL11b and TDP-43 are normalised against GAPDH (A-D) or XPNPEP1 (E-H) with the R^2 and *p*-values calculated in GraphPad using its linear regression analysis. These results are subdivided into ALS (A&B, E&F) and non-ALS control samples (C&D and G&H).

When analysing for significant correlations between expression data and PMD, Age or RIN in ALS and no-cancer controls there was only a single significant difference in the geometric means of Male and Female expression data in HERV-K *RT* (p= 0.0120) and TDP-43 (p= 0.0190). All other comparisons between gene target expression data and PMD, Age and RIN values showed no statistically significant differences (Supplementary Figures S82-S90).

4.2.7 Utilising the Pfaffl Method for Analysis of Relative Expression of BCL11b, TDP-43, HERV-K env & RT Using Post-Mortem Premotor Cortex Brain Tissue from ALS and non-ALS, non HERV Associated, Controls.

When the RT-qPCR assay data for HERV-K *env*, HERV-K *RT*, BCL11b and TDP-43 is normalised against a geometric mean of GAPDH and XPNPEP1 using the Pfaffl gene quantification method there was no significant difference between ALS and no-cancer control post-mortem premotor cortex brain tissue samples (Table 4.8, Figure 4.12).

Geometric means for ALS and control data shown in Table 4.8 were derived from the Pfaffl method calculations done in excel while the p-values were calculated in GraphPad PRISM 8.0 using the Mann-Whitney t-test for non-parametric data sets. As shown in the table there is little difference between the geometric means of ALS and no-cancer control samples, the largest being HERV-K *RT* (0.212 difference between means).

Table 4.8. Geometric Mean of BCL11b, TDP-43, HERV-K *env* & *RT* Relative Expression in ALS and non-ALS control cases, using the Pfaffl gene normalisation method to a geometric mean of derived GAPDH or XPNPEP1 expression data.

The table displays the geometric means of the Pfaffl derived differential expression values for n=18 ALS and n=14 non-ALS control samples against each of the gene targets used in the RT-qPCR expression assay. These were normalised to a geometric mean of GAPDH and XPNPEP1 taking into account the amplification efficiency of primer pairs.

	HERV-K *env*	HERV-K *RT*	TDP-43	BCL11b
ALS	1.039	1.223	1.099	1.033
Control	1.008	1.011	1.012	1.048
p-value	0.4935	0.0930	> 0.9999	0.9754
Statistical Significance	NS	NS	NS	NS

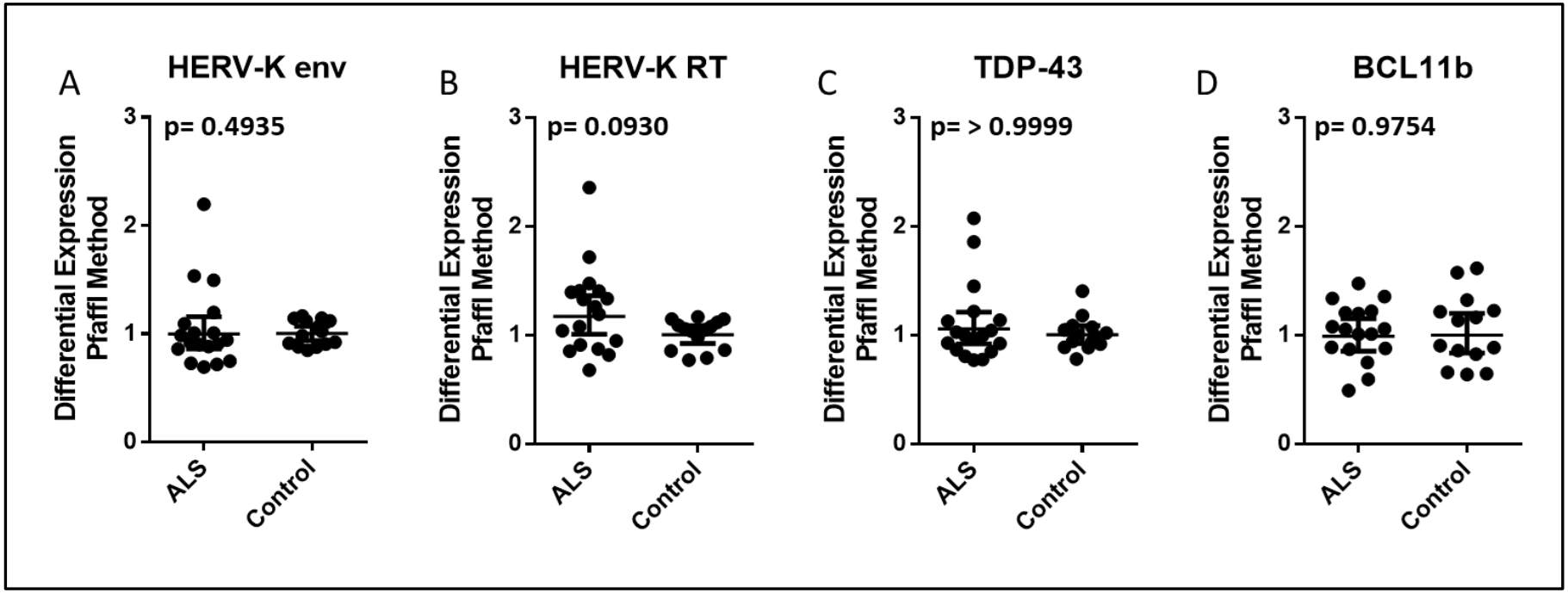


Figure 4.12. Differential Expression levels for HERV-K *env* and *RT,* TDP-43, BCL11b as Calculated Using the Pfaffl Method Compared Between ALS and No-Cancer Controls.

The graphs displayed in the figure above show Pfaffl Differential Expression levels of HERV-K *env* and RT, TDP-43 and BCL11b transcripts in ALS and non-ALS control cases. The expression data is normalised against a geometric mean of GAPDH and XPNPEP1 reference genes with the R^2 and *p*-values calculated in GraphPad using its linear regression analysis. The outlier reading seen in ALS HERV-K *env* & *RT* (A & B) is sample A381/11, the same as the highest relative expression seen in the TDP-43 gene target.

When correlating gene target expression against HERV-K *env* and sample RIN, patient PMD and Age at time of death there were similar significant results as detailed in section 4.2.6. This includes a significant difference between the geometric means of male and female control samples for HERV-K *RT* (p= 0.0445) and TDP-43 (p= 0.0120) gene targets (Supplementary figure S98) and a statistically significant negative correlation between HERV-K *env* and BCL11b in ALS samples (p= 0.0031). Also shown in Figure 4.13E is a lack of correlation between HERV-K transcripts in control tissue. When comparing HERV-K *RT* transcript expression to TDP-43 and BCL11b expression levels there was a significant positive correlation in ALS to TDP-43 and a significant positive correlation in Control samples to BCL11b. As increased TDP-43 has been observed to have a positive effect on HERV-K expression and BCL11b is thought to have a role in retrovirus latency these results would lend evidence to that effect in premotor cortex tissue samples in this sample set.

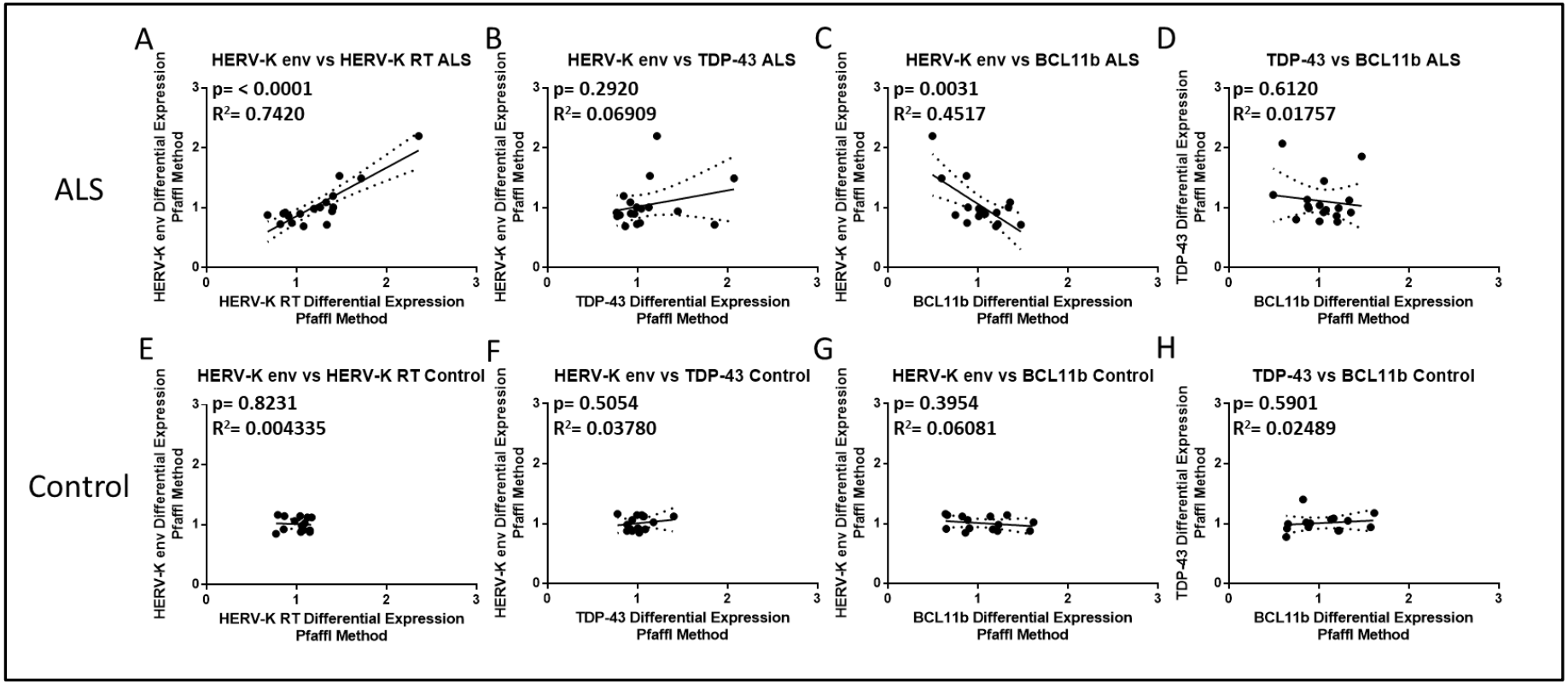


Figure 4.13. Differential Expression levels for Comparison of HERV-K *env* and *RT,* TDP-43, BCL11b transcripts as Calculated Using the Pfaffl Method

The graphs displayed in the figure above show 2^-ΔΔCt Differential Expression levels of HERV-K *env a*nd RT, TDP-43 and BCL11b transcripts in ALS and non-ALS control cases compared for correlation between their relative expression levels. The expression data is normalised against a geometric mean of GAPDH and XPNPEP1 reference genes with the R^2 and *p*-values calculated in GraphPad using its linear regression analysis. The outlier reading seen in ALS HERV-K *env* & *RT* (A, B & C) is sample A381/11, the same as the highest relative expression seen in the TDP-43 gene target.

4.3 Discussion

In the research paper published by Li *et.al.* (2015), describing increased levels of HERV-K expression in the frontal cortex of ALS brain tissue samples drew much attention as there is currently no diagnostic marker available or effective treatment for ALS to-date.

The aim of this research study was to see if we could independently confirm the research findings by Li *et al*. (2015) but in a UK cohort and on a larger sample size for statistical purposes as well as incorporating more than one reference gene for normalisation of gene expression purposes to comply with MIQE guidelines. The findings in this study failed to confirm that HERV-K expression is elevated in ALS post-mortem premotor cortex tissue specimens relative to controls matched as close as possible for sex, age, and Post-mortem Delay (PMD). These findings are in line with at least 3 other recent publications that did not find elevated levels of HERV-K in ALS post-mortem brain tissue samples.

Mayer *et.al* (2018) analysed RNA-Seq data from tissue samples obtained from brain and spinal cord and did not find elevated HERV-K transcript expression. In this study a cohort of 16 Cerebellum samples (n=9 ALS, n=7 Control), 30 spinal cord samples (n=15 ALS, n=15 Control), 35 from the motor cortex (n=23 ALS, n=12 Control) and 19 samples from the occipital cortex (n=14 ALS, n=5 Control) was tested for HERV-K (HML-2) expression. Despite the transcript levels varying between tissue types there was no significant difference in expression between ALS and control samples in any tissue type as measured by the same method used to analyse samples in this thesis, the 2^-ΔΔCt method.

In a recent paper I contributed to by Garson *et.al.* (2019), we looked in a different UK cohort of ALS patient samples and non-ALS controls (n=34 ALS, n=24 Control) from the premotor cortex; obtained from the same UK Neurodegenerative disease brain bank as the samples we analysed in this study. This study also used the 2^-ΔΔCt method and reported no significant difference in HERV-K transcript expression in the premotor cortex of ALS patients compared to control cases. It is important to note that despite the disparity in sample sizes between my own cohort and one used in our paper there was no overall difference in the significance, even when cancer samples were removed from either study. This region was the same used in the Li *et.al.* study and correspondence with the research team behind that paper confirmed that the premotor cortex was ideal for research into

HERV-K (HML-2) expression. Digital PCR was also used on the sample set to test for differences in absolute copy number in ALS vs Controls (unpublished data).

A more recent paper has also been published, looking at HERV-K expression in the premotor cortex of a Japanese cohort consisting of 29 motor cortex tissue samples (n=13 ALS, n=16 Control) and 15 spinal cord tissue samples (n=6 ALS, n=9 Control). Similar to our own published data in Garson *et.al.* (2019) this research group found good correlation between the HERV-K *gag, pol* & *env* transcripts indicating that a full provirus was expressed in their tissue samples. While the study found that HERV-K transcripts were expressed in both tissue types they were unable to find any significant change in expression in ALS vs controls for HERV-K (HML-2) (Ishihara et al., 2022).

There are multiple factors, such as PH level, preservation, storage and prolonged stress prior to death that can effect expression data from post-mortem brain tissue (Stan *et al.*, 2006; Weis *et al.*, 2007; Durrenberger *et al.*, 2010). Degradation of RNA in tissue samples has also been shown to have a negative effect on the performance of RT-qPCR, requiring an estimation of sample quality prior to use in an assay (Angela Pérez-Novo *et al.*, 2005; Derveaux, Vandesompele and Hellemans, 2010). For this data set only samples with RIN values in excess of 4.0 were used for the final analysis. A minimum value of between 3.95 and 5.0 has been suggested for RNA quality as a cut-off to ensure accuracy of data obtained from patient samples, with RINs of 8.0 being considered as “perfect” but not always feasible due to the delay in obtaining post-mortem tissue following death (Fleige and Pfaffl, 2006; Weis *et al.*, 2007; Rydbirk *et al.*, 2016). The source of the RNA can also have an effect on estimation of RNA quality, with relative integrity of mRNA in post-mortem brain tissue reported as much lower than other tissues (Koppelkamm *et al.*, 2011). It has also been shown that RNA integrity in brain tissue can be a poor measure of RNA quality, requiring appropriate reference genes to be selected to ensure accurate expression estimates (Sonntag *et al.*, 2016). This can be directly observed in the outlier ALS sample A381/11, shown as the having the highest expressed HERV-K transcript across Sections 4.2.2-4.2.5 and Sections 4.2.6-4.2.7 expression assay runs. This sample has a RIN of 6.6 and extracted RNA concentration of 328ng/µl, within the average scores of both sample sets; if there were a difference in expression due to RNA quality then this sample would have the highest RIN value. The only consistent significant relationship to increasing RIN was observed in

HERV-W *env* which had a significant negative correlation to RIN in the principal assay (Sections 4.2.2-4.2.5); meaning a decrease in expression with increasing quality of RNA as measured by RIN.

HERV-K expression has been found to be differentially expressed in relation to age and found to be more highly expressed in older individuals (Wallace *et al.*, 2014; Balestrieri *et al.*, 2015). This increase in HERV-K expression as age increases in individuals could also be due to the action of transcriptional suppression due to DNA methylation (Sverdlov, 1998; Yu, Zhao and Zhu, 2013; Buzdin, Prassolov and Garazha, 2017). As we age the methylation of DNA in the human genome is reduced, taking away this important suppressor of HERV transcription potentially causing the increase in HERV-K expression seen in earlier studies (Johnson *et al.*, 2012). We found evidence of this in ALS samples when normalised to GAPDH where we found HERV-K *pol* and *env* transcripts had a positive correlation to age (Supplementary Figure S43). However, this was not observed when normalised against XPNPEP1, though when analysed using a geometric mean of reference genes using the Pfaffl method the significant positive correlation of these transcripts was recorded (Supplementary Figure S55). This highlights the importance of ensuring that experimental samples and controls are matched for age (Nevalainen *et al.*, 2018). Great attention was paid in ensuring that all aspects of the assay complied with MIQE guidelines for conducting RT-qPCR experiments. This included using only reference genes that had been validated and stably expressed in our post-mortem premotor cortex brain tissue samples. This is coupled with evidence that RNA integrity may affect genes expression in differing ways, with stability changing dependent on the overall RNA quality of samples analysed (Angela Pérez-Novo *et al.*, 2005), however as these assays have used a properly validated set of reference genes for normalisation, this can potentially be ruled out. The 2^-ΔΔCt method for relative quantification of HERV-K is a useful tool for the analysis of RT-qPCR data though it has a limitation in not accounting for the differing efficiencies of primer sets used in the assays. The Pfaffl method for relative gene quantification accounts for these differences in the mathematical model for a more accurate interpretation of data (M. W. Pfaffl, 2001). When analysed by the Pfaffl method the results confirmed those observed in the 2^-ΔΔCt method with no significant differences in relative expression of HERV-K transcripts between ALS and non-ALS controls. There was also no significant expression observed

between ALS and controls observed with HERV-W transcripts when normalised against either GAPDH or XPNPEP1 (Figure 4.3).

Aside from the effects of Age in ALS when normalised to GAPDH there were other factors that affected the expression data of HERV-K transcripts when ALS and Control samples were analysed separately. When the data was normalised to GAPDH there were significant positive correlations with RIN in control samples, and a negative correlation with PMD, similar to the data when normalised with XPNPEP1. This negative correlation with PMD could be attributed to the degradation in the quality of RNA with increasing PMD (Angela Pérez-Novo *et al.*, 2005; Nagy *et al.*, 2015). This negative correlation with PMD in control tissue is not seen in all transcripts and HERV-K *gag* when normalised to XPNPEP1 in ALS samples shows a positive association with PMD, which relates to an increase in relative expression with increasing PMD. This positive association with HERV-K *gag* and PMD in ALS samples was also seen when using the Pfaffl method with control samples showing a significant negative correlation with HERV-K *gag, pol* and *RT* transcript expression (Supplementary Figure S55). Degrading RNA could potentially be a cause of these variations in measuring expression between ALS and non-ALS controls, however, RIN values are not a completely reliable measure of integrity in post-mortem brain tissue (Sonntag *et al.*, 2016). The paper by Sonntag *et.al.* (2016) found discrepancies between reported RINs of individual samples and their measured RNA concentrations. The expectation of these analyses is that with decreasing RIN you would find decreasing RNA concentration and this was not the case in the study. Instead, they found full RNA sequences in lower RIN samples indicating that some degradation of rRNA subunits may happen in the brain as opposed to other tissue making RIN an unreliable metric from brain tissue samples. This can be observed in the ΔΔCt data when normalised against XPNPEP1 where there was only a significant negative correlation in HERV-W, HERV-K *pol* and HERV-K *RT* transcripts in ALS tissue with control tissue showing a significant positive correlation in HERV-K *pol* expression when normalised against GAPDH (Supplementary Figures S47-48 & S52). The only significant result for RIN that correlated with the Pfaffl method for relative quantification was the negative correlation in ALS tissue with HERV-W gene expression (Supplementary Figure S61). This highlights the need for proper normalisation against multiple reference

genes for accurate interpretation of results as Li *et.al.* (2015) only used a single reference gene, GAPDH.

To analyse the effect of control samples where the donor patient has cancer, which is commonly linked to increased HERV-K expression, assays were set up to test HERV-K transcripts with those samples with linked conditions removed. As HERV-K has been shown to be upregulated and associated with certain cancers, Autoimmune disorders and other non-ALS neurodegenerative conditions these samples were removed from consideration and additional no-cancer controls added to ensure statistical validity in further analysis we undertook looking at relative expression levels of TDP-43 and BCL11b in sporadic ALS compared to non-ALS controls (Brodziak *et al.*, 2012; Gonzalez-Cao *et al.*, 2016; Chen, Foroozesh and Qin, 2019). TDP-43 and BCL11b have been reported in the literature to act as transcriptional regulators of retroviral infection and increased TDP-43 expression has been observed in both familial and sporadic ALS, and why we chose to investigate these gene targets in our analysis to see if they were differentially expressed and/or correlated with HERV-K expression in sporadic ALS (Desplats *et al.*, 2013; Ajroud-Driss and Siddique, 2015; Li *et al.*, 2015). From our research findings, when TDP-43 and BCL11b expression was normalised against either XPNPEP1 or GAPDH using the 2^-ΔΔCt method there was no significant difference in relative expression between ALS and no-cancer controls for both targets, similarly this was the case when the analysis was performed using the Pfaffl method.

BCL11b when normalised against XPNPEP1 in the 2^-ΔΔCt method and when normalised against both reference genes using the Pfaffl mathematical model showed a significant negative correlation with HERV-K *env* transcripts in ALS samples (Figure 4.10C & Figure 4.7G) but not in control samples (Figure 4.10G & 4.8C+G). This suggests a potential inhibitory effect of increase in BCL11b on HERV-K expression in ALS. Observations in latent HIV infections of the CNS showed binding of BCL11b to LTR regions of the HIV provirus, effectively silencing expression in the neuronal cells (Desplats *et al.*, 2013). As BCL11b is also involved in T-Cell development there is a potential involvement in the inflammatory response in the disease, though its role in HIV latency is better categorised (Lennon *et al.*, 2016).

While there was no significant difference between ALS and Control samples for TDP-43 expression the correlation with HERV transcript expression within ALS and Control sample sets yielded interesting results. Experiments measuring the relationship between the two in the literature have seen a positive correlation between their expression, possibly relating to the DNA binding nature of the TDP-43 protein (Li *et al.*, 2015; Douville and Nath, 2017). This apparent discontinuity between the established role of TDP-43 in sporadic ALS and the lack of differential expression seen in this study compared to other research groups is of particular interest. This is highlighted in Figure 4.10A&D which shows the relationship between HERV-K *RT* and TDP-43 in ALS samples when normalised against GAPDH (4.11A) and XPNPEP1 (4.11B). When normalised against either using the ΔΔCt normalisation method there was a positive correlation between them, meaning as one factor increases expression so does the other. This was also seen when using a combination of the reference genes in the Pfaffl method of measuring differential gene expression in ALS patient samples (Figure 4.13A). This information shows that it is possible an interaction may exist in pathogenic motor neurons that maybe independent of HERV-K expression despite other studies showing a link between TDP-43 and the endogenous retrovirus family. This could be very informative to other studies as TDP-43 has been shown to strongly co-localise in affected cells with HERV proteins and is also being investigated for its use as a novel biomarker in ALS (Majumder *et al.*, 2018; Dolei *et al.*, 2019).

An interesting observation from the brain tissue expression data that was analysed using both the ΔΔCt and Pfaffl methods of relative quantification is the lack of statistically significant correlation of HERV-K transcripts compared with no-cancer control tissue (Figures 4.8A&E and 4.10E). This could potentially be due to the removal of non-ALS control tissue that were derived from individuals with other clinical conditions not associated with ALS but certain cancers for example in which overexpression of a number of HERVs have been observed. Hence, overexpression of HERV's in this non-ALS group could be a reason why we did not see differential expression that was statistically significant in the ALS cohort. (Golan *et al.*, 2008; Johanning *et al.*, 2017; Saito *et al.*, 2017; Grandi and Tramontano, 2018). However, we obtained a similar result when we re-analysed the expression data and removed cancer controls from the non-ALS group. Another potential factor limiting our ability to detect changes in these cohorts was the lower sample size of

the analyses compared to the larger sample set as this subset may not truly represent the population of those affected by ALS.

Endogenous Retroviruses remain of interest in relation to ALS, particularly as there has been evidence of increased retroviral activity in ALS serum samples compared to non-ALS controls (McCormick *et al.*, 2008). Although we did not find statistically significant differences in the level of HERV-K or HERV-W expression, the next step will be to continue to look for differential expression of other HERV family members using RNA NGS analysis so that we can undertake a broad screening of the expression profiles of all HERV families in ALS and controls to look for any significant differences in HERV expression that we might have missed in our RT-qPCR assay which was focused only on two HERV families (HERV-K and HERV-W). We have yet to confirm difference in expression at the protein level however, with future work planned to confirm this by immunostaining of FFPE tissue for HERV-K gag and env proteins.

5.0 Confirmation of RNA-Seq Identified HERV-K3 Transcripts by RT-qPCR

5.1 Introduction

RT-qPCR remains a highly sensitive tool for the estimation of gene expression in multiple tissue types, however it is limited to detecting targeted sequences known prior to the analysis (Kimberly R. Kukurba and Montgomery, 2015). Analysis of gene expression by Next-generation sequencing (NGS) by comparison does not require this prior knowledge and provides a useful tool for discovering novel gene expression in tissue (Kimberly R. Kukurba and Montgomery, 2015). Aside from quantifying mRNA expression levels, NGS RNA sequencing (RNA-Seq) method has the ability to look into pre-mRNA and non-coding RNAs such as micro RNA (miRNA) and ncRNA (Kimberly R. Kukurba and Montgomery, 2015). An additional benefit is the ability to look into the expression of all of these RNA types in the same samples and datasets; for example a study looking at ALS in Peripheral Blood Mononuclear Cells (PBMC's) and Brain tissue was able to identify 13 dysregulated genes common to both tissues along with a number of miRNAs (Rahman *et al.*, 2019). RNA-seq methods also have the capacity to detect long non-coding RNA's (lncRNA) in PBMCs, lending evidence to the utility of RNA-seq in profiling whole transcriptomes of tissue types under investigation (Zucca *et al.*, 2019).

Another advantage of RNA-seq over RT-qPCR is the ability to provide a high throughput method of analysing expression data. While another method of high-throughput gene expression analysis exists in the form of microarray assay, RNA sequencing has been shown to have higher reproducibility in the results obtained along with higher concordance with RT-qPCR results in the same tissues (Li *et al.*, 2016). Using RNA-sequencing, Li et al. (2016) identified 23 differentially expressed genes identified in both RNA-Seq and Microarray methods (Li *et al.*, 2016). Another recent study citing poor reproducibility of microarrays was able to use publicly available RNA-seq data to confirm the differential expression of several genes involved in the study (Patel, Dobson and Newhouse, 2019). Another advantage of RNA-Seq over Microarray analysis is that Microarrays still require some prior knowledge of gene sequences, reducing the likelihood of novel gene expression discovery (Lowe *et al.*, 2017).

RNA-Seq utilises many NGS platforms for the characterisation of the transcriptome (Li *et al.*, 2014). In RNA-Seq the NGS library preparation differs slightly from DNA in that total RNA needs to be purified from extracted patient total RNA before fragmentation and cDNA synthesis (Atamian and Kaloshian, 2012). Beyond the identification of novel transcripts and RNA splice variants, RNA-Seq allows for the quantification of gene expression across the whole transcriptome and allows the expression of individual alleles to be identified (Kimberly R Kukurba and Montgomery, 2015). Analysing the RNA-Seq data generated by the NGS platforms often requires powerful computational and software resources to adequately analyse the data (Han *et al.*, 2015).

While HERV-K transcripts have been found to be differentially regulated in RNA-Seq data from studies such as Prudencio et.al. (2017) they are from different Human Endogenous MMTV-like (HML, Garcia-Montojo et al., 2018) groups to the HERV-K (HML-2) initially identified by Li et.al. (2015) and evaluated by RT-qPCR in Chapter 4.0. HERV-K3 (HML-6) has been shown to be upregulated in post-mortem primary motor cortex tissue in an RNA-Seq analysis performed by Jones et al., (2021). This research paper found a transcript of HERV-K3 (HML6) located at locus 3p21.31c (chr3:46426676–46433564) to be upregulated in the primary motor cortex. HERV-W, also annotated as HERV17 in dfam.org and University of California, Santa Cruz (UCSC) databases, has been a RT-qPCR target previously in this thesis (Chapter 4.0). Briefly, the normal function of HERV-W in healthy tissue is primarily centred around its *env* protein Syncytin-1 which is used to form the cell-cell tight junctions as part of the placental barrier during pregnancy (Grandi and Tramontano, 2017; Wang, Huang and Zhu, 2018). In human disease MRSV transcripts that are similar to HERV-W's Syncytin-1 have been shown to be upregulated in MS patients (Mameli *et al.*, 2009; Antony *et al.*, 2011; Dolei *et al.*, 2015; Grandi and Tramontano, 2017; Wang, Huang and Zhu, 2018).

We will be using the differential expression data from the paper by Jones *et.al.* (2021)to evaluate a selection of primary motor cortex tissue samples provided by the author to see if we can observe a difference in expression of this ERVK3 locus and HERV-W *env* by RT-qPCR. In addition to this sample set we will also be using RT-qPCR on a larger cohort of n=54 ALS and n=37 Non-ALS Control post-mortem premotor cortex tissue samples, with additional samples coming from those tested by (Garson *et al.*, 2019), to analyse whether the HERV-K3 transcript is significantly expressed in the premotor cortex. These RT-qPCR

experiments will use the previously identified reference genes XPNPEP1 and GAPDH as these have been validated on post-mortem brain tissue in Chapter 3.0.

5.2 Results

5.2.1 Genomic DNA Amplification of HERV-K3 *env* Transcripts by Polymerase Chain Reaction.

As the HERV-K3 (HML-6) family member that was detected by Jones et.al. (2021) was on a specific locus within the genome a primer set chosen based on previous methodology would not be suitable. This is because the previous primer sets were chosen to target a wide array of family members within the genome and a HERV-K *env* primer set based on that methodology would not be specific enough to be sure of detecting the HERV-K family member at the 3p21.31c locus. The method used to design the primers for HERV-K *env* is described in full in methods section 2.2.17. Briefly, an alignment of the HERV-K3 sequence present at 3p21.31c locus was performed against a HERV-K3 consensus sequence obtained from Dfam.org and a primer set designed so that the forward primer flanked the end of a 1200bp insertion unique to the HERV-K3 sequence present in the locus.

As with Chapter 4 quality control for the HERV-K3 *env* primers follows the methods outlined in Chapter 3. RNA was extracted from N=10 ALS and N=10 Non-ALS Control primary motor cortex tissue samples and quantified prior to initial assays (Summarised in Table 5.1 below, full information provided in Supplementary Table 15). While initial assessment of amplification efficiency was successfully measured on an ALS patient sample (Supplementary Table 16, Supplementary Figure S197-198), subsequent expression assays showed that the standard deviation of replicates for the primer set were poor and multiple amplicon sizes were detected in the melt curve and electrophoresis analysis (Supplemetary Figures S199-S201).

Table 5.1. Summary of the Quantification of Total RNA Extracted from n=10 ALS and n=10 Non-ALS Post-Mortem Primary Motor Cortex Brain Tissue Samples.
In the table below summary information is given on RIN values obtained from Agilent Bioanalyser 2100 along with RNA yield as measured using the Qubit BR (Broad Range) assay. Qubit means were derived from duplicate/triplicate values that were within 40ng/µl of each other, and only those values used for the mean quantification of RNA yield are given in the table. Full data is shown in Supplementary Table 15.

Variable	Summary Statistic	Value
RIN	Median	6.6
	Range	3.4-8.1
QuBit Derived Conc. (ng/µl)	Median	477
	Range	199.5-673

The variations in size of amplicons in the HERV-K3 locus on chromosome 3 seen in the validation steps can potentially be explained by allelic variation. In mendelian genetics the inherited genetic variation in the locus of a gene which is a causative factor in human disease can lead to a difference in the phenotype of a disease (Sidransky, 2006). This means that the variation in amplicon size at this specific locus on chromosome 3 could potentially have an effect on pathogenesis in ALS. To measure the potential variation in this HERV-K3 locus n=20 ALS and n=19 non-ALS controls were selected from available post-mortem Premotor Cortex brain tissue samples. Briefly, genomic DNA (gDNA) was extracted from these samples using DNeasy Blood and Tissue kit (Qiagen #69504) with proteins removed by Proteinase K digestion from the kit. RNA was removed by the use of RNase A (Qiagen #158922) and DNA purified from the extracted material using on-column DNA purification as provided by the Blood and Tissue Kit.

Figure 5.1 below shows a gel electrophoresis output for the 39 gDNA samples and identifies whether the sample comes from the ALS or non-ALS control groups. These patient samples were also matched for gender and age at time of death. As we can see from the gel electrophoresis image the majority of patient samples are homozygous for the larger (176bp) amplicon of HERV-K *env* with only 3 samples showing homozygosity for the smaller (approx. 156bp) amplicon. For examples of heterozygous inclusion of both amplicons there are 8 samples showing both amplicon sizes. These are a mix of ALS & control and male & female samples therefore, these factors do not seem to be contributing to the pathogenesis of ALS based on the subset of samples tested.

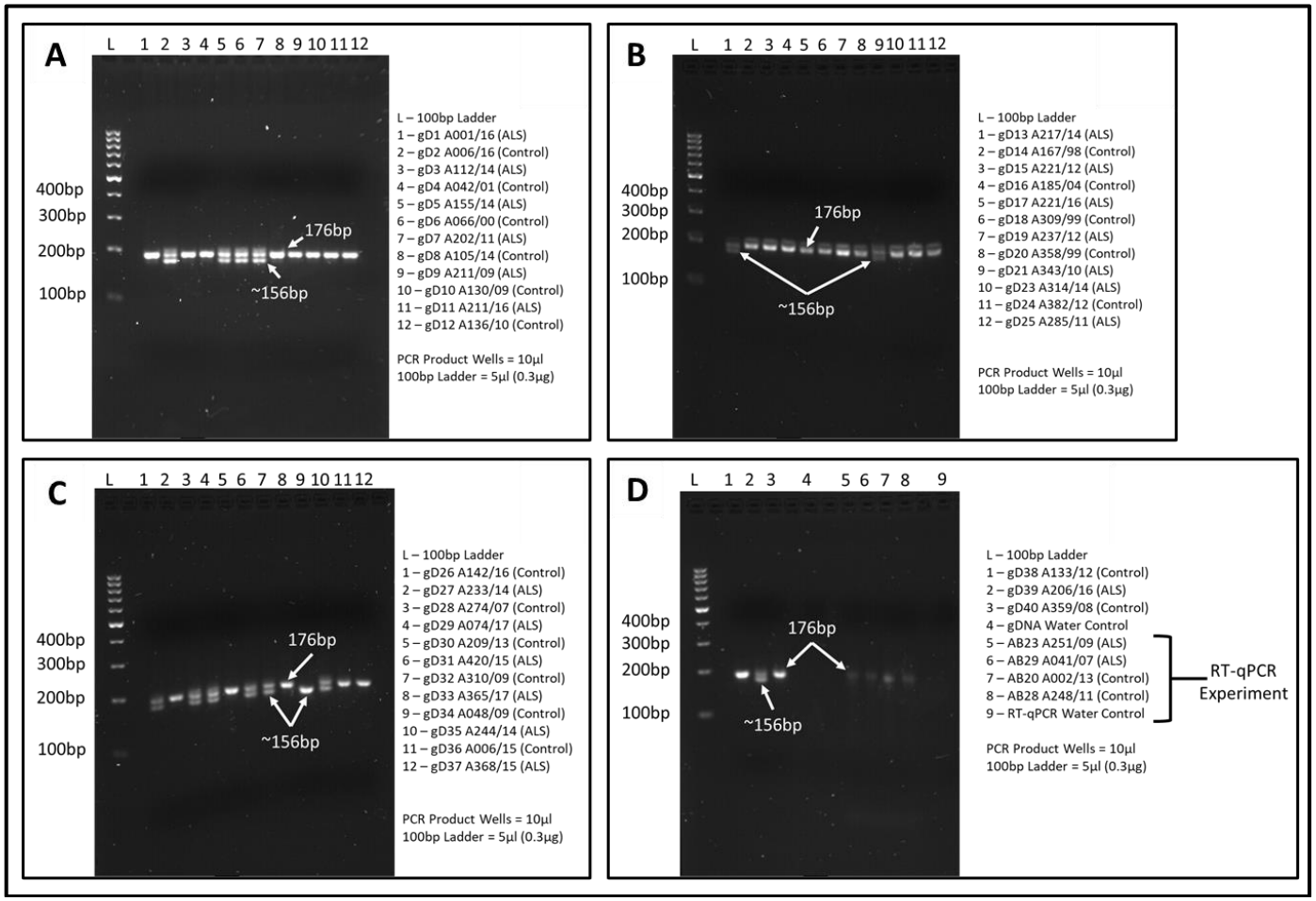


Figure 5.1 Agarose Gel Electrophoresis Analysis of HERV-K3 env Amplicons Produced by Conventional PCR Utilising Genomic DNA from n=20 ALS and n=19 Non-ALS Control Post-Mortem Premotor Cortex samples.

The figure above shows the gel electrophoresis results for HERV-K3 env Genomic DNA (gDNA) amplification by Polymerase Chain Reaction. The gel in the image above was made to a 2% concentration in TBE buffer with the 100bp ladder Generuler (ThermoFisher Scientific, SM0241).

5.2.2. Differential Expression of HERV-K3 *pol* Transcripts in Post-Mortem Primary Motor Cortex Tissue Samples from n=10 ALS and n=10 Non-ALS Controls

Due to the amplification profile shown by HERV-K3 *env* in the SYBR green assay (Supplementary Figures S199-201) it was determined that the targeted section of the HERV-K3 locus at 3p21.31c was not optimal for assessing the differential expression of the transcript via RT-qPCR. As the transcription of this specific locus is low both TaqMan and SYBR green chemistry were used during the validation of a primer set targeting the *pol* region of HERV-K3. For this new primer set the *pol* region was targeted for primer design using the paper by Pisano *et al.* (2019) wherein a 29bp deletion in the *pol* region of the HERV-K3 provirus in the specific locus of interest (referred to as 3p21.31b rather than c in the paper) was observed with the forward primer being targeted for that location. The reverse primer was positioned to give an optimal amplicon length and primer sequence characteristics. While TaqMan chemistry normally has the advantage of highly specific amplification of a primer target sequence the method was unable to successfully provide a favourable amplification effieciency for the primer set. SYBR Green chemistry however was able to pass all quality control methods and the primer amplicon was successfully sequenced (Supplementary Tables 17-18, Supplementary Figures S202-S203).

Following confirmation of the specificity of HERV-K3 *pol* primer sets, we tested blinded the set of n=10 ALS and n=10 non-ALS postmortem primary motor cortex tissue samples for HERV-K3 pol expression and the samples were unblinded after the RT-qPCR assay was performed. These samples were taken from the larger Kings College sample set in which RNA-Seq was undertaken previously as they gave a range from low to high levels of expression of HERV-K3 transcript located in chromosome 3 by RNA-Seq analysis (Jones et al. (2021) and we wanted to see if we could confirm this experimentally using our optimised RT-qPCR assay. Melt curves and amplification plots showing distribution of amplification curves and specificity of primer set for a single amplicon are provided in Supplementary Figure S204.

Samples were tested in triplicate using a concentration of 50ng/µl of Poly-A Carrier RNA in the diluent to ensure the low copy number of the transcripts were accurately measured for each sample. Estimation of differential expression values for HERV-K3 pol were calculated in Microsoft Excel (Microsoft, Washington, USA) using the 2^-(ΔΔCt) method for relative quantification of RT-qPCR data (Schmittgen and Livak, 2008). For the calculation of the initial ΔCt value the data was normalised against GAPDH and XPNPEP1 reference genes separately and against a geometric mean of the 2 reference genes. A mean of ΔCt values for control samples was used as a calibrator for the ΔΔCt step. The summary for the difference in expression in ALS as measured by 2^-ΔΔCt compared to geometric mean of the control samples to ALS is shown in Table 5.2. Also shown in this table is the difference in expression when calculated by the Pfaffl method (M. W. Pfaffl, 2001), this method works differently to 2^-ΔΔCt as it also takes into account the amplification efficiency of the primers so the geomean of the control is not exactly 1. As we can see in this table there is a clear increase in expression of the HERV-K3 pol transcript in ALS compared to controls and all results are shown as statistically significant utilising the Mann Whitney u-test in GraphPad Prism. To confirm that this significant result is not the effect of other variables within the patient metadata binomal logistic regression was performed using IBM SPSS Version 24.0 (Table 5.3). AS we can see in the equation variables the only covariate which has a significant effect on the difference between ALS and non-ALS Control samples is the differential expression score (highlighted in yellow). The R^2 results in the model summary show that this model fits the majority of the data, with over 50% of the data covered when using the Cox method and over 70% shown as covered using the Nagelkerke method. The geometric mean method of determining 2^-ΔΔCt differential expression has been shown in Table 5.3 as an example of positive results with the other tables listed in Summary Tables 21–23.

Table 5.2. Geometric Mean of HERV-K3 *pol* Relative Expression in ALS and non-ALS control cases, Normalised to GAPDH or XPNPEP1 Reference Genes.

The table displays the geometric means of the 2^-ΔΔCt differential expression values from n=10 ALS and n=10 non-ALS control samples for HERV-K3 *pol* gene targets used in the RT-qPCR expression assay.

	GAPDH	XPNPEP1	GeoMean	Pfaffl
	HERV-K3 *pol*	HERV-K3 *pol*	HERV-K3 *pol*	HERV-K3 *pol*
ALS	5.521	4.745	5.151	9.126
Control	1	1	1	1.172
P-Value	0.0007	0.0052	0.0039	0.0029

Table 5.3 Binary Logistic Regression Analysis of HERV-K3 *pol* 2^-ΔΔCt Using a Geometric Mean of XPNPEP1 and GAPDH Reference Genes

The combined table below shows the R^2 model summaries for the binary regression followed by the p-value significance (Sig.) that each variable is related to the difference between ALS and Non-ALS control samples.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	12.317[a]	.537	.716

Variables in the Equation

		B	S.E.	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	RIN	1.601	1.543	1	.299	4.960	.241	102.101
	Sex(1)	1.754	1.567	1	.263	5.775	.268	124.624
	ddCT	2.630	1.512	1	.082	13.875	.717	268.667
	Constant	-15.823	11.677	1	.175	.000		

Figure 5.2 below shows the Mann-Whitney U test results and dot plot graphs for the 2^-ΔΔCt and Pfaffl methods of determining differential gene expression. As we can see from the graphs the HERV-K3 *pol* amplicon has been found to be significantly expressed (p-value cut-off 0.05) when normalised against a single reference gene or a geometric mean of the two. This data was initially generated using the ΔΔCt differential expression method in Microsoft Excel and P-Values determined using GraphPad PRISM 6.0 utilising the Mann-Whitney U-test for non-parametric data. Additionally, the data was also analysed to see whether there was any difference in expression according to gender of the patient and there were no significant results in ALS or non-ALS control samples (data not shown).

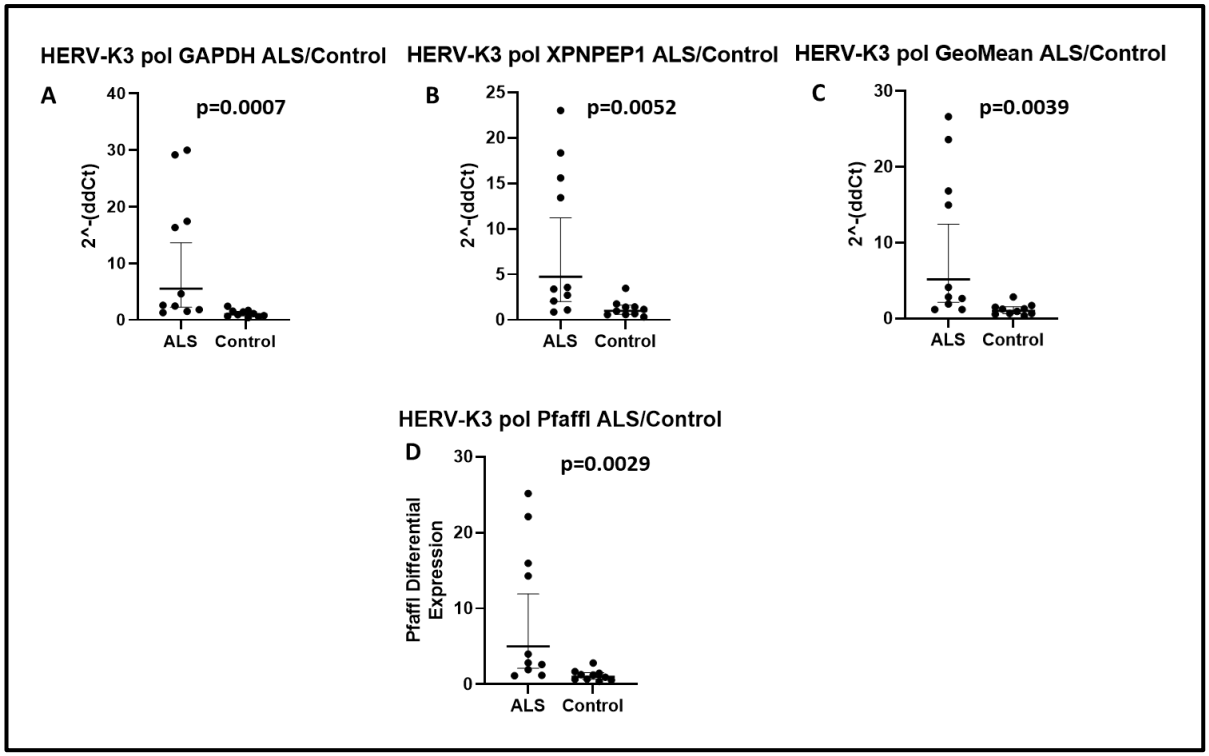


Figure 5.2. Dot Plots Showing Comparison of n=10 ALS and n=10 Non-ALS Control Samples Obtained from Postmortem Primary Motor Cortex Tissue Samples.

The Figure above shows the plotted 2^-(ΔΔCt) (A-C) and Pfaffl method (D) values for HERV-K3 *pol* expression in n=10 ALS and n=10 non-ALS control postmortem primary motor cortex samples . The thick line in the middle of each group represents the geometric mean with the lines above and below representing the 95% confidence interval of the geometric mean. Also included in each graph is the Mann-Whitney U test p-value for the analysis.

Figure 5.3 below shows a scatter plot of GAPDH normalised 2^-ΔΔCt values for each postmortem primary motor cortex patient sample plotted against their GAPDH normalised counts localised to the 3p21.31c locus for HERV-K3 *pol*. In order to normalise the RNA-Seq HERV-K3 counts against the GAPDH counts the relevant values were obtained from the HERV and cellular genes counts matrices then HERV-K3 counts were divided by the GAPDH counts. As we can see from the graph below the increasing counts localised to the region in patient samples matches the increased 2^-ΔΔCt seen in the RT-qPCR analysis. One outlier was removed from the graph, sample A081_91 which showed a high 2^-ΔΔCt despite having a low normalised counts value, however the positive relationship between the 2^-ΔΔCt result and the normalised counts was still present, though at a lower value of R^2=0.3674 with a P Value of 0.0046.

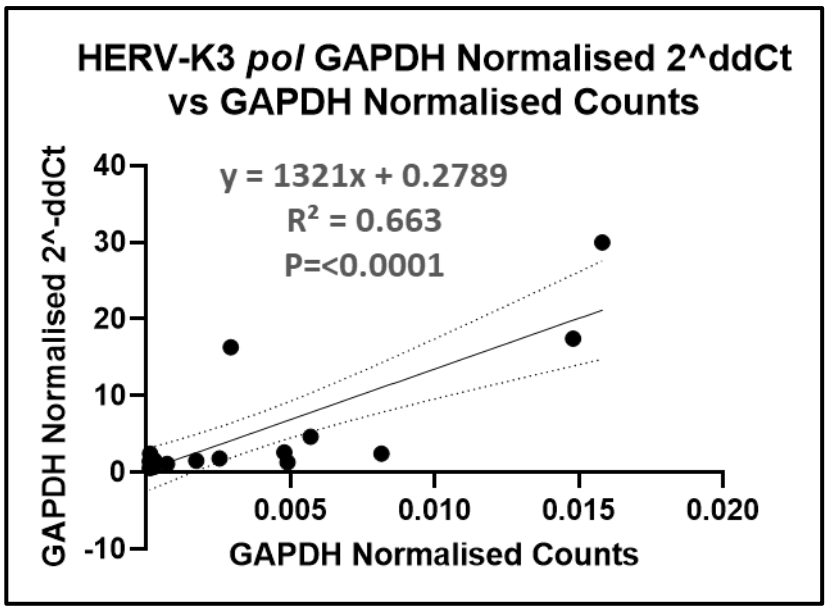


Figure 5.3. Linear Regression Analysis Scatter Plot of GAPDH Normalised 2^-ΔΔCt Values and GAPDH Normalised RNA-Seq Counts.

The scatter plot above shows the linear regression analysis of HERV-K3 *pol* differential expression results as calculated by the 2^-ΔΔCt method, when normalised against the GAPDH reference gene, against their GAPDH normalised RNA-Seq Counts.

5.2.3 Open Reading Frame Protein Analysis for HERV-K3 (HML6) Located at 3p21.31c

In order to analyse the HERV-K3 (HML6) located at the 3p21.31c locus for functional proteins the FASTA formatted nucleotide sequence was downloaded from UCSC and analysed in UGene for intact open reading frames. There were 2 open reading frames within the sequence which were able to be confirmed by BLASTn, a *pol* reading frame from position 3163-3194 and an *env* sequence from position 4422-5030. These sequences were translated in UGene to their amino acid sequence and predicted models built using the ExPASy SWISS model online tool. The HERV-K3 *env* sequence was unable to provide any useful 3D model, with only a fragment being predicted covering a small area of the open reading frame (Data not shown).

The open reading frame in the *pol* region was identified by ExPASy as an Integrase sequence, with the amino acid sequence forming a protein dimer (Figure 5.4). In order to see if our HERV-K3 *pol* amplicon lay within this predicted open reading frame a search was performed in UGene on the 3p21.31c sequence for amplicon sequences obtained from Sanger sequencing (Supplementary Table 18). While not an exact sequence match, most of the sequence for the A292/09 and A151/10 amplicon were observed in the open reading frame with the latter part of the translated sequence found within the predicted model (Red circled region in Figure 5.4A). Within the HERV-K3 3p21.31c sequence the Sanger sequenced PCR amplicons align at codon positions 3169-3210 (A292/09) and 3178-3213 (A151/10). This 3D protein model was further analysed using an active site prediction tool provided by FTSite (Boston University, USA) and can be seen in Figure 5.4B. As we can see in the active site prediction image a wire frame area between the protein dimer has been predicted as a likely active site for the enzyme.

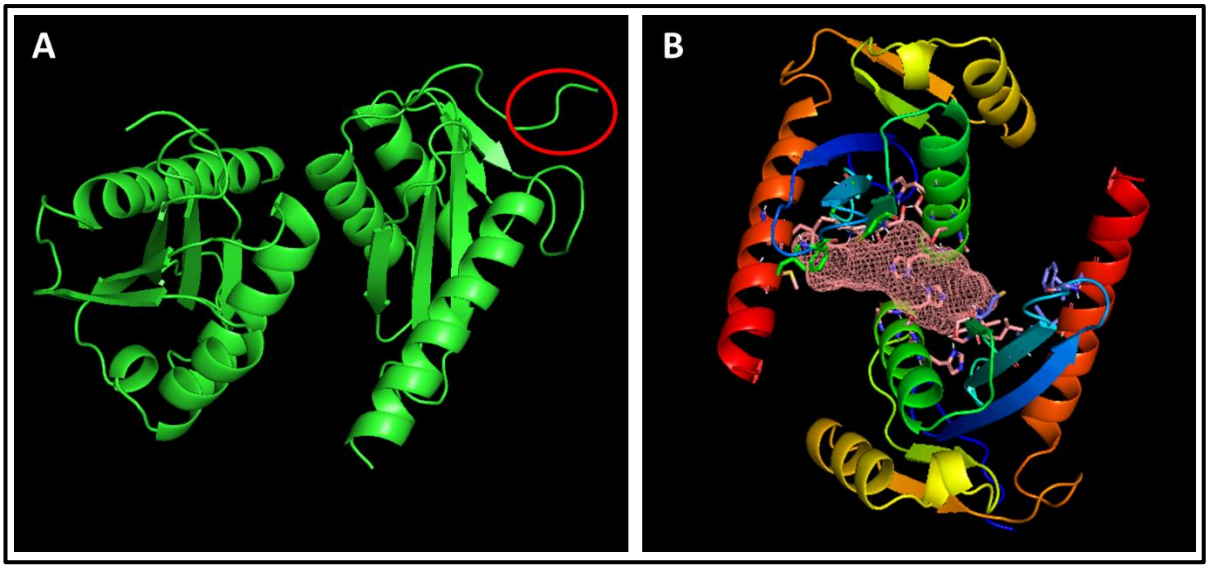


Figure 5.4. PyMol Generated Visualisations of HERV-K3 *pol* Integrase Open Reading Frame

The figure above shows 3D models of the predicted protein sequence produced by the Integrase open reading frame present at bp 3163-3194 of the UCSC sequence for HERV-K3 3p21.31c. The predicted protein sequence produced by ExPASy SWISS model is shown in A with the red circle highlighting the translated amplicon sequence location for the HERV-K3 *pol* primers. B shows the FTsite predicted active site for the predicted 3D model produced by ExPASy SWISS model.

In order to confirm the predicted protein structure is Integrase the amino acid sequence was entered into SMART (Simple Modular Architecture Research Tool) (Schultz *et al.*, 1998) and HMMER (Potter *et al.*, 2018) web tools for identifying proteins and functional regions within the protein sequences. Both tools confirmed the amino acid sequence as Integrase, with SMART giving an E-value (probability of sequence matching by chance) of $7x10^{-30}$ and HMMER giving the highest Homo Sapiens match an E-value of $4.0x10^{-48}$. SMART was also able to identify motifs within the integrase sequence, ZnF_C2HC, Zinc Finger DNA binding domain, MIT, a motif that has microtubule trafficking function and SWAP (Suppressor-of-White-APricot splicing regulator) domain which has RNA interaction and processing functions.

5.2.4 2^-ΔΔCt and Pfaffl Analysis of HERV-K3 *pol* Expression in n=47 ALS and n=29 Non-ALS Derived Postmortem Premotor Cortex Brain Tissue.

As we have confirmed the RNA-Seq findings from Jones *et.al.* (2021) by RT-qPCR it was determined that we should see whether the HERV-K3 3p21.31c locus was differentially expressed in a larger cohort of premotor cortex tissue samples (Materials Tables 2.5-2.7). This larger cohort includes all samples tested in Chapter 4, alongside additional samples from our paper Garson *et.al.* (2019) to make a total of 91 postmortem premotor cortex tissue samples (n=54 ALS, n=47 Control). As HERV-K3 *pol* amplicons had previously been shown to have poor duplicate values the cDNA samples were diluted in 50µg/ml of Poly-A carrier RNA prior to adding to the qPCR reaction plate and performed in triplicate. Quality control performed on GAPDH, XPNPEP1 and HERV-K3 *pol* assays provided in Supplementary Figures S205-S207 show that all amplicons produce a single defined peak indicating a lone amplicon.

A reduced sample set was eventually used for the evaluation of HERV-K3 *pol* transcript differential expression in postmortem premotor cortex tissue samples. A total of 16 samples (8 ALS and 8 non-ALS controls) were excluded from analysis due to having SD values outside of the cut-off of 0.3SD. The relative expression of the HERV-K3 *pol* transcript in the premotor cortex brain tissue samples was lower, though not statistically significant, in ALS when compared to the non-ALS controls when analysed by the 2^-ΔΔCt method (Table 5.4). This was consistent whether the HERV-K3 *pol* samples were normalised by GAPDH, XPNPEP1 or a geometric mean of the two reference genes. The only analysis which showed a positive difference was the Pfaffl method, though this difference between the two recorded values is due to the presence of the outlier samples A151/10 (ALS) and A265/08 (Control). Even when these outliers are removed from the Pfaffl analysis the positive difference remains though with a smaller value (p=2.661). These two samples are the same outliers in the other analyses, the outlier values are due to the two samples having lower (3-5Ct) Ct means than the rest of the ALS and Control groups.

Table 5.4. Geometric Mean of HERV-K3 *pol* Relative Expression in ALS and non-ALS control cases, Normalised to GAPDH or XPNPEP1 Reference Genes.
The table displays the geometric means of the 2^-ΔΔCt differential expression values from n=47 ALS and n=29 non-ALS control samples for the HERV-K3 *pol* gene targets used in the RT-qPCR expression assay. For the Pfaffl results, a standard mean for ALS and Control Values.

	GAPDH	XPNPEP1	GeoMean	Pfaffl
	HERV-K3 *pol*	HERV-K3 *pol*	HERV-K3 *pol*	HERV-K3 pol
ALS	0.740	0.677	0.708	5.602
Control	1	1	1	1.049
P-Value	0.4234	0.3055	0.3614	0.3531

Figure 5.5 below shows the differential expression data plotted as points on a dot plot. The data was analysed on GraphPad with the geometric mean and 95% confidence interval plotted though this is not visible due to the presence of outlier values. Whether normalised against GAPDH, XPNPEP1 or a Geometric mean of the two by the 2^-ΔΔCt method or by the Pfaffl method of differential expression there was no statistically significant difference in HERV-K3 pol expression between ALS and non-ALS control groups and did not alter when the outlier values were removed from the analysis.

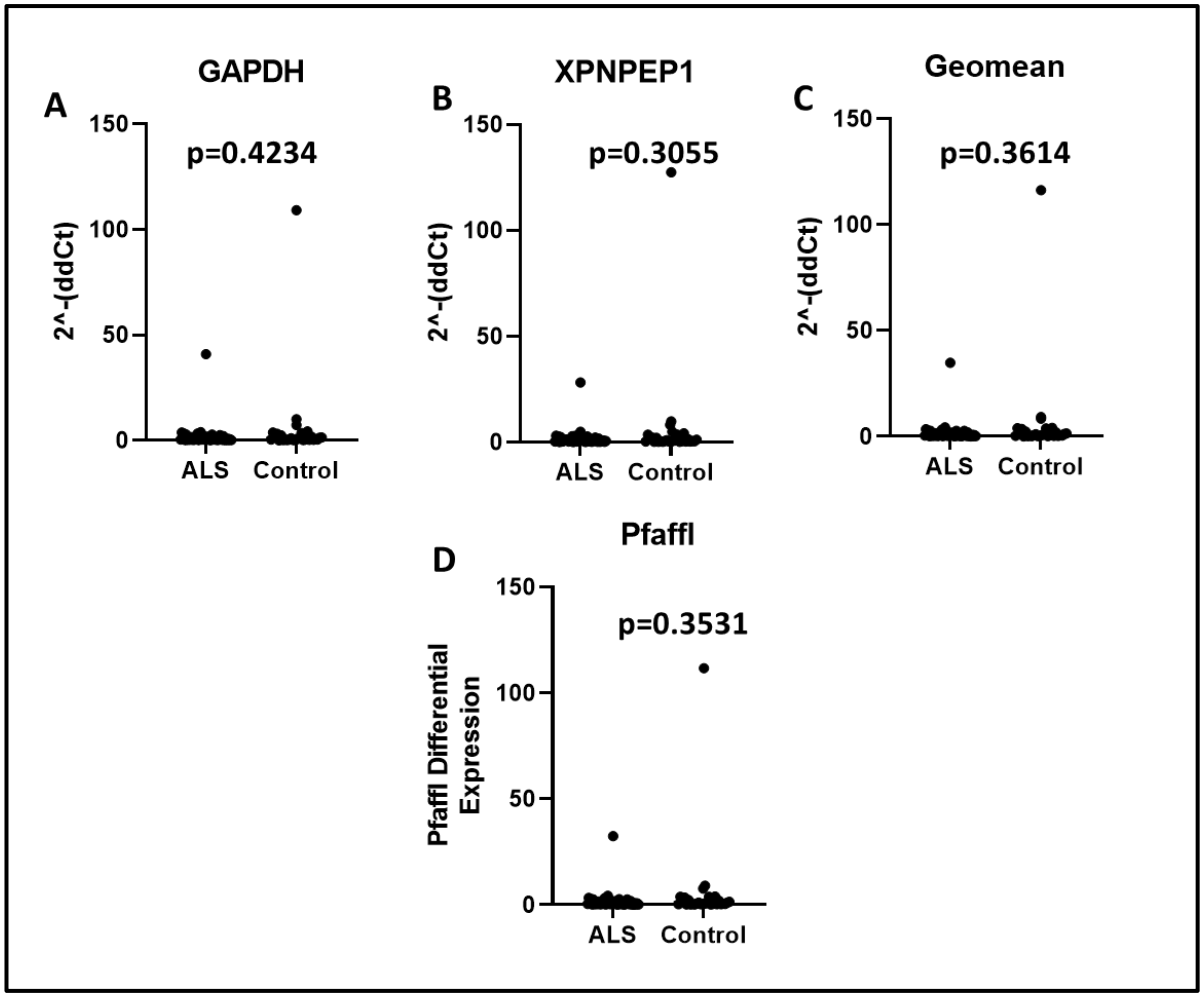


Figure 5.5. Dot Plots Showing Comparison of n=47 ALS and n=29 Non-ALS Control Samples Obtained from Postmortem Premotor Cortex Tissue Samples.

The Figure above shows the plotted 2^-(ΔΔCt) (A-C) and Pfaffl method (D) values for HERV-K3 *pol* expression in n=47 ALS and n=29 non-ALS control postmortem premotor cortex samples. A line in the middle of each group represents the geometric mean with the lines above and below representing the 95% confidence interval of the geometric mean (not visible due to outliers). Also included in each graph is the Mann-Whitney U test p-value for the analysis.

5.3 Discussion

RNA sequencing (RNA-seq) has allowed researchers to identify and quantify expression of elements within the transcriptome without the need to know the sequence beforehand (Rutter *et al.*, 2019; Simoneau *et al.*, 2021). The development of this method represented a significant advance over microarrays, the previous method of high throughput analysis of gene expression which still requires the user to have prior knowledge of the target sequence (Simoneau *et al.*, 2021). The popularisation of RNA-seq combined with the cost of assays lowering has allowed customisation of the process to target specific lengths and types of RNA sequences such as HERVs. The ERVMap pipeline used by Jones *et.al.* (2021) is a good example of this customisation, with custom perl codes and a database of known HERV sequences across the human genome it allows for the specific targeting and analysis of these low expression gene sets (Tokuyama *et al.*, 2018a). While this is a very useful tool for analysing differential expression of genes many different sources of interference in the preparation of samples for sequencing and the sequencing process can potentially interfere with the accurate measurement of gene expression (Simoneau *et al.*, 2021). Because of this it remains good practice to confirm the *in-silico* results with more standard laboratory practices such as RT-qPCR. Additionally, the data supplied to us by our collaborators at King's College was obtained from the primary motor cortex region of the brain which were not the target of our initial study. We had chosen the premotor cortex to focus our attention due to the pathology of ALS disease as well as from communication from Avindra Nath who published the Li et al. (2015) paper and found higher expression of HERV-K (HML-2) transcripts in ALS and controls in this anatomical region of the brain. Hence, this meant that the newly identified HERV-K3 provirus at locus 3p21.31c alongside one HERV which had already been studied in Chapter 4 (HERV-W) would need to be confirmed by RT-qPCR on post-mortem primary motor cortex tissue samples that we were provided by our collaborators at KCL in which RNA-Seq analysis had found HERV-K3 (HML6) to be upregulated in their published study (Jones et al., 2021)To this end they provided us with n=10 ALS and n=10 non-ALS control post-mortem primary motor cortex samples from their dataset provided by the MRC neurodegenerative disease brain bank. The HERV-K3 *env* primer sequences provided by our collaborator Dr. Jeremy Garson to target the specific locus have had mixed results when testing with extracted RNA. Amplification efficiency data was only able to be extracted from a single ALS sample at a lower 1:2 dilution series

compared to our standard 1:4 with control tissue unable to produce a consistent single amplicon across multiple dilutions (Supplementary Figure S197B). This pattern was also seen in the amplification of cDNA for differential expression quantification by RT-qPCR (Supplementary Figures S200C & S201). In the RT-qPCR assay the HERV-K3 *env* primer set failed to consistently produce a single amplicon band across both ALS and Control samples with a minimum Ct value of 35. The next stage of testing for this HERV-K3 locus would be the primer/probe method of TaqMan which should only bind specifically to the target sequence and not fluoresce when attached to dsDNA such as primer-dimers, as in the case of SYBR green chemistry (Heid *et al.*, 1996; Stephenson and Stephenson, 2016).

As the amplicon for the HERV-K3 *env* at the 3p21.31c locus appears to have two distinct bands, the main predicted size of 176bp based on in-silico analysis of the HERV-K3 *env* primer set and the smaller approximately 156bp band, the amplification of both of these alleles could potentially interfere with the accurate measure of their expression by RT-qPCR. This difference in band size is likely due to a previously described polymorphism in the HERV-K3 gene in this locus. The polymorphism, annotated as SNP: rs71098403 is a 14bp indel of which the insertion is the more common in the population though has no known clinical effect reported to ClinVar (NCBI, dbSNP). As we see from our amplicons generated from genomic DNA in Figure 5.1, the sequence at the chromosome 3 locus appears to have a mendelian inheritance pattern. We can see the existence of both the longer and shorter sequence in eight of the post-mortem premotor cortex tissue samples while the larger or smaller amplicon size appear by themselves in other samples. This variation in gene sequence in heterozygous and homozygous patients who possess the shorter HERV-K3 allele could potentially have an effect on disease phenotype (Douville and Nath, 2014). In a recent study of genes in Frontotemporal Degeneration patients found a heterozygous deletion of 13bp in a gene which was not present in 200 control patient samples (Adrião *et al.*, 2021). This provides a solid example of how a heterozygous change in sequence could affect developmental diseases. This does not appear to be the case with this heterozygous change between the post-mortem premotor cortex samples we investigated as the heterozygous change in amplicon length appears in both ALS and non-ALS controls and has a mix of Male and Female patients across differing age bands (Figure 5.1).

The alternative HERV-K3 *pol* gene target provided a much more reliable primer set than the HERV-K *env* primer set, resulting in a single PCR amplicon across all dilutions in the amplification efficiency experiment (Supplementary Figure S203 B&D). This primer set also performed much better in the subsequent differential expression experiment (Supplementary Figure S204) with a single amplicon being produced by all samples. It was also possible to accurately measure an increase in expression of this transcript in ALS compared to controls (Table 5.2) which was significant whether measured by an individual reference gene, a geometric mean of the two or by the Pfaffl method. This confirms the findings by Jones *et.al.* (2021) that the HERV-K3 provirus present at the 3p21.31c locus is upregulated in the primary motor cortex in ALS. However, it should be noted that the samples the research team behind the paper sent us were picked based on the samples which had the highest (from ALS samples) and lowest (from non-ALS control samples) number of reads mapped to the locus by RNA-Seq. This preferential selection of samples could be the reason that we are seeing a confirmation of the results presented in the paper so repeating the assay with a higher sample count would be beneficial to ensure accuracy in reporting a significant result. As this RT-qPCR assay showed that the HERV-K3 *pol* primers were specific to the locus by SYBR green the consistent failure of the primers when using the TaqMan chemistry must be for a reason other than specificity to the primer target.

Analysis of the HERV-K3 3p21.31 sequence for protein open reading frames was performed to see if the provirus coded for functional proteins. Of the several predicted open reading frames only 2 were able to be identified as human, a sequence within the *pol* region and one within the *env* region. Subsequent analysis was able to determine (Figure 5.4A) that the HERV-K3 *pol* primers used in the differential expression of postmortem primary motor cortex tissue samples targeted the start of the Integrase open reading frame of the HERV-K3 *pol* polyprotein region. This was initially shown via ExPASy SWISS model, with an active site between the protein dimers predicted by FTSite (Figure 5.4B. This identity was further confirmed by SMART (Schultz *et al.*, 1998) and HMMER (Potter *et al.*, 2018) web tools. SMART further identified protein motifs within this region which have expected functions within the Integrase enzyme, primarily the DNA binding zinc finger protein as the enzyme function is to insert proviral DNA into the host genome. As the HERV-K3 *pol* primer sequences targeted a potentially functional integrase protein it may provide some evidence

that integrase is of clinical significance in ALS. This result ties into a current clinical trial using Triumeq which contains RT and integrase inhibitors (Abacavir, Lamivudine, and Dolutegravir) , which is being conducted by the research group headed by Julian Gold to target HERV-K expression in ALS (Gold *et al.*, 2019; Garcia-Montojo *et al.*, 2021). If this trial is successful further research should be conducted as to the pathological mechanism the HERV-K Integrase sequence contributes to.

When the differential expression of the HERV-K *pol* transcript was measured on a larger premotor cortex sample set there were no significant difference in expression recorded between ALS and non-ALS controls (Table 5.4). While this sample set comes from a region to the front of the primary motor cortex there has been previous experiments studied on gene expression in ALS patient brain tissue which shows similarity in expression across brain regions. In a study by Phan *et al.* (2021) they found HERV-K transcripts significantly expressed in the Frontal Cortex compared to the cerebellum which shows that the difference is not exclusive to the motor cortex. A study by Lederer *et al.* (2007) also showed increases in similar genes across both the motor cortex and spinal cord which gives further evidence to differential expression of gene sets being distributed across multiple areas of the CNS in pathology. While these studies did find similarity in gene expression across different regions of the brain the study whose findings we are verifying in the primary motor cortex, Jones *et al.* (2021), found that the HERV-K3 (HML6) was differentially expressed in the primary motor cortex and frontal cortex but not the cerebellum. Despite these similarities in expression of HERVs and other genes across brain regions in ALS it is curious to see a lack of significant expression between ALS and non-ALS controls in the premotor cortex. It could be that the HERV-K3 provirus at locus 3p21.31c is too low in copy number to be effectively measured for differential expression in premotor cortex tissue samples by RT-qPCR. This is due to the inherent limitations of RT-qPCR assays as when a gene target is below a certain copy number the PCR reaction can generate random, unreproducible results including non-specific melt curves and failed reactions (Bernardo, Ribeiro Pinto and Albano, 2013). The low log2fold change seen in Jones *et.al.* (2021) translates to less than a cycle difference which could lead to errors in reporting even if primers are validated correctly. As the HERV-K3 *pol* primers were designed to target the specific locus expressed in the primary motor cortex there is a possibility that there is a

HERV-K3 family member differentially expressed at a different locus that the primer set is missing due to its specificity for the 3p21.31c locus. A possible solution to the issue of low copy number transcripts interfering with the viability of RT-qPCR is using a digital PCR technique which allows the quantification of exact copy number of target transcripts including low copy number genes such as HERV-K (Hindson *et al.*, 2011).

6.0 Differential Expression Analyses of Human Endogenous Retroviruses in ALS using RNA-seq Analysis.

6.1 Introduction

Endogenous retroviruses (ERVs) provide a unique challenge for RNA-seq data analysis because they are highly repetitive elements distributed across the genome (Tokuyama *et al.*, 2018b). For sequence reads, inferred sequences of base pairs matching all or a section of a single DNA fragment (Whiteford *et al.*, 2005), to be useful they must be long enough for the sequence to be mapped specifically to the nucleotide template (Whiteford *et al.*, 2005). This means as the length of a read decreases, the likelihood that the generated sequence information will map to multiple places in the reference genome increases (Whiteford *et al.*, 2005). This can prove to be problematic when dealing with repeat elements such as HERVs which exist in fragments throughout the human genome (Whiteford *et al.*, 2005; Bhardwaj *et al.*, 2015; Mayer *et al.*, 2018). Most analysis pipelines will also filter out ERV reads due to a single family member mapping to multiple locations across the genome; HERV-K113, for example, maps to 10 separate locations across chromosomes 4 and 19 (Wildschutte *et al.*, 2016; Tokuyama *et al.*, 2018b). ERVMap presents a solution to this problem by introducing a pipeline which preferentially filters ERV sequences and provides a database of ERV locations within the genome to analyse expression within RNA-Seq data (Tokuyama *et al.*, 2018b). This database consists of 3220 identified ERVs with the pipeline able to accurately assign reads to a specific locus for the individual repeat sequence (Tokuyama *et al.*, 2018b). This data is compiled from in silico analysis from nine separate papers, for those loci that overlapped between studies, the curators of the ERVmap database selected the loci with the most coverage of the reference sequence (Tokuyama *et al.*, 2018b). The counts generated by the gene counts analysis by this database are then converted to size factors and used to normalise gene expression in the ERV counts data to provide normalised expression data for analysis by other programs such as DESeq2 (Love, Huber and Anders, 2014).

In general, NGS data (whether DNA or RNA-Seq) is aligned against the relevant reference genome so accurate transcription information can be assigned to the correct genomic locus. There are non-reference alignment protocols available which create de novo sequences for novel transcripts and to map new viral/bacterial sequences. The latest

human genome assembly at this time, GRCh38, was initially released in 2013 with the most recent patch, introduced on March 1st 2019 (Genome Reference Consortium, 2020). Patches to the human genome assembly represent scaffolds which have been applied to fix variations without disrupting the chromosome base pair co-ordinates in the human reference genome (Genome Reference Consortium, 2020). This genomic sequence is primarily built from clones that were sequenced during the original Human genome project and features major improvements from the previous assembly (GRCh37) (Genome Reference Consortium, 2020). The GRCh38 build was also the first to benefit from the Illumina high throughput next generation sequencing technology, which replaced the microarray based method primarily used to update the previous genome assembly (Guo *et al.*, 2017). The patches to this build have, to date, covered an additional 113 gaps from the initial assembly with another 12 being closed thanks to extra-long reads generated by the Oxford Nanopore NGS system, all of which have been included in the most recent patch (Jain *et al.*, 2018; Genome Reference Consortium, 2020). These factors contribute to the GRCh38.p13 build being the most accurate representation of the human genome to date which includes the addition of annotation of centromere regions compared to the previous build (Guo *et al.*, 2017).

The work described below utilises the ERVMap pipeline, with updated tools used in place of older versions in the original ERVMap protocol including STAR align which replaces TopHat to speed up gene alignment. This method was used to analyse the expression of ERVs in ALS (Table 2.8 in Chapter 2) and non-ALS controls, using postmortem primary motor cortex brain and a publicly available dataset of postmortem Frontal Cortex and Cerebellum samples (Chapter 2 Table 2.10) (Prudencio *et al.*, 2017). By utilising the RNA-Seq analysis pipeline we will be able to look at the expression of individual ERVs and ERV families across the entire transcriptome, as opposed to HERV-K transcripts previously analysed by RT-qPCR (Chapter 4). In addition, co-expression analysis will be used to look at the relationships between ERV transcription and other known modifiers of retroviral transcription such as genes TAR DNA-binding protein 43 (*TARDBP*) and B-Cell CLL/Lymphoma 11B (*BCL11b*), which we investigated earlier by RT-qPCR (Chapter 4.0).

6.2. Results

6.2.1 Differential Expression of Endogenous Retroviruses (ERVs) Between ALS and Non-ALS Controls in Postmortem Primary Motor Cortex Tissue samples

Postmortem primary motor cortex tissue samples from n=11 ALS and n=14 non-ALS controls were supplied by the MRC London Neurodegenerative Diseases Brain Bank and was sequenced by Source Bioscience (Nottingham, UK) in collaboration with King's College London using an Illumina HiSeq 4000. RIN values were also measured by Source Bioscience and ranged between 4.8 and 7.8, full patient metadata can be seen in Table 2.8 These RINs were controlled for in the DESeq2 analysis considering some of the lower RIN ranges which were still above the cut-off of 4.5. The FASTQ files for each sample sequencing run were processed through a modified ERVmap pipeline to generate count files for ERV and gene expression using bioinformatics tools bedtools, samtools, STAR align, BBMap and pythons htseq-counts module then processed through DESeq2 Differential Expression analysis to estimate the change in gene expression between ALS and non-ALS control postmortem primary motor cortex tissue (see section 2.2.15). In order to prevent spurious results from occuring in the data from low expressed ERVs these ERV members were filtered out (ERVs must have at least 10 counts in at least 5 samples to be included) during the analysis process. The primary output for this differential expression analysis was log2 fold expression changes in ALS tissue compared to controls for each ERV identified on ERVmap's bed file.

This data is displayed in Table 6.1 below shows ERVs identified in the ERVmap.bed file (a reference file for Bedtools which identifies ERVs by chromosome and base pair region) that showed significant changes in gene expression by uncorrected p-value from ALS and Non-ALS Controls. The adjusted p-value (cut-off $p<0.05$) calculated by DeSeq2 uses the Benjamini-Hochberg procedure which helps prevent the reporting of false positives by testing the p-value across multiple comparisons. This allows us to control for the likelihood of type 1 errors occuring in the data (incorrect rejection of the null hypothesis), allowing for the "p-value" results to be discounted from the results set. Using this adjusted p-value (cut-off $padj<0.05$) we can see that none of the ERVs identified as being significant by p-value ($p<0.01$) passed validation by testing for multiple comparisons. However, these ERVs were still analysed for potential relationships with ALS and Neurological conditions.

Table 6.1 DESeq2 Differential Expression Results for Statistically Significant Endogenous Retroviruses in Postmortem Primary Motor Cortex ALS and Non-ALS Controls.

The table below shows Log2 fold changes in expression of ERVs identified as being significant in non-adjusted p-value between ALS and non-ALS controls. Base mean is the mean counts of the gene from all samples analysed by Deseq2, lfcSE is the standard error of the Log2Fold change, padj refers to the adjusted p-value of the differential expression result, and this pvalue is the probability that the log fold is due to random chance.

ERVmap ID	Chromosome base pair Location (GRCh38 Assembly)	Chr Locus ID	UCSC Identified ERV, UCSC ERV Family	Base Mean	log2 Fold Change	lfcSE	p-value	p-adj
1412	chr4:55,938,234-55,948,225	4q12	HERV4_I, ERV1 (5,879 bp)	13.40860	2.99299	0.86257	0.00052	0.25211
4160	chr14:57,367,152-57,374,327	14q22.3	HERV-K9, ERVK (7,176 bp)	18.33260	-1.37095	0.39664	0.00055	0.25211
2458	chr7:33,158,829-33,174,607	7p14.3	MER57, ERV1 (15,779 bp)	66.98121	-0.76481	0.22643	0.00073	0.25211
2658	chr7:143,105,116-143,110,504	7q34	HERV-H, ERV1 (5,389 bp)	13.72757	-2.42340	0.72481	0.00083	0.25211
4757	chr19:51,804,687-51,811,880	19q13.41	HERV-K3 (7,194 bp)	200.50423	0.75984	0.24346	0.00180	0.43942
4506	chr18:11,130,174-11,134,262	18p11.21	HERV-L, ERVL (4,089 bp)	21.07724	-1.67846	0.56611	0.00303	0.61512
4864	chr21:15,979,607-15,985,359	21q21.1	HERV-H, ERV1 (5,753 bp)	11.05253	-1.91797	0.66996	0.00420	0.65932
4639	chr19:20,362,039-20,371,673	19p12	HERVS71, ERV1 (9,635 bp)	47.43545	-1.01826	0.35880	0.00454	0.65932
2699	chr8:7,152,779-7,159,835	8p23.1	HERV-K11 (7,057 bp)	104.55859	0.73910	0.26547	0.00537	0.65932
2010	chr5:144,609,328-144,616,900	5q31.3	HERV-H, ERV1 (7,573 bp)	5.17784	2.20854	0.79397	0.00541	0.65932
935	chr3:44,542,481-44,550,763	3p21.31	HERV9, ERV1 (8,283 bp)	16.96176	1.20548	0.44838	0.00718	0.73572
2294	chr6:114,010,995-114,018,008	6q21	HERV-H, ERV1 (7,014 bp)	18.17520	-1.05851	0.39562	0.00746	0.73572
4843	chr20:44,670,228-44,679,205	20q13.12	HERV-H, ERV1 (8,978 bp)	40.28646	-0.81154	0.30525	0.00785	0.73572
2548	chr7:86,087,732-86,093,417	7q21.11	HERV-L, ERVL (5,686 bp)	14.80064	1.38627	0.52963	0.00886	0.77141

The ERVs identified as having a significant uncorrected p-value (cut-off p<0.01) as shown in Table 6.1 were analysed further, to look for potential differentially regulated genes up or downstream of the ERV site and whether they were related to neurodegenerative disorders. There were no genes which were up or downstream of the proviral insert site related to neurodegenerative disease. Other ERVs close to genes which were found to be unrelated to developmental diseases or neurodegeneration, as described by their entry on GeneCards, include ERV 2658 (HERV-H, 21kbps upstream of PIP, Prolactin Induced Protein), ERV 4757 (HERV-K3, in intron of FPR3, Formyl Peptide Receptor 3), ERV 4864 (HERV-H in intron of lncRNA MIR99AHG, Mir-99a-Let-7c Cluster Host Gene also 99kbps downstream of USP25, Ubiquitin Specific Peptidase 25), ERV 4639 (HERVS71, situated approximately 160kbp upstream of ZNF737 (Zinc Finger Protein 737) and approximately 160kbp downstream of ZNF486 (Zinc Finger Protein 486) both involved in NA binding and HSV1 infection), ERV 2699 (HERV-K11, 94kbp downstream of DEFA5, Defensin Alpha 5), ERV 2010 (HERV-H, 122 kbps downstream of KCTD16, Potassium Channel Tetramerization Domain Containing 16), ERV 935 (HERV9, 4kbps upstream of ZKSCAN7, Zinc Finger With KRAB And SCAN Domains 7), ERV 4843 (HERV-H, 18kbp downstream of ADA, Adenosine Deaminase) and ERV 4160 (HERV-K9, 16kbps upstream of NAA30, N-Alpha-Acetyltransferase 30, NatC Catalytic Subunit).

The only ERVs that have significant uncorrected p-value log2fold changes that are close to genes known neurological condition determinants are ERV 2548 (HERV-L, 547kbp Upstream of GRM3, an L-glutamate receptor in neural transmission implicated in Bipolar Disorder and Schizophrenia), ERV 2294 (HERV-H, 36kbp upstream of HS3ST5 (Heparan Sulfate-Glucosamine 3-Sulfotransferase 5), associated with Mental Retardation) and ERV 4506 (HERV-L, within intron of PIEZO2 (Piezo Type Mechanosensitive Ion Channel Component 2) associated with Marden-Walker Syndrome a developmental disease of the CNS). While all ERVs transcripts with significant non-adjusted p-values appear to be larger fragments of, or close to full length, HERV genomes, they do not have any open reading frames for viral proteins in their genetic code (Supplementary Figures S104-S117).

As the relationship between retroviral transcription and *TARDBP* activity has been previously explored in the literature and included during our RT-qPCR analysis of HERV expression, the differential expression of *TARDBP* (*TARDBP*) and *BCL11b* (which has been

found to effect ERV translation in CNS HIV infections) was also investigated in the RNA seq data. Table 6.2 shows the differential expression of these genes as analysed by the DESeq2 differential expression algorithm. As we can see by both p-value and adjusted p-value the expression for both *TARDBP* and *BCL11b* between ALS and controls was not significant.

Table 6.2. DESeq2 Differential Expression for *TARDBP* and *BCL11b* in Postmortem Primary Motor Cortex tissue samples Between ALS and Non-ALS Controls.

The table below shows Log2 fold changes in expression of *TARDBP* and *BCL11b* between ALS and non-ALS controls. Base mean is the average of the normalized count values, dividing by size factors, taken over all samples of the gene from all samples analysed by Deseq2, lfcSE is the standard error of the Log2Fold change, padj refers to the adjusted p-value of the differential expression result, and pvalue is the probability of the log fold change occuring due to random chance.

Ensembl Gene ID	Gene	baseMean	log2Fold Change	lfcSE	pvalue	padj
ENSG00000120948	*TARDBP*	430.8253722	-0.21696	0.16136	0.17876	1
ENSG00000127152	*BCL11b*	425.3931435	-0.03798	0.19506	0.84561	1

In order to visualise the spread of log2 fold changes of ERV genomic data between ALS and non-ALS controls an MA plot was constructed (Figure 6.1.). ERVs that have a p-value of less than 0.01 would appear as red points in Figure 6.1, with ERV members that fall outside of the range indicated by an upwards or downwards facing arrow at the top and bottom edges of the MA plot. The mean of normalised counts are normally distributed. Of those 15 samples that pass the cut-off on p-value, while not being significant based on adjusted p-value 2 seem close to or above the length of theoretically intact ERV sequences, with ERV 2458 being 15779bp in length and ERV 4649 being 9635bp in length. Other ERVs identified appear to be approximately between 5389 to 8978 bp in length, shorter than the predicted size for full gag-pol-env HERV genome.

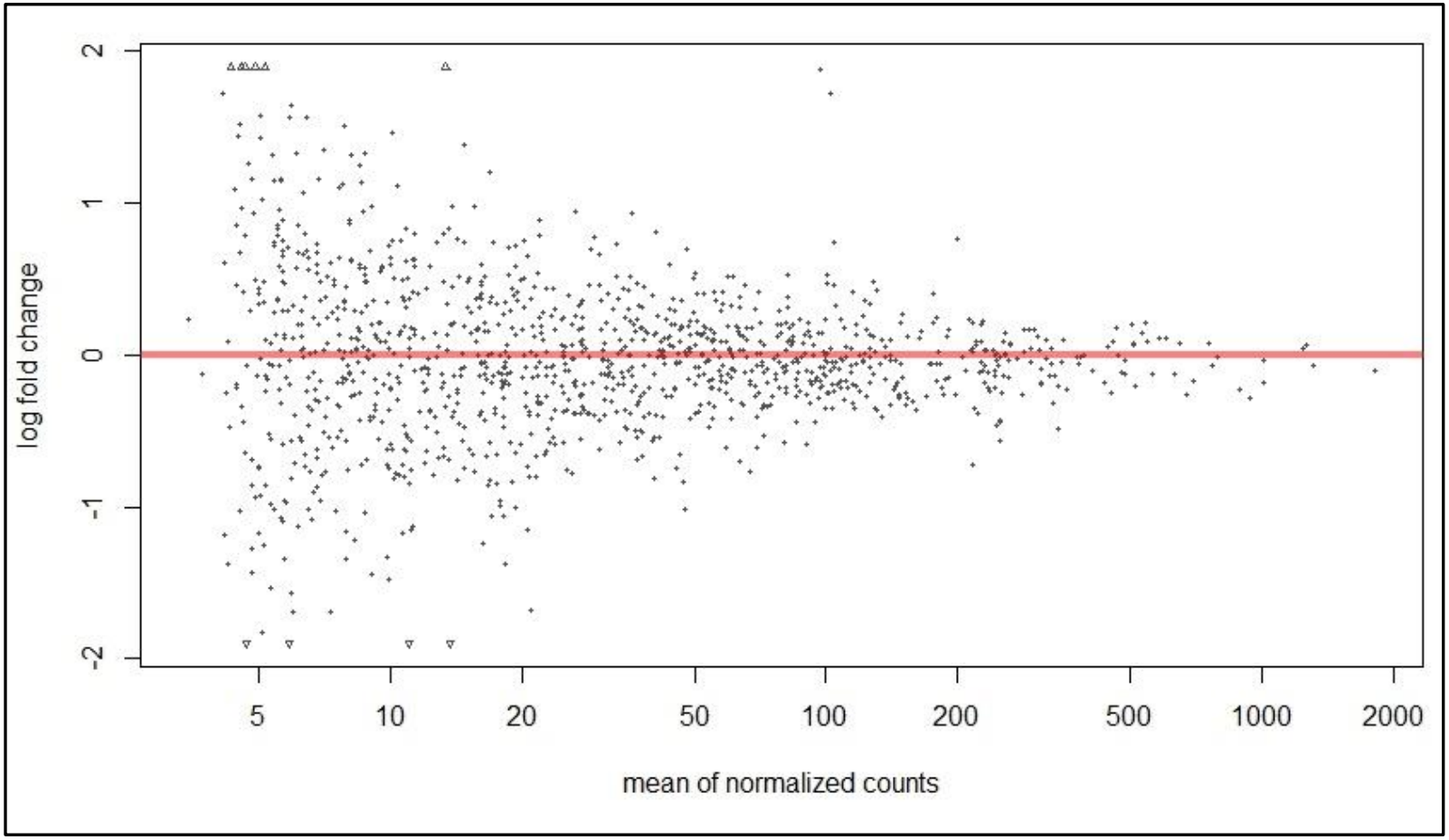


Figure 6.1 MA Plot of Log2 Fold Changes in Expression between in Postmortem Primary Motor Cortex ALS and Non-ALS Controls for ERVs Identified in the ERVMap.bed file.

The MA plot in the figure above shows the distribution of Log2 fold changes against the mean of normalised counts for individual ERVs identified in ALS and non-ALS controls. Statistically significant ERVs (p-value less than 0.01) would be identified by red points if present.

6.2.2 Variation in DESeq2 Differential Expression Data Derived from Postmortem Primary Motor Cortex Tissue Samples.

In order to verify the validity of our DESeq2 normalised RNA-Seq data we can look at how much the count values deviate from the mean. Figure 6.2 shows the dispersion estimate plot for the RNA-Seq data, the individual ERVs are plotted on the graph in black dots and their shrinkage (blue arrows pointing towards the red mean line in Figure 6.2) towards the mean line is plotted with blue arrows (too small to see on plot). These represent the distance of the dispersion estimate for the individual ERV across all samples to the mean line. Shrinkage of the values also helps with eliminating potential false positives from the data. As we can see in the figure below the majority of samples lie close to the mean line with the few outliers plotted above main distribution of data (shown by black dots with blue circles around them).

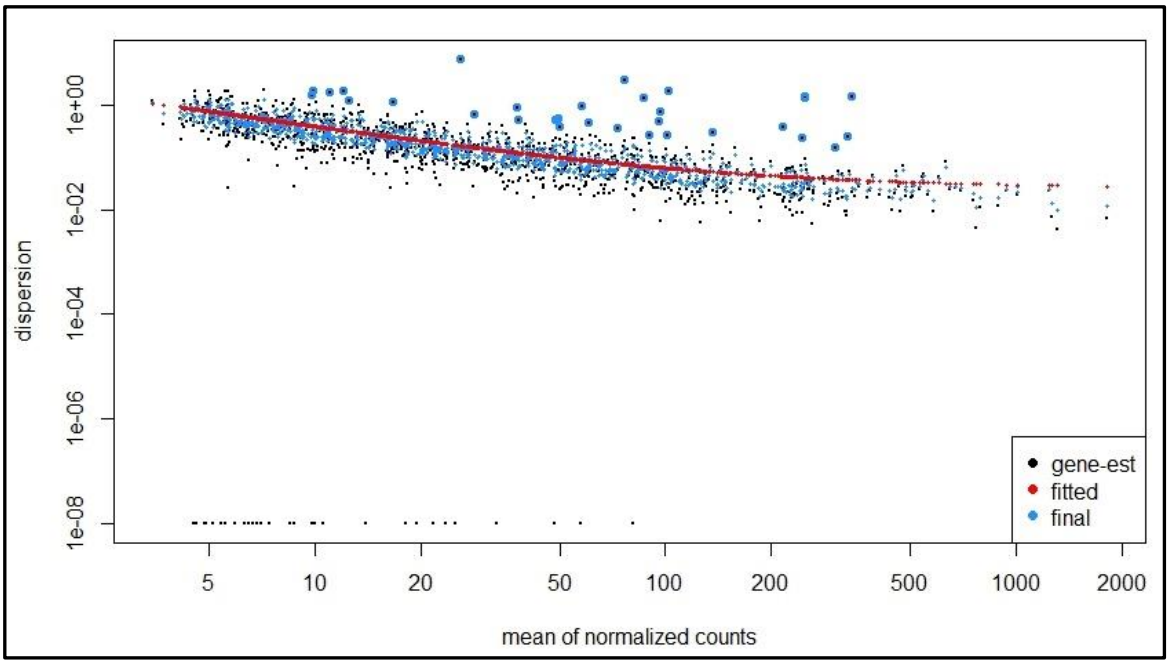


Figure 6.2. Dispersion Estimate Plot for Differential Expression Data Generated by DESeq2 from Postmortem Primary Motor Cortex ALS and Non-ALS Controls.
The Figure above plots dispersion estimates for endogenous retrovirus differential expression data from ALS and Non-ALS control tissue using *post-mortem* primary motor cortex samples over the mean count of those ERVs This is shown as shrinkage towards the red line (consensus) for non-outlier data points. The black points circled in blue above the main data set are defined as dispersion outliers and not shrunk towards the consensus line.

In Figure 6.3 below we can see the frequency distribution of p-values within the RNA-Seq data. With differential expression analyses you could expect to see a binomial distribution of p-values, with peaks at both extremes of the p-value frequencies. This is due to some analyses incorporating no reads mapped in any sample recording a probability value of 1.0, shifting the weight of p-value frequencies towards 1.0. However, the data displayed in the graph shows uniform distribution, with most of the p-values evenly spread across all p-values. This provides further evidence that ERVs are not significantly differentially expressed in ALS when compared with controls.

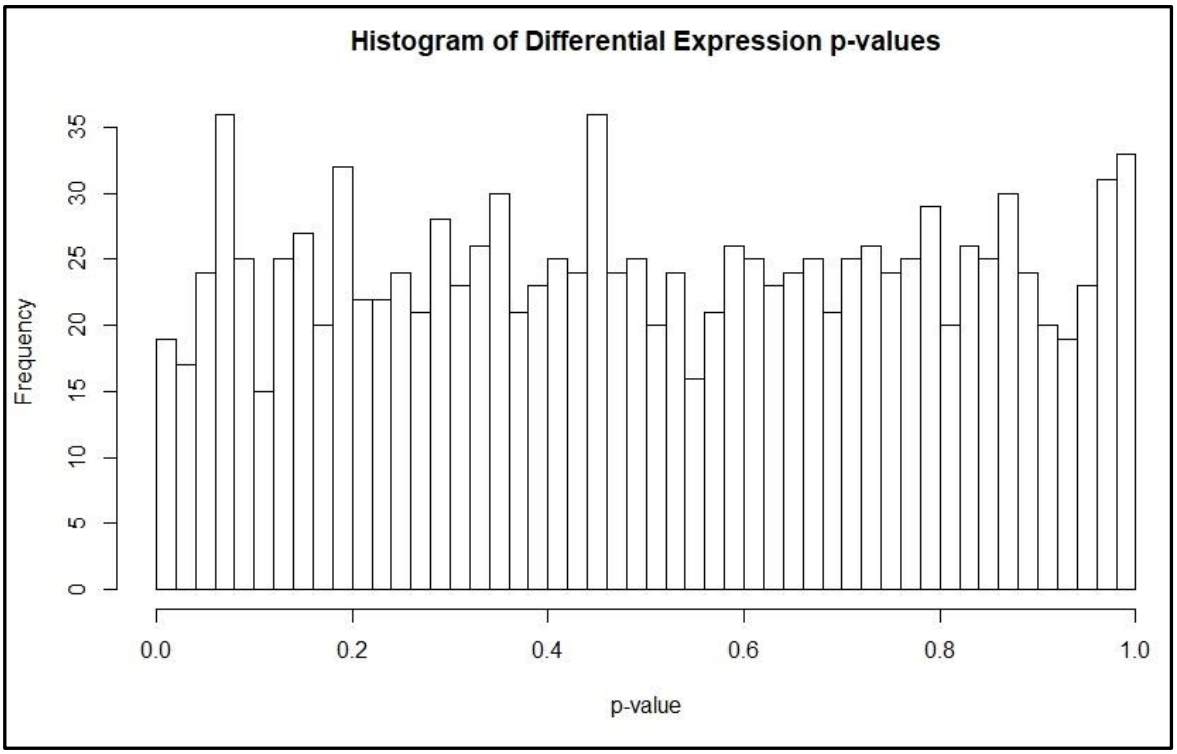


Figure 6.3. Histogram of p-Value Frequency within DESeq2 Differential Expression Analysis

The figure above displays the frequency at which a given p-value occurs within the ALS vs Non-ALS control differential expression sample set. The p-values, while varied at different values broadly shows a uniform distribution in the sample set. This indicates that the data has no statistically significant differential expression values for ERVs in ALS when compared with controls.

6.2.3 Analysis of ERV Expression Profiles Between ALS and Non-ALS Controls.

In addition to looking at the spread of values within the RNA-Seq dataset we can also look at the clustering of expression data to see if we can observe any trends between samples or families of ERVs. A method of interpreting RNA-Seq expression data is using a heatmap to visualise the difference in expression of significant genes between samples. In Figure 6.4 below we have a heatmap of normalised counts for ERVs identified by the ERVmap.bed file. The data displayed in Figure 6.4 has been hierarchically clustered to show similarities between samples (columns) and ERVs (rows), with most of the normalised counts data appears to be relatively low. This is due to the low spread of values in the counts (5.5-8.5) amongst the filtered ERVs. Additionally, while some ALS and non-ALS control samples are clustered together (orange and purple cells at the top of the heatmap), there are still some ALS samples which are closer to their non-ALS controls than other ALS samples.

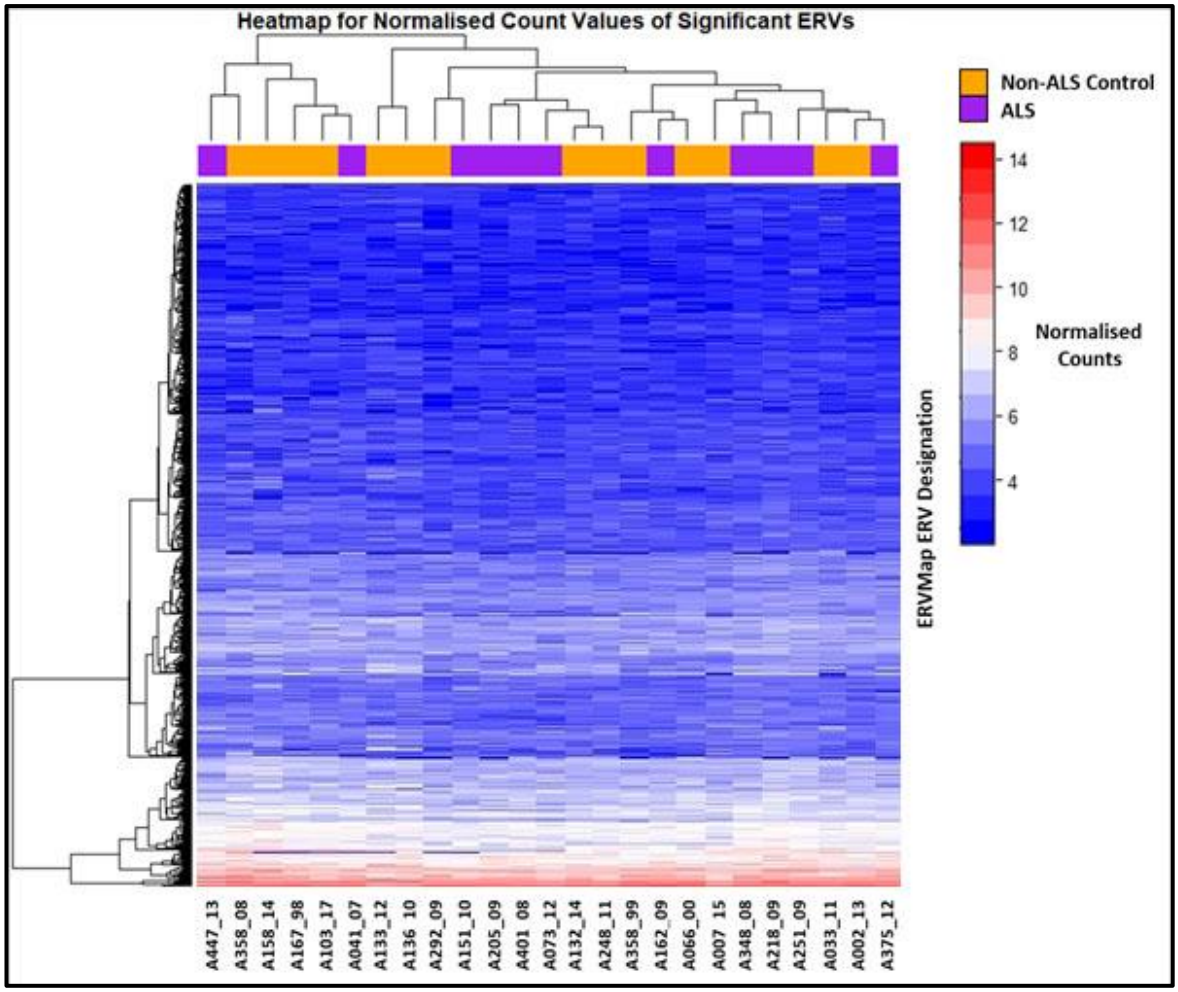


Figure 6.4. Heatmap of Normalised counts for ERVs in Postmortem Primary Motor Cortex ALS and Non-ALS Control Samples.

The heatmap displayed in the figure above shows the normalised counts data for ERVs identified by the ERVmap.bed file with low expressed ERV members filtered out. The rows and columns are hierarchically clustered to group together samples and ERVs with similar expression profiles based on normalised counts data generated from DESeq2. Also included in the cells above the counts matrix identifies those samples which are from the ALS (purple) and non-ALS control (orange) sample sets.

A different way of viewing the grouping of samples for disease status and other phenotypic data is the principal component analysis (PCA) plot. This plot reduces the normalised counts (3220 per sample for ERV counts) down to their principal components or the relationship of the weight of counts for each ERV within each sample. This allows us to easily group our samples based on how similar their gene expression profiles are for each ERV, as shown in Figure 6.5. The PCA plots in Figure 6.5 below show a combined comparison for disease status and sex of patient (Figure 6.5A) and differing RIN (RNA integrity number, Figure 6.5B) in the patient samples. Figure 6.5A serves as a representative example for the rest of the

comparisons, including age of patient at time of death and postmortem delay all of which show no discernible grouping in any of these factors (Supplementary Figures S169-170). The only notable difference between in sample groups is when we look at the distribution of RIN values in the PCA plot. As we can see in Figure 6.5B the samples appear to be grouped in regions correlating to their samples derived RIN value, indicating that the RNA integrity of the samples taken from the primary motor cortex influences the measured expression of ERV members (Section 2.2.3).

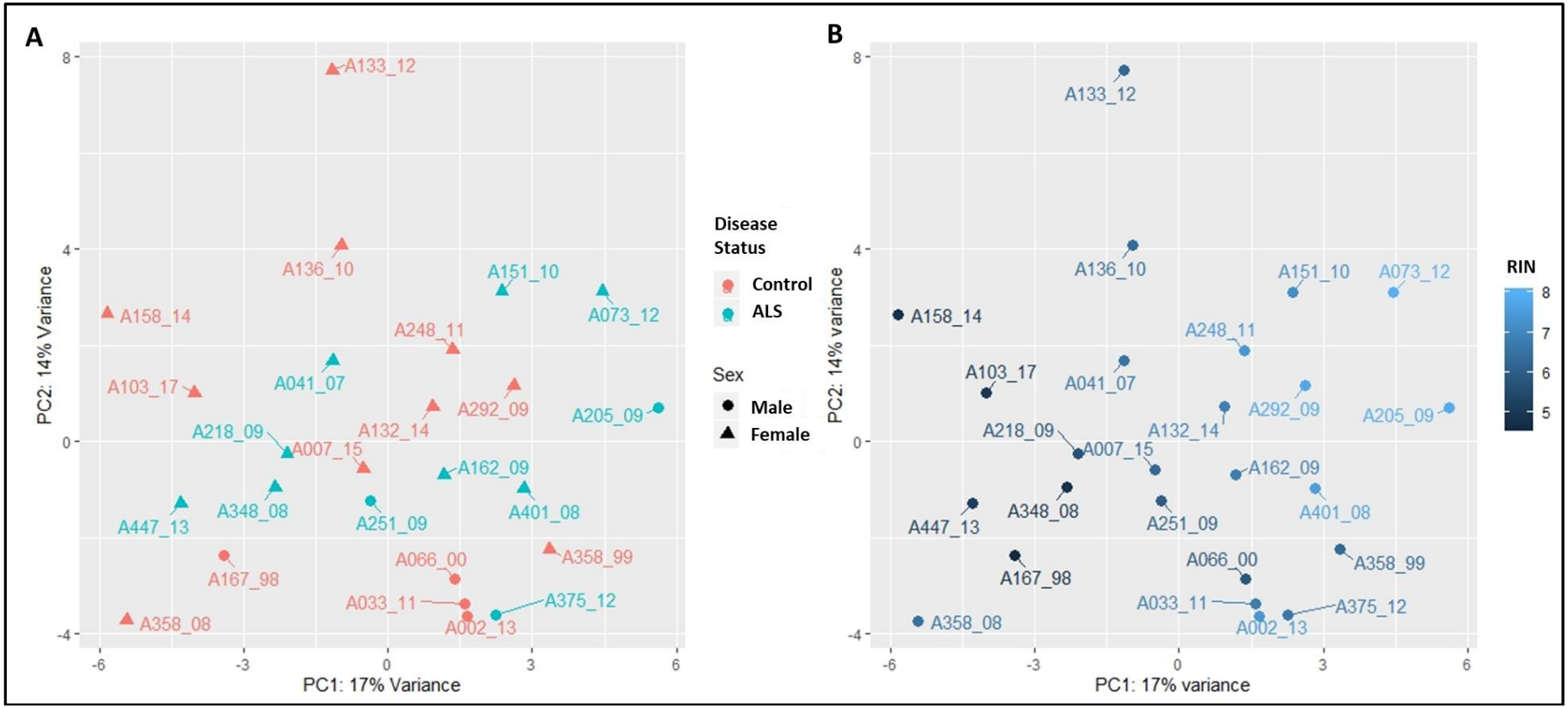


Figure 6.5. Principal Component Analysis (PCA) plots for DESeq2 Generated Normalised Counts from Differential Expression Analysis of Postmortem Primary Motor Cortex Tissue from ALS and Non-ALS Controls.

The principal component analysis (PCA) shown in the figure above plots variation between all ERVs in each sample using normalised counts for each gene per patient sample. These principal component values are then plotted against each other to group samples together based on similar expression profiles of ERVs. **PCA plot A identifies samples based on both sex and disease** status while **PCA plot B shows RIN values** for the samples.

Another way of visualising normalised count data between samples, in order to look for variation in counts recorded between samples is the box plot. Figure 6.6 below plots the distribution of counts for those ERVs that were preferentially filtered by DeSeq2 as having expression above the cut-off in more than 5 of the samples within the set as part of the pipeline. As shown in Figure 6.6 there is slight variation the normalised counts between samples with an even mix of ALS and non-ALS control samples (A132_14 (Non-ALS), A218_09 (ALS), A292_09 (Non-ALS) and A251_09 (ALS)) appearing to be slightly shifted in comparison to the distribution in other samples. While these samples appear to have significantly different spreads in count data there is no factor in the patient metadata (Age, Sex, PMD, RIN) to explain this variation. While this variation is present the general variation in the data is not shifted too far from the group to indicate an issue with the analysis.

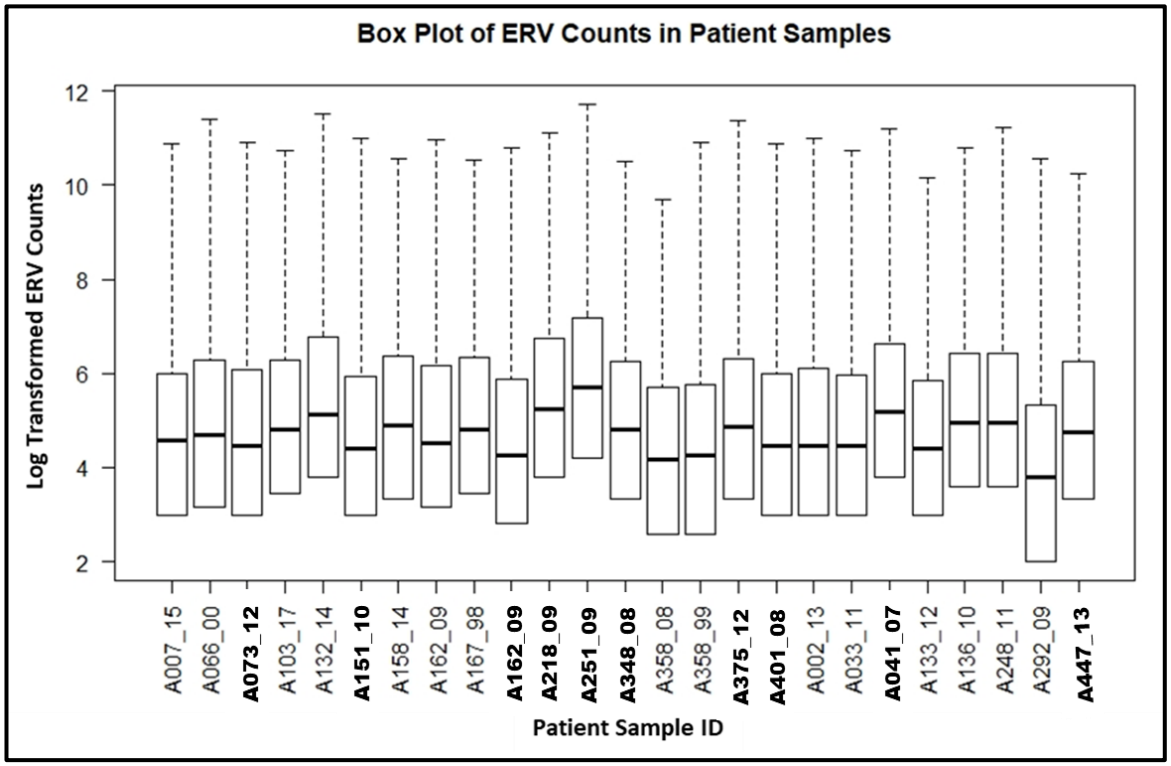


Figure 6.6. Box Plot of Endogenous Retrovirus Normalised Counts between n=11 ALS and n=14 Non-ALS controls.

The figure above displays statistical information on the counts data for ERVs within the n=11 ALS and n=14 non-ALS control sample set. The thick line inside each of the plots shows the median value for the data while the upper and lower end of the bars represent the interquartile ranges. The "whisker" above the plot, represented by the dotted line, shows the highest count recorded for an individual ERV within the sample. ALS samples in the Box plot have been differentiated by the sample ID presented in bold type.

6.2.4 Analysis of Publicly Available Central Nervous System (CNS) RNA-Seq Datasets for Endogenous Retrovirus (ERV) Expression

Raw fastq files for *post-mortem* frontal cortex and cerebellum samples were obtained using the NCBI (National Centre for Biotechnology Information, USA) Sequence Read Archive (SRA) publicly available ALS vs non-ALS Control dataset generated from the study detailed in the paper by Prudencio *et al.* (2017). The dataset includes n= 18 ALS and n=9 non-ALS control patients, the Frontal Cortex and Cerebellum samples were both taken from the same patient from the specific areas of the brain allowing for the 2 datasets, with n=8 ALS samples identified as having the C9orf72 mutation. The samples were extracted from frozen frontal cortex and cerebellum tissue, with 20-30mg used for RNA extraction. RNA was purified from the samples using RNeasy Plus Mini Kit (QIAGEN, Germany) and quantified using Agilent Bioanalyser 2100, with samples exceeding a RIN value of 7.0 used for library preparation. Sequencing was performed on Illumina's HiSeq 2000 platform (Prudencio *et al.*, 2017).

For the modified ERVMap analysis pipeline those frontal cortex and cerebellum samples that were identified as sporadic ALS only were selected, excluding those ALS samples identified as having the C9orf72 mutation as this analysis is looking for differential expression of genes in sALS cases with no known mutation. This is due to published results stating that the C9orf72 mutation only accounts for 5%-10% of patients with the sporadic form of the disease so the initial analysis looks at the expression data without the bias of the n=8 C9orf72 ALS samples (Umoh *et al.*, 2016). This resulted in n=10 ALS & n=9 non-ALS control samples from the frontal cortex and n=10 ALS & n=8 non-ALS control samples from the cerebellum.

The cerebellum revealed 2 ERVs that were differentially expressed by our **adjusted p-value** cut-off as shown in Table 6.3. However, ERV 4760 shows a particularly high log2fold change (6.64) which is likely the result of low counts mapped to the particular ERV across multiple samples (as evidenced by the low sample mean). Table 6.4 shows there is a single ERV that is differentially expressed by both **adjusted p-value** and regular **p-value** cut-offs in the frontal cortex tissue. Looking at the mean of average counts for these individual ERVs it appears that the HERV-K22 identified in the cerebellum has a higher count mean, indicating

that its expression is higher than the HERV-H found in the Frontal Cortex. A potential reason for this difference in mean of average counts is seen in the log2fold column, the minus value in the HERV-H table indicates that this ERV is downregulated in the ALS samples compared to controls while the HERV-K22 is upregulated.

As there are many similarities in loci in sex chromosomes due to their shared evolutionary history it is possible that differences in reads aligning to these chromosomes could potentially lead to errors in the reporting of differential expression in DESeq2 (Olney *et al.*, 2020). In order to test whether the ERVS listed in the Table 6.3 & 6.4 below are present due to similarities in sex chromosomes those ERVs loci present on sex chromosomes had their counts removed from the counts matrix and DESeq 2 analysis re-run. In this modified analysis ERV 2152 and 3316 were still present in the analysis though 4760 was no longer significant (data not shown).

Quality control for the dataset was performed for both tissue types following the template set out in sections 6.2.1-6.2.3. Figures detailing the quality control aspects are shown in Supplementary Figures S208-S220 and while they do show some variation in comparison to the preceding section the dispersion of the data does not show anything which could compromise the validity of the results shown in the table below. The PCA plots for the expression data also shows no distinctive grouping based on disease status.

Table 6.3. DESeq2 Differential Expression Results for Statistically Significant Endogenous Retroviruses in Postmortem Cerebellum Tissue Between ALS and Non-ALS Controls.

The table below shows Log2 fold changes in expression of the single significant ERV identified from an Adjusted P-value cut-off of 0.05 with p-values below 0.01. Base mean is the mean counts of the gene from all samples analysed by Deseq2, lfcSE is the standard error of the Log2Fold change, padj refers to the adjusted p-value of the differential expression result, and this pvalue is the probability of the log fold change occuring due to random chance.

ERVmap ID	Chromosome base pair Location (GRCh38 Assembly)	Chr Locus ID	UCSC Identified ERV, UCSC ERV Family (base pair length)	Base Mean	log2 Fold Change	lfcSE	p-value	p-adj
4760	chr19:52,300,606-52,307,977	19q13.31	HERV-K9, ERVK (7,372bp)	22.82609	5.637699	1.394141	5.26E-05	0.029059
2152	chr6:57,069,286-57,080,996	6p12.1	HERV-K22, ERVK (11,711bp)	402.8291	1.708243	0.438566	9.82E-05	0.029059

Table 6.4. DESeq2 Differential Expression Results for Statistically Significant Endogenous Retrovirus in Postmortem Frontal Cortex Tissue Between ALS and Non-ALS Controls.

The table below shows Log2 fold changes in expression of the single significant ERV identified from p-values below 0.01. Base mean is the mean counts of the gene from all samples analysed by Deseq2, lfcSE is the standard error of the Log2Fold change, padj refers to the adjusted p-value of the differential expression result, and this pvalue is the probability of the log fold change occuring due to random chance.

ERVmap ID	Chromosome base pair Location (GRCh38 Assembly)	Chr Locus ID	UCSC Identified ERV, UCSC ERV Family (bp size)	Base Mean	log2 Fold Change	lfcSE	p-value	p-adj
3316	chr10:89,284,719-89,291,763	10q23.31	HERV-H, ERV1 (7,045bp)	163.6134	-1.58329	0.289473	4.51E-08	$1.51x10^{-05}$

6.2.5 Gene Set Enrichment and Functional Pathway Analysis for Genes within 1MB Up/Downstream of Proviral Insertion Site for Cerebellum and Frontal Cortex Tissue Data

The single HERV-K family member identified as being differentially expressed by the analysis of Cerebellum tissue samples and was shown to be significantly upregulated by adjusted p-value (cut-off <0.05). This ERVMap ID 2152, identified as HERV-K22 while its genome is close to the theoretical length of an intact HERV-K provirus, the translated protein sequence does not show any intact open reading frames for functional proteins. A summary of cellular genes within 1MB upstream and downstream of the differentially expressed proviral sequences is given in Table 6.5 below (Macfarlane and Badge, 2015). The closest upstream gene to the provirus is KIAA1586 (13kbp upstream), unrelated to neurological conditions. The closest downstream gene to the proviral insertion site isZNF451 (Zinc Finger Protein 451, 15kbp downstream) which is associated with a neurological condition, Peroneal Neuropathy, but is not known to be differentially regulated in ALS. While this thesis has previously focused on the HML-2 group of HERV-K family members to compare to the 2015 Li *et.al.* paper the HERV-K22 family member has been documented as belonging to the HML-5 group of proviruses. The other ERV found to be differentially expressed in Cerebellum tissue is HERV-K9 (ERVMAP ID 4760). The majority of the genes within 1MB up/downstream of the proviral insertion site have no disease associated or disease linked to neurological conditions (Macfarlane and Badge, 2015). This distance is due to the effect of proviral sequences on the transcription of genes within this distance of the insertion site. The exceptions are MicroRNA 125a (Alzheimers), PPP2R1A (Mental Retardation), ZNF611 (X-Linked Intellectual Disability).

The ERV found to be significantly differentially expressed in the Frontal cortex is ERVMap ID 3316, identified as HERV-H. The proviral sequence is contained within the intron of a longer LIPA (Lipase A, Lysosomal Acid Typ) transcript, and its principal function is in lipoprotein metabolism. The closest gene outside of alternate LIPA transcripts in the region is 9kbp downstream from the insertion site and is annotated as IFIT2 (Interferon Induced Protein With Tetratricopeptide Repeats 2), while the known diseases associated with this protein are unrelated to neurological conditions the protein has function in RNA binding, potentially useful in proviral replication (*IFIT2 Gene - GeneCards | IFIT2 Protein | IFIT2 Antibody*, 2022).

Table 6.5. Annotated Genes within 1MB Up/Downstream of Proviral Insertion Sites for Significant Differentially Expressed ERVs in Frontal Cortex and Cerebellum Publicly Available RNA-Seq Data

The data given in the table below uses the UCSC genome browser to track annotated genes within 1MB up/downstream of the proviral insertion site. The table shows the annotation for the gene and its distance from either the 5' end of the provirus for upstream genes or the 3' end of the provirus for genes appearing downstream of the insertion site.

ERV	Closest Gene Upstream (kbp distance from 5' LTR)	Closest Gene Downstream (kbp distance from 3' LTR)	Genes Within 1MB Upstream	Genes Within 1MB Downstream
4760	Within intron of ZNF480, Overlaps start of of element AC010320.4		AC010320.3 (2kbp), ZNF766 (4kbp), MIR643 (19kbp), Y_RNA (41kbp), AC010320.2 (68kbp), PPP2R1A (72kbp), MIR6801 (79kbp), AC010320.1 (122kbp), ZNF836 (128kbp), AC011468.2 (129kbp), AC011468.3 (157kbp), ZNF616 (161kbp), ZNF841 (203kbp), AC011468.1 (234kbp), ZNF432 (249kbp), AC011468.5 (250kbp), ZNF614 (271kbp), ZNF615 (292kbp), ZNF350 (311kbp), ZNF350-AS1 (321kbp), ZNF613 (354kbp), ZNF649 (396kbp), ZNF649-AS1 (397kbp), ZNF577 (411kbp), AC006272.1 (459kbp), FPR3 (475kbp), AC018755.5 (527kbp), FPR2 (530kbp), FPR1 (546kbp), HAS1 (575kbp), SPACA6 (594kbp), MIR99B, MIRLET7E, MIR125A, SPACA6P-AS (606kbp, all lie within 200-300bp region), SIGLEC5, AC018755.2 (649kbp), SIGLEC14 (652kbp), AC018755.4 (660kbp), AC018755.1 (699kbp), ZNF175 (706kbp), SIGLEC6 (766kbp), AC020914.1 (777kbp), SIGLEC12 (797kbp), CEACAM18 (808kbp), SIGLEC8 (838kbp), SIGLEC10 (879kbp), SIGLEC10-AS1 (881kbp), AC008750.1 (885kbp), NKG7 (927kbp), AC008750.7 (929kbp), ETFB (931kbp), AC008750.4 (932kbp), AC008750.3 (944kbp), AC008750.2 (952kbp), VSIG10L (954kbp), IGLON5 (968kbp)	ZNF610 (28kbp), ZNF880 (61kbp), ZNF528-AS1 (79kbp), ZNF528 (91kbp), ZNF534 (120kbp), ZNF578 & AC010332.3 (143kbp), AC022150.3 (202kbp), ZNF808 (220kbp), ZNF701 (246kbp), AC022150.1 (267kbp), ZNF83 (287kbp), AC022150.2 (289kbp), AC022150.4 (340kbp), ZNF611 (394kbp), ZNF600 (455kbp), ZNF28 (489kbp), ZNF468 (528kbp), ZNF320 (549kbp), ZNF888 (595kbp), ZNF816-ZNF321P (617kbp), ZNF816 (639kbp), AC010328.3 (647kbp), AC010328.1 (697kbp), ERVV-1 (706kbp), ERVV-2 (735kbp), ZNF160 (755kbp), ZNF415 (797kbp), ZNF347 (813kbp), ZNF665 & AC092070.1 (853kbp), ZNF677 (925kbp), AC092070.3 (930kbp), VN1R2 (946kbp), VN1R2 (954kbp), AC092070.4 (955kbp)

Table 6.5. (Continued) Annotated Genes within 1MB Up/Downstream of Proviral Insertion Sites for Significant Differentially Expressed ERVs in Frontal Cortex and Cerebellum Publicly Available RNA-Seq Data

The data given in the table below uses the UCSC genome browser to track annotated genes within 1MB up/downstream of the proviral insertion site. The table shows the annotation for the gene and its distance from either the 5' end of the provirus for upstream genes or the 3' end of the provirus for genes appearing downstream of the insertion site.

ERV	Closest Gene Upstream (kbp distance from 5' LTR)	Closest Gene Downstream (kbp distance from 3' LTR)	Genes Within 1MB Upstream	Genes Within 1MB Downstream
2152 (Cerebellum)	KIAA1586 (13kbp)	ZNF451 (15kbp)	BEND6 (40kbps), DST (215kbp), COL21A1 (674kbp), AL031779.1 (734kbps)	BAG2 (96kbp), RAB23 (112kbp), PRIM2 (239kbp), MIR548U (314kbp), FO680682.1 (418kbp), AL021368.3 (782kbp), AL021368.5 (804kbp), AL021368.1 (827kbp), AL021368.2 (833kbp), AL021368.4 (846kbp), AL445250.1 (886kbp)
3316 (Frontal Cortex)	Within last Intron of longer LIPA transcript, overlaps start of longer IFIT2 transcript and encompasses nearly all of element AL353751.1 (all annotated overlapping this ERV)		LIPA (Shorter Transcripts, 32kbp), FAS (268kbp), ACTA2 (294kbp), STAMBPL1 (360kbp), ANKRD22 (432kbp), LIPM (462kbp), LIPN (506kbp), LIPK (531kbp), LIPF (605kbp), LIPJ (677kbp), RNLS (701Kbp)	IFIT2 (shorter transcript, 9kbp), IFIT3 (38kbp), IFIT1B (85kbp), IFIT1 (103kbp), IFIT5 (121kbp), SLC16A12 (139kbp), PANK1 (288kbp), FLJ37201 (399kbp), KIF20B (Pseudogene, 410kbp), LINC00865 (538kbp), LINC01374 (564kbp), LINC01375 (625kbp), AL139340.1 (768kbp), RN7SKP143 (Pseudogene, 873kbp)

While the list of genes within the 1Mb window of the proviral insertion site is informative it does not give any information as to whether the genes are associated with the expression of the provirus at this particular locus in the human genome. In order to test for the gene set relationship to the provirus we can use gene set enrichment analysis (GSEA) to see which genes are overrepresented within this region by their log2fold change scores as calculated by DESeq2. In order to test these DESeq2 scores which are added on a genome wide bases irrespective of actual significance in the dataset to the gene set enrichment analysis program (Broad Institute, Massachusetts Institute of Technology, CA, USA) (Mootha *et al.*, 2003; Subramanian *et al.*, 2005). Gene sets were added to the program and run alongside the GSEA analyses to crop out the genes of interest for the enrichment plots. Only a single ERV provirus was found to be significant by the false discovery rate (FDR) cut-off of 0.25 (FDR = 0.19), ERV 3316 otherwise annotated as HERV-H (Figure 6.7). The gene enrichment plot in Figure 6.18 below shows the enrichment score (ES) of the genes found within 1Mb of the proviral insertion site (Table 6.5). From this wider set of genes only 2 genes were enriched, Ankyrin repeat domain 22 (ANKRD22, Running ES: 0.5408) whose gene ontology reported function is protein binding and Solute carrier family 16 member 12 (SLC16A12, Running ES: 0.3319) associated with transmembrane transporter activity. To test whether these 2 genes had any significant functional relationship they were analysed further using DAVID (Database for Annotation, Visualization and Integrated Discovery, Laboratory of Human Retrovirology and Immunoinformatics, USA) online tool which showed no functional relationship between the two genes (Huang, Sherman and Lempicki, 2009b, 2009a).

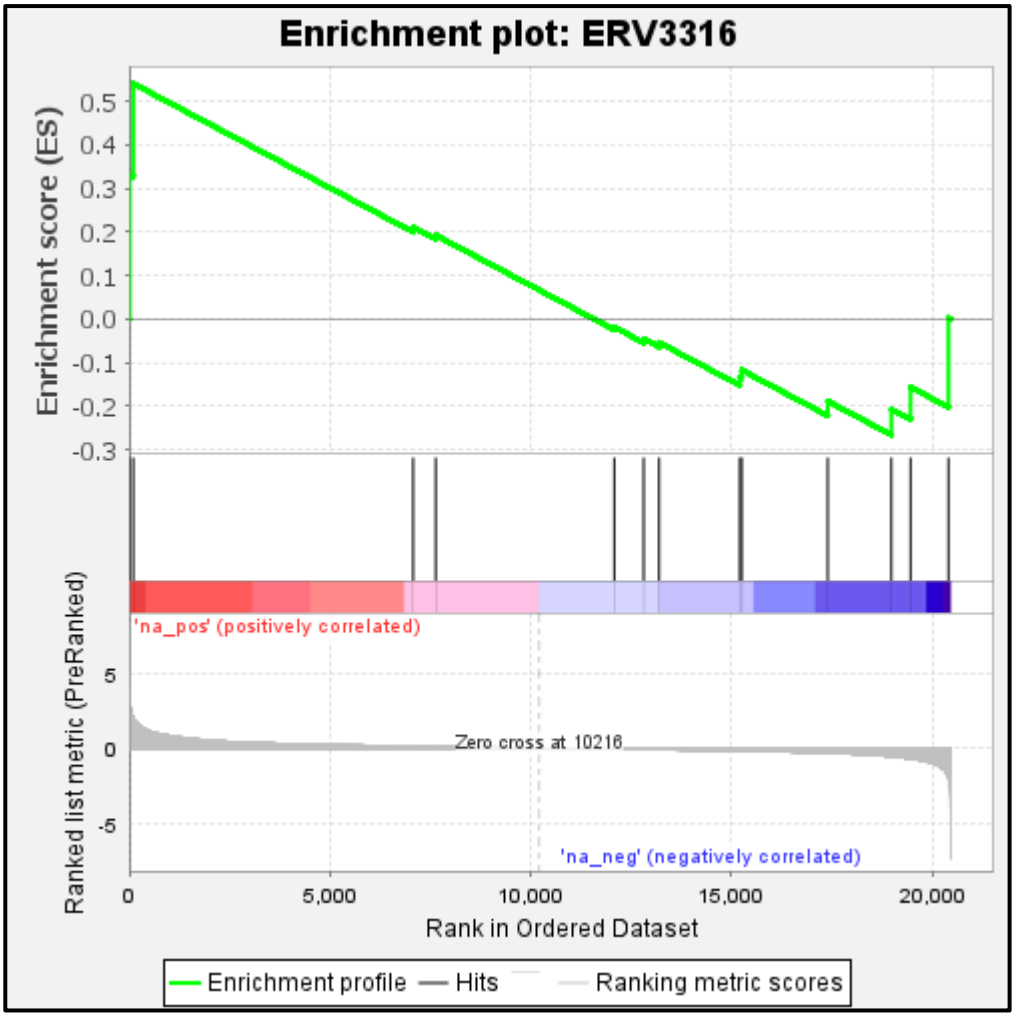


Figure 6.7. Gene Enrichment Plot for ERV3316 (HERV-H) in Frontal Cortex Dataset

The figure above shows the gene enrichment plot for the sole significant enriched gene set in the analysis from the publicly available RNA-Seq data from Prudencio *et.al.* (2017). This plot shows the positively correlated genes with the peak to the left of the graph representing the enrichment score (0.54) with the bold black lines below representing the list of genes within 1Mb of the proviral insertion site. The pre-ranked scores represent the log2fold change values of the individual genes in the full dataset.

As these genes were enriched in the region of a differentially expressed ERV locus their log2fold change statistics were obtained from the DESeq2 analysis and presented in Table 6.6 below. Despite being enriched in the sample set the 2 genes are not differentially expressed in ALS patient's vs controls.

Table 6.6. DESeq2 Differential Expression Results for ERV3316 Enriched Genes in Postmortem Frontal Cortex Tissue Between ALS and Non-ALS Controls.

The table below shows Log2 fold changes in expression for the enriched genes found within 1Mb of the 3316 proviral insertion site. Base mean is the mean counts of the gene from all samples analysed by Deseq2, lfcSE is the standard error of the Log2Fold change, padj refers to the adjusted p-value of the differential expression result, and pvalue is the probability of the log fold change occuring due to random chance.

Gene Symbol	Ensembl Reference	Base Mean	log2 Fold Change	lfcSE	p-value	p-adj
SLC16A12	ENSG00000152779	11.72296	3.179544	1.260698	0.011667	0.734807
ANKRD22	ENSG00000152766	14.63411	2.038803	1.402864	0.146136	1

As HERVs are known to have an effect on gene transcription within 1Mb up or downstream of their insertion site the next stage of the analysis was to look at whether these genes are co-expressed with the HERV-H (ERV3316) locus. As we can see from Table 6.7 below, despite not being significantly expressed in the DESeq2 differential expression analysis SLC16A12 showed a significant negative correlation with HERV-H expression. This means that as HERV-H expression is downregulated the expression of SLC16A12 in the locus is upregulated.

Table 6.7. R^2 and P-Values for Correlation Analysis Between Differentially Expressed ERV 3316 and Enriched Genes Within 1Mb of the Proviral Insertion site in Frontal Cortex Tissue Samples.

The table below shows the Pearson's r correlation of analysis results for significantly expressed HERV-H family member when compared to the expression of SLC16A12 and ANKRD22 between ALS and non-ALS controls and the p-value of the comparisons.

Ensemble ID	Gene Symbol	R2	P-value
ENSG00000152779	SLC16A12	-0.4869	0.0345
ENSG00000152766	ANKRD22	-0.1408	0.5654

6.2.6 Co-expression Analysis of RNA-seq data to look for correlation between HERV Expression and Transcriptional Regulators *TARDBP* and *BCL11b* in Frontal Cortex and Cerebellum Tissue.

Previous studies have shown a relationship between increased expression of nucleic acid binding protein *TARDBP* and increasing expression of ERVs, HERV-K in particular in sporadic ALS. The link between *BCL11b* and ERVs has not been categorised however, *BCL11b* has been shown to be highly expressed in the CNS and has a suppressive effect on HIV infection in the spinal cord. As we have looked at the potential expression of these genes in Chapter 4.0 alongside HERV-K (HML-2) expression we can apply co-expression analysis to see if there is a link in expression between these genes and differentially expressed ERVs from publicly available RNA-Seq data sourced.

This co-expression data for HERV-K22 (HML-5) n cerebellum tissue compared to *TARDBP* and *BCL11b* has been displayed in Table 6.8 and for the differentially expressed HERV-H family member from Frontal Cortex tissue in Table 6.9 below. As we can see from this data while there does appear to be a positive correlation between the expression of *TARDBP* & HERV-K22 (HML-5, ERVID 2152) and a negative relationship between *BCL11b* and ERVID 2152, these co-expression results are significant. The results for the comparison of ERVID 3316 (HERV-H) also shows significance for *TARDBP* in Table 6.9 with a negative correlation to *TARDBP* (indicating as the level of *TARDBP* increases HERV-H expression falls).

Table 6.8. Co-expression Analysis Results Comparing ERVID 2152, *TARDBP* and *BCL11b* in Cerebellum Tissue Between ALS and Non-ALS Controls.

The table below shows the Pearson's r correlation of ERVID 2152 and 4760 (R^2) when compared to the expression of *TARDBP* and *BCL11b* between ALS and non-ALS controls and the p-value of the comparisons.

		TDP-43		BCL11b	
ERVMap ID	HERV	R2	P-value	R2	P-value
4760	HERV-K9	0.7610	0.0002	-0.0974	0.7007
2152	HERV-K22	0.5791	0.0118	-0.2185	0.3838

Table 6.9. Co-expression Analysis Results Comparing ERVID 3316, *TARDBP* and *BCL11b* in Frontal Cortex Tissue Between ALS and Non-ALS Controls.

The table below shows the Pearson's r correlation of ERVID 3316 (R^2) when compared to the expression of *TARDBP* and *BCL11b* between ALS and non-ALS controls and the p-value of the comparisons.

Ensembl Gene ID	Gene	R^2	P-Value
ENSG00000120948	*TARDBP*	-0.6178	0.0048
ENSG00000127152	*BCL11b*	-0.1026	0.6761

Table 6.10 below shows the DESeq2 differential expression data for *TARDBP* and *BCL11b* in Cerebellum tissue samples in ALS compared to Controls. As we can see the expression of these genes is not significant by either p-value or adjusted p-value cut-offs ($p<0.01$ and adjusted $p<0.05$ respectively). Additionally, there was also no significant differential expression of these two genes within the frontal cortex tissue sample dataset comparing ALS to Controls (Table 6.11).

Table 6.10. DESeq2 Differential Expression Results for *TARDBP* and *BCL11b* in Cerebellum Tissue Between ALS and Non-ALS Controls.
The table below shows Log2 fold changes in expression of *TARDBP* and *BCL11b* between ALS and non-ALS controls.

Ensembl Gene ID	Gene	baseMean	log2FoldChange	lfcSE	pvalue	padj
ENSG00000120948	*TARDBP*	5630.628	0.515963	0.329545	0.117423	0.8449
ENSG00000127152	*BCL11b*	40.31356	-0.08601	1.494724	0.954115	1

Table 6.11. DESeq2 Differential Expression Results for *TARDBP* and *BCL11b* in Frontal Cortex Tissue Between ALS and Non-ALS Controls.
The table below shows Log2 fold changes in expression of *TARDBP* and *BCL11b* between ALS and non-ALS controls.

Ensembl Gene ID	Gene	baseMean	log2FoldChange	lfcSE	pvalue	padj
ENSG00000120948	*TARDBP*	2657.09	0.210973	0.139157	0.1295	1
ENSG00000127152	*BCL11b*	835.3407	0.189728	0.279949	0.497946	1

6.2.7 Analysis of HERV-H and HERV-K22 Open Reading Frame for Intact functional Proteins

The ERV sequences for the single significant HERV-K22 (2152) identified in the Cerebellum tissue and the single significant HERV-H (3316) identified in Frontal Cortex Tissue were obtained from UCSC genome browser (this service uses the GRch38 genome build). A consensus sequence for the internal regions of the HERV-K22 was downloaded from DFam and a consensus sequence for HERV-K22 which included 5' and 3' LTRs was constructed from the information given in Lavie *et.al.* (2004). In a similar manner the consensus sequence for HERV-H was also downloaded from dfam, as there was no alternate consensus sequence found in a paper this was the only one used for the alignment. The ERVMap sequences were aligned in MegaX using the ClustalW alignment algorithm against their consensus sequences to look for regions of high similarity. The open reading frames for ERV 2152 and ERV 3316 were then identified using UGene analysis software. The UGene software was also used to identify the open reading frames for the HERV-K22 and HERV-H consensus sequences, the open reading frames for *gag*, *prot*, *pol* and *env* identified, where present, and confirmed by NCBI nucleotide BLAST tool. The translated amino acid sequence for the consensus sequence and gene regions for ERV 2152 and 3316 were entered into ExPASy SWISS-Model protein modelling tool to obtain 3D models for alignments. These alignments are given QMEAN scores which relate to how good of a fit the predicted model is to the amino acid sequence. QMEAN scores closer to 0 are the ideal for models that are taken as good, also denoted by a "thumbs up" symbol on the tool. On the other end of the scale QMEAN scores of -4 or lower are an indication that model is an increasingly poor fit to the amino acid sequence. The *pol* region was the sole region used for ERV 2152 as this was the only partially intact open reading frame to show a recognisable conserved motif in the form of the reverse transcriptase functional site. For HERV-H, the single ORF which translated to a similar protein sequence to the consensus was the RNAseH sub-region of *pol*, though this was initially identified as part of the reverse transcriptase sub-region by SWISS-MODEL. Unfortunately, there were no ORFs large enough to be identified in the ERVMAP 4760 HERV-K9 sequence so no protein model could be predicted.

Figure 6.8 shows the 3D models for the pol consensus sequence (Figure 6.8A), the pol sequence from the ERVMap 2512 HERV-K22 sequence (Figure 6.8B) and their alignment

using the PyMOL protein software. This protein structure model for the HERV-K22 *pol* fragment has a QMEAN score of -3.85, below the -4 cut-off for a poor protein model with the only alternative given having a QMEAN score of -5.04. As we can see in the Figure 6.8C the pol fragment ORF in the ERVMap 2152 HERV-K22 sequence encodes for part way through the RT active site coding for MDD (due to the methionine in the second position of the FMDD motif), RNaseH and integrase regions. This represents a relatively intact *pol* polyprotein with few stop codon interruptions. SMART was unable to identify the open reading frame as reverse transcriptase, with none of the sub domains seemingly related to reverse transcriptase features (the closest being a GTPase binding domain). However, HMMER was able to identify the "thumb" region of reverse transcriptase within the sequence with an e-value of $2.8x10^{-39}$.

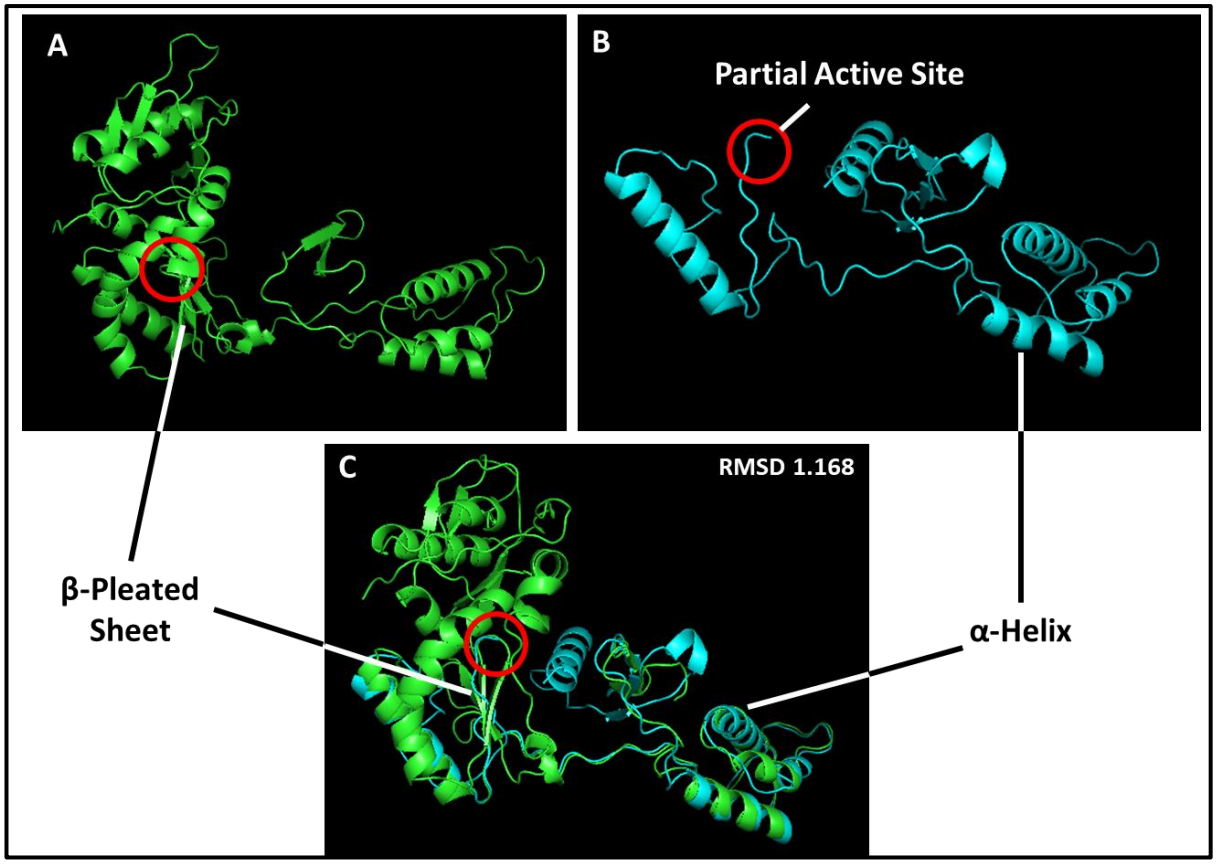


Figure 6.8 SWISS-Model 3D Protein Models for *pol* Open Reading Frames Identified in HERV-K22 Consensus Sequence and ERVMap 2152 HERV-K22 sequence.
The figure above shows the SWISS-Model 3D protein models for translated HERV-K22 consensus sequence *pol* region (A) and the ERVMap 2152 HERV-K22 *pol* fragment (B). The 3D model for the ERVMap 2152 sequence (blue) was aligned against the *pol* consensus region (green) and a Root-mean-square deviation of atomic positions (RMSD) score given for the alignment (C). The red circled region in each of the images identifies the reverse transcriptase active site in each alignment.

Figure 6.9 below shows a single ORF which was identified as forming a protein by SWISS-MODEL (ExPASy, Switzerland) and BLASTp (NCBI, USA) for the differentially expressed HERV-H found in the Frontal Cortex tissue. This ORF fragment from HERV-H has a QMEAN score of -2.50 on SWISS model, compared to the only other structure generated which showed a QMEAN of -5.04, the same as the discarded model for the HERV-K22 *pol* model. The ORF fragment from HERV-H (green in the image below) aligns closely to the consensus sequence for RNaseH with an RMSD score of 0.047. While this was initially identified in SWISS-MODEL as part of the reverse transcriptase section of *pol* it aligns closer with the segment of the consensus sequence identified as RNaseH. SMART was unable to find any similarity between the protein sequence and HERV-H, with no additional domains being

found within the sequence. HMMER was able to identify the sequence as a RNAseH sub region within *pol* though no human sequences were listed in the HMMER results.

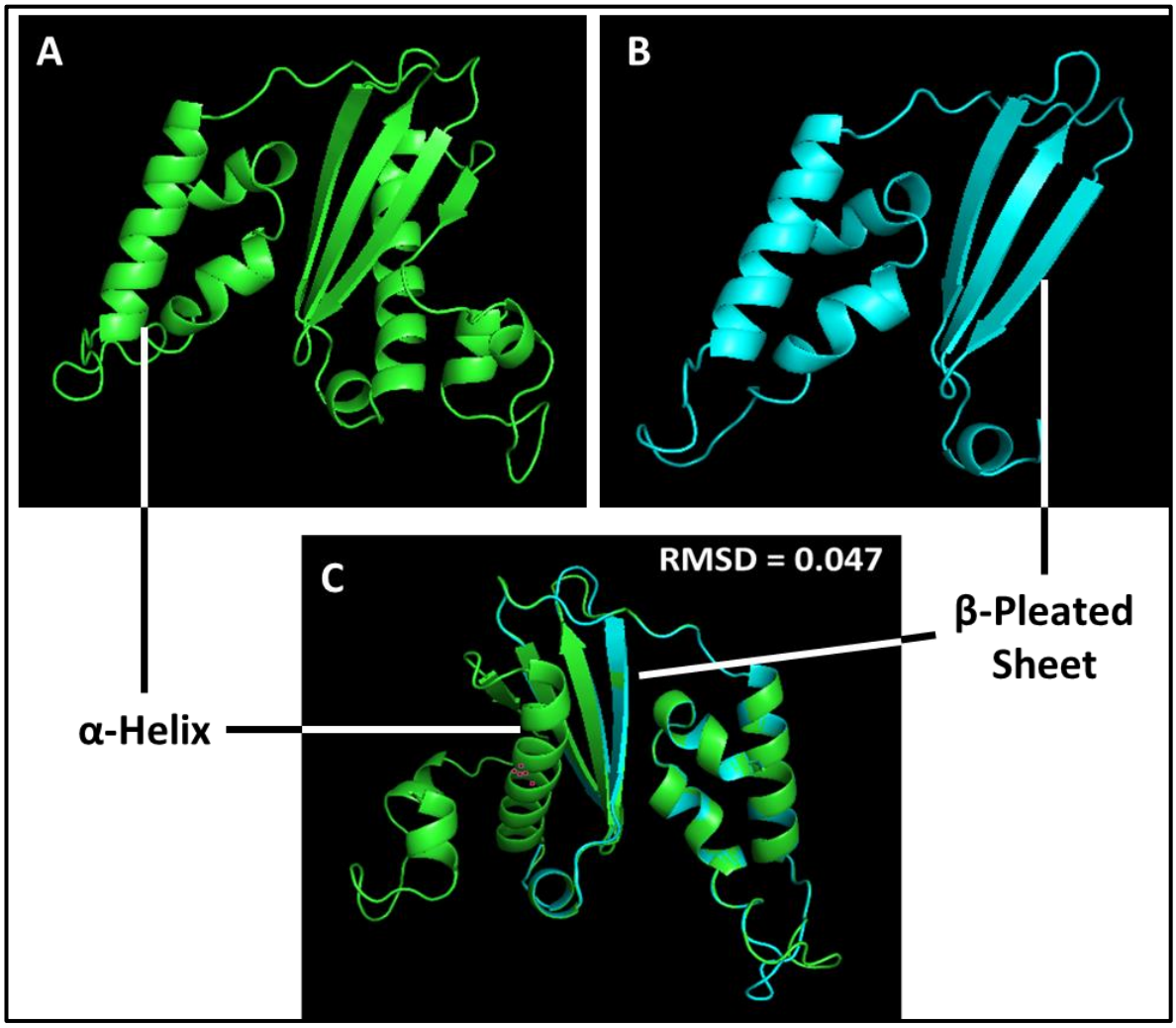


Figure 6.9 SWISS-Model 3D Protein Models for *RNaseH* Open Reading Frame Identified in HERV-H Consensus Sequence and ERVMap 3316 HERV-H sequence.
The figure above shows the SWISS-Model 3D protein models for translated HERV-H consensus sequence *RNaseH* region (A) and the ERVMap 3316 HERV-H *RNaseH* fragment (B). The 3D model for the ERVMap 3316 sequence (blue) was aligned against the *RNaseH* consensus region (green) and a Root-mean-square deviation of atomic positions (RMSD) score given for the alignment (C).

6.2.8 Investigation of HERV-K22 in the cerebellum and HERV-H in the frontal cortex, for LTR Promotor Sequences

The HERV-K22 sequence identified as being differentially expressed by DESeq2 for the Cerebellum and the HERV-H sequence differentially expressed in the Frontal Cortex were analysed for the presence of flanking LTR sequences as the majority of HERVs exist a solo LTR's.. Table 6.12 shows information about LTR regions associated with both the ERVMap 2152 HERV-K22 family member and ERVMap 3316 HERV-H family member, including whether the LTRs are present at the 5' or 3' end of the internal ERV sequences, the type of LTR present and which promotors are present in the LTR region. A paper by Manghera and Douville (2013) provided multiple promotor sequences identified in the LTR region of full length (5'LTR-gag-prot-pol-env-3'LTR) HERV-K sequences, from which transcriptional promoters were recorded for use in the analysis. Additionally, to this the canonical sequence for the TATA promotor box was recorded as this is a commonly known promotor sequence. A single Hormone specific Androgen sequence was included in the analysis as a there has been a significant difference in expression based on patient sex identified in previous PCA plot (Figure 6.5) from this chapter. These promotor sequences identified in the LTR regions of the HERV-H and HERV-K22 proviruses are from various transcription regulatory sequences, including, Ying Yang 1 (YY1), which acts as both a promotor and suppressor of transcription, Poly-adenylation signal (adds a Poly-A tail to mRNA sequences), Short Inverted Repeat (SIR) a regulatory sequence with a reverse complement downstream of its insertion, Interferon Regulatory Factor, which activates or suppresses gene transcription in the presence of interferon, T-Cell Factor-1 a T-cell–specific mediator of Wnt signaling (immune cell maintenance pathway), Upstream Transcription Factor which allows for the binding of dual α-helix protein structures, X-box binding protein 1 factor which allows for the binding of the XB1 protein and GC boxes which are found upstream of the TATA site and 110bp upstream of transcription start sites. In addition to these, also found in the LTR sequences below are CCAAT-Enhancer Binding Protein which has a role in the activation of Myeloid-derived suppressor cells (Wang *et al.*, 2019), E-twenty six (Erythroblast Transformation Specific) which has a role in cancer gene expression (Zhang *et al.*, 2020), GATA binding protein which acts a binding site for GATA-1 (Wilson, Dorfman and Orkin, 1990), Ikarose-1 found to regulate dendritic cell development

(Cytlak *et al.*, 2018), Polyomavirus Enhancer Activator 3 a epididymal transcription factor which has also seen involvement in breast cancer (HF *et al.*, 2011), Myc Associated Zinc finger protein which has been shown to have the capability to block progression of RNA polymerase II along the genome resulting in alternative RNA splicing patterns (Xiao, Li and Felsenfeld, 2021), Specificity Protein 1 which plays a role in embryonic development (Safe *et al.*, 2014), Vitamin D Receptor which in complement with Vitamin D regulates expression of over 900 genes (Kongsbak *et al.*, 2013), Nuclear Factor of Activated T cells which promotes the expression of interleukin-2 in activated T-Cells (Lee, Kim and Choi, 2018), and Integrase promotor sequences.

Table 6.12 Promotor Sequences Appearing in LTR Regions Flanking Differentially Expressed HERVs from Cerebellum and Frontal Cortex Tissue Samples.

The table below displays information about LTR sequences found in the ERVMap Region associated with differentially expressed endogenous retrovirus family members identified by DESeq2.

ERV	LTR's Present	LTR Type	5' Promoters (Amount if > 1)	3' Promoters (Amount if > 1)
4760	5', 3'	MER9B	GC Box, TATA Box (2), SIR Signal (5), CCAAT-Enhancer Binding Protein, E-twenty six (3), GATA binding protein, Ikarose-1 (2), Interferon Regulatory Factor (2), Polyomavirus Enhancer Activator 3 (3), Integrase Promotor Sequence (11)	GC Box, TATA Box (2), SIR Signal (7), E-twenty six (3), GATA binding protein, Interferon Regulatory Factor (3), Polyomavirus Enhancer Activator 3 (3), Integrase Promotor Sequence (16)
2152	5', 3'	LTR22C0	GC Box (2), TATA Box, SIR Signal (4), Activating Protein 1, E-twenty six (4), GATA binding protein, Ikarose-1, Interferon Regulatory Element, Myc Associated Zinc finger protein, Polyomavirus Enhancer Activator 3 (4), Specificity Protein 1, Vitamin D Receptor (2), Integrase Promotor Sequence (14)	TATA Box (4), Polyadenylation Signal, SIR Signal (4), Activating Protein 1, E-twenty six (3), GATA binding protein (2), Ikarose-1, Interferon Regulatory Factor (3), Myc Associated Zinc finger protein, Nuclear Factor of Activated T cells, Polyomavirus Enhancer Activator 3 (3), Vitamin D Receptor (2), Integrase Promotor Sequence (6)
3316	5', 3'	LTR7	YY1 (4), GC Box (2), TATA Box (3), Polyadenylation Signal, SIR Signal (3), Interferon Regulatory Factor (3), Integrase Promotor Sequence (15)	YY1, GC Box (2), TATA Box (2), Polyadenylation Signal, SIR Signal (3), Interferon Regulatory Factor (5), T-Cell Factor 1, Upstream Transcription Factor, XBOX binding protein, Integrase Promotor Sequence (12)

6.2.9 Analysis of Prudencio *et al.* (2017) Cerebellum and Frontal Cortex RNA-Seq Data Inclusive of C9orf72 Samples

The additional C9orf72 positive ALS samples for the Prudencio et al. (2017) dataset were obtained using the NCBI (National Centre for Biotechnology Information, USA) Sequence Read Archive (SRA) as detailed in the previous section. The additional samples from the dataset add post-mortem 8 ALS samples for both the Frontal Cortex and Cerebellum as they are both taken from the same patient from the mentioned areas of the brain. As with the previous samples in the dataset Prudencio et al. (2017) details that the samples were extracted from frozen frontal cortex and cerebellum tissue, with 20-30mg used for RNA extraction. RNA was purified from the samples using RNeasy Plus Mini Kit (QIAGEN, Germany) and quantified using Agilent Bioanalyser 2100, with samples exceeding a RIN value of 7.0 used for library preparation. Sequencing was performed on Illumina's HiSeq 2000 platform (Prudencio *et al.*, 2017). This inclusion of the additional ALS samples for the analysis resulted in n=18 ALS & n=9 non-ALS control samples from the frontal cortex and n=18 ALS & n=8 non-ALS control samples from the cerebellum.

As with the previous analysis the Cerebellum revealed a single ERV that was differentially expressed by our **adjusted p-value** cut-off as shown in Table 6.13, this ERV however is different from the one initially identified in the previous analysis indicating that the inclusion of the C9orf72 samples has an effect on the analysis. The Frontal Cortex results detailed in Table 6.14 also shows a difference in the differentially expressed ERVs identified in the tissue region. Instead of the single HERV-H (3316) that was present in the previous analysis there are now 8 ERVs differentially expressed by both **adjusted p-value** and regular **p-value** cut-offs in the frontal cortex tissue. The single ERV (5387, HERV-K3) that is differentially expressed in the Cerebellum can also be seen in the Frontal Cortex results. In both instances they are upregulated, with the fold change in the Cerebellum showing a much larger increase in expression of the transcript compared to the Frontal Cortex Region. Interestingly, when counts localised to the sex chromosomes were removed from this analysis no HERV loci were identified as being significantly differentially expressed in ALS compared to controls, apart from ERV 3316 (HERV-H) in the frontal cortex. (data not shown).

Quality control for the dataset was performed for both tissue types following the template set out in sections 6.2.1-6.2.3. Figures detailing the quality control aspects are shown in Supplementary Figures S221-S234 and while they do show some variation in comparison to the preceding section the only major difference lies within the box plot of gene counts. The lack of boxes defining the upper and lower interquartile ranges for the sample counts in Supplementary Figure S233 indicates that the median of the normalised samples appears to be outside of the range of counts shown by the dotted line. This error in the box plot may be due to the log2 transformation of the normalised counts. The PCA plots for the expression data also shows no distinctive grouping based on disease status.

Table 6.13. DESeq2 Differential Expression Results for Statistically Significant Endogenous Retrovirus in Postmortem Cerebellum Tissue Between ALS and Non-ALS Controls.

The table below shows Log2 fold changes in expression of the single significant ERV identified from an Adjusted P-value cut-off of 0.05 with p-values below 0.01. Base mean is the mean counts of the gene from all samples analysed by Deseq2, lfcSE is the standard error of the Log2Fold change, padj refers to the adjusted p-value of the differential expression result, and this pvalue is the probability of the log fold change occuring due to random chance.

ERVmap ID	Chromosome base pair Location (GRCh38 Assembly)	Chr Locus ID	UCSC Identified ERV, UCSC ERV Family (base pair length)	Base Mean	log2 Fold Change	lfcSE	p-value	p-adj
5387	chrX:53,160,068-53,162,218	Xp11.22	HERV-K3, ERVK (2151 bp)	148.6909	3.427949	4.210661	2.55E-05	0.017849

Table 6.14. DESeq2 Differential Expression Results for Statistically Significant Endogenous Retrovirus in Postmortem Frontal Cortex Tissue Between ALS and Non-ALS Controls.

The table below shows Log2 fold changes in expression of the single significant ERV identified from p-values below 0.01. Base mean is the mean counts of the gene from all samples analysed by Deseq2, lfcSE is the standard error of the Log2Fold change, padj refers to the adjusted p-value of the differential expression result, and this pvalue is the probability of the log fold change occuring due to random chance.

ERVmap ID	Chromosome base pair Location (GRCh38 Assembly)	Chr Locus ID	UCSC Identified ERV, UCSC ERV Family (bp size)	Base Mean	log2 Fold Change	lfcSE	p-value	p-adj
3316	chr10:89,284,719-89,291,763	10q23.31	HERV-H, ERV1 (7,045 bp)	174.5914	-1.4609	0.334864	1.28E-05	0.005743
ERVW-13	chr2:112,039,344-112,044,989	2q13	HERV17, ERV1 (5,646 bp)	102.1939	-1.75385	0.454775	0.000115	0.018017
5481	chrX:74,168,248-74,182,990	Xq13.2	HERV-K3, ERVK (14,743 bp)	178.0848	-0.7092	0.184485	0.000121	0.018017
1023	chr3:93,938,045-93,955,754	3q11.1	HERV-H, ERV1 (17,710 bp)	37.82272	1.54096	0.413646	0.000195	0.021799
2710	chr8:7,706,517-7,713,055	8p23.1	HERV-E, ERV1 (6,539 bp)	7.168665	3.909432	1.169434	0.000829	0.048078
4744	chr19:47,047,397-47,055,397	9q13.32	HERV-H, ERV1 (8,001 bp)	401.1711	-0.8391	0.251357	0.000843	0.048078
5387	chrX:53,160,068-53,162,218	Xp11.22	HERV-K3, ERVK (2151 bp)	24.70208	1.857808	0.557269	0.000857	0.048078
6123	chr1:183,613,209-183,622,443	1q25.3	HERV-H, ERV1 (9,235 bp)	40.32405	-1.20467	0.361484	0.00086	0.048078

6.2.10 Gene Set Enrichment and Functional Pathway Analysis for Genes within 1MB Up/Downstream of Proviral Insertion Site for Cerebellum and Frontal Cortex Tissue Data Inclusive of C9orf72 Samples.

Table 6.15 below shows data on genes 1MB (mega-base or $1x10^6$ base pairs) up/downstream of the individual ERV insertion site. Of the multiple ERVs identified the single ERV that was differentially expressed in both Cerebellum and Frontal Cortex Tissue, ERVMap ID 5387, has the most genes that have a relation to conditions that effect the brain. These are Lysine Demethylase 5C (KDM5C) which is involved in mental retardation and diseases of mental health and G Protein-Coupled Receptor 173 (GPR173), HECT, UBA And WWE Domain Containing E3 Ubiquitin Protein (HUWE1), PHD Finger Protein 8 (PHF8) and Family with Sequence Similarity 120C (FAM120C) which all have links to X-linked intellectual disability. In addition to this FAM120C also has been listed as having an involvement in Alzheimer's. Other ERVs in the table that are differentially expressed exclusively in Frontal Cortex samples, of these only 4 ERVs have genes in their 1MB up/downstream regions which are listed on GeneCards as being involved in diseases effecting the central nervous system/brain; 5481, 2710, 4744 and 6123. The genes near ERV 5481 insertion site are MicroRNA 545 (MIR545) which is linked to diseases of mental health and Neurite Extension And Migration Factor (NEXMIF) which has links to X-linked mental retardation. ERV 2710 has a single gene with links to CNS diseases, Ubiquitin Specific Peptidase 17 Like Family Member 3 (USP17L3), which is listed as having links to Partington X-Linked Mental Retardation Syndrome and Developmental Epileptic Encephalopathy 1. The genes appearing within 1MB of the ERV 4744 insertion site are BRD4 Interacting Chromatin Remodelling Complex Associated Protein (BICRA) which is linked to mental retardation syndrome and diseases of mental health and the sole gene which has a link to ALS in the list NOP53 Ribosome Biogenesis Factor (NOP53) whose function in healthy tissue is translational control and PI3K / Akt Signalling. Finally, the single gene involved in CNS related disorders in ERV 6123 is Regulator of G Protein Signalling 8 (RGS8) which is involved in Spinocerebellar Ataxia. No other gene or long non-coding RNA (lncRNA) element near the ERV insertion sites is listed as having links to neurological conditions.

Table 6.15. Annotated Genes within 1MB Up/Downstream of Proviral Insertion Sites for Significant Differentially Expressed ERVs in Frontal Cortex and Cerebellum Publicly Available RNA-Seq Data

The data given in the table below uses the UCSC genome browser to track annotated genes within 1MB up/downstream of the proviral insertion site. The table shows the annotation for the gene and its distance from either the 5' end of the provirus for upstream genes or the 3' end of the provirus for genes appearing downstream of the insertion site.

ERV	Closest Gene Upstream (kbp distance from 5' LTR)	Closest Gene Downstream (kbp distance from 3' LTR)	Genes Within 1MB Upstream	Genes Within 1MB Downstream
5387 (Frontal Cortex & Cerebellum)	KANTR (11kbp)	KDM5C (12kbp)	AL591212.1 (65kbp), TSPYL2 (69kbp), GPR173 (77kbp), FAM156A (162kbp), AC234031.1 (227kbp), FAM156B (249kbp), XAGE3 (288kbp), XAGE5 (337kbp), SPANXN5 (359kbp), SSX2B (367kbp), SSX2 (450kbp), SSX7 (502kbp), XAGE1B (639kbp), XAGE1A (654kbp), XAGE2 (780kbp), MIR8088 (821kbp), AC231759.2 (939kbp), AC245177.1 (954kbp), MAGED4 (962kbp), SNORA11D (965kbp)	MIR6895 (32kbp), MIR6894 (35kbp), IQSEC2 (67kbp), Y_RNA (159kbp), SMC1A (207kbp), MIR6857 (239kbp), RIBC1 (258kbp), HSD17B10 (268kbp), AC233728.1 (269kbp), VTRNA3-1P (296kbp), HUWE1 (370kbp), MIR98 (390kbp), MIRLET7F2 (391kbp), PHF8 (769kbp), FAM120C (902kbp)
3316 (Frontal Cortex)	Within last Intron of longer LIPA transcript, overlaps start of longer IFIT2 transcript and encompasses nearly all of element AL353751.1 (all annotated overlapping this ERV)		LIPA (Shorter Transcripts, 32kbp), FAS (268kbp), ACTA2 (294kbp), STAMBPL1 (360kbp), ANKRD22 (432kbp), LIPM (462kbp), LIPN (506kbp), LIPK (531kbp), LIPF (605kbp), LIPJ (677kbp), RNLS (701Kbp)	IFIT2 (shorter transcript, 9kbp), IFIT3 (38kbp), IFIT1B (85kbp), IFIT1 (103kbp), IFIT5 (121kbp), SLC16A12 (139kbp), PANK1 (288kbp), FLJ37201 (399kbp), KIF20B (Pseudogene, 410kbp), LINC00865 (538kbp), LINC01374 (564kbp), LINC01375 (625kbp), AL139340.1 (768kbp), RN7SKP143 (Pseudogene, 873kbp)

ERV	Closest Gene Upstream (kbp distance from 5’ LTR)	Closest Gene Downstream (kbp distance from 3’ LTR)	Genes Within 1MB Upstream	Genes Within 1MB Downstream
ERVW-13 (Frontal Cortex)	Within intron of AC093675.1		MERTK (8kbp), RN7SL297P (109kbp), ANAPC1 (154kbp), MIR4771-2 (268kbp), AC017002.3 (338kbp), AC017002.2 (402kbp), AC017002.1 (428kbp), MIR4435-2HG (514kbp), SOCAR (527kbp), MIR4435-2 (716kbp), AC068491.1 (755kbp), AC068491.2 (765kbp), AC108463.3 (825kbp), AC108463.2 (830kbp), BCL2L11 (871kbp), ACOXL (916kbp), ACOXL-AS1 (923kbp)	AC093675.2 (17kbp), FBLN7 (92kbp), AC092645.1 (133kbp), ZC3H8 (172kbp), ZC3H6 (228kbp), RGPD8 (324kbp), TTL (334kbp), POLR1B (498kbp), Y_RNA (534kbp), CHCHD5 (537kbp), AC012442.2 (543kbp), AC012442.1 (545kbp), AC079922.2 (596kbp), SLC20A1 (599kbp), NT5DC4 (677kbp), CKAP2L (692kbp), IL1A (726kbp), AC079753.2 (773kbp), AC079753.1 (780kbp), IL1B (785kbp), IL37 (866kbp), IL36G (931kbp), IL36A (962kbp), IL36B (977kbp)
5481 (Frontal Cortex)	Within Intron of FTX, Multiple shorter transcripts of the same gene appear within 100kbp downstream and 200kbp upstream of the insertion point, overlapping with other gene transcripts		AL359740.1 (55kbp), JPX (165kbp), AL353804.2 (207kbp), AL353804.1 (228kbp), AL353804.3 & AL353804.5 (229kbp), AL353804.7 & XIST (325kbp), AL353804.6 (348kbp), TSIX (350kbp), AL353804.4 (356kbp), CHIC1 (490kbp), CDX4 (718kbp), RNU6-1044P (775kbp), NAP1L2 (959kbp),	Z83843.1 (28kbp), MIR421, MIR374B & MIR374C (35kbp, all 3 appear within 200bp of each other), RN7SL648P (60kbp), AC004386.6 (91kbp), AC004386.2 (97kbp), AC004386.3 (98kbp), AC004386.4 (99kbp), MIR545 & MIR374A (104kbp), AC004386.5 (110kbp), ZCCHC13 (121kbp), RN7SL790P (207kbp), SLC16A2 (237kbp), NEXMIF (550kbp), ABCB7 (867kbp), UPRT (973kbp)

ERV	Closest Gene Upstream (kbp distance from 5’ LTR)	Closest Gene Downstream (kbp distance from 3’ LTR)	Genes Within 1MB Upstream	Genes Within 1MB Downstream
1023 (Frontal Cortex)	Within intron of PROS1		RNU6-488P (93kbp),	Y_RNA (23kbp), RNU6-511P (32kbp), ARL13B (36kbp), STX19 (71kbp), DHFR2 (113kbp), NSUN3 (117kbp), LINC00879 (991kbp)
2710 (Frontal Cortex)	FAM90A23P (124kbp)	FAM90A14P (4kbp)	PRR23D1 (162kbp), DEFB107B (196kbp), DEFB105B (214kbp), DEFB106B (220kbp), SPAG11B (237kbp), DEFB103B (278kbp), DEFB4B (286kbp), ZNF705G (319kbp), FAM66B (349kbp), USP17L4 (366kbp), USP17L1 (371kbp), DEFA5 (648kbp), DEFA3 (683kbp), DEFA1B (705kbp), DEFA1 (724kbp), DEFA4 (763kbp), AF233439.1 (773kbp), DEFA6 (776kbp), DEFB1 (824kbp), GS1-24F4.2 (834kbp), XKR5 (867kbp), AF233439.2 (914kbp), AGPAT5 (944kbp), MIR4659A & MIR4659B (958kbp), MCPH1-AS1 (995kbp)	PRR23D2 (67kbp), DEFB107A (100kbp), DEFB105A (111kbp), DEFB106A (113KBP), SPAG11A (137kbp), DEFB103A (170kbp), DEFB4A (182kbp), ZNF705B (214kbp), FAM66E (243kbp), USP17L8 (260kbp), USP17L3 (264kbp), AC105233.4 (346kbp), MIR548I3 (375kbp), FAM85B (454kbp), AC068020.1 (475kbp), PRAG1 (604kbp), AC103957.1 (700kbp), AC103957.2 (743kbp), AC114550.2 (838kbp), AC114550.3 (847kbp), AC114550.1 (848kbp), RN7SL178P (870kbp), AC087269.3 (952kbp), CLDN23 (986kbp)

ERV	Closest Gene Upstream (kbp distance from 5' LTR)	Closest Gene Downstream (kbp distance from 3' LTR)	Genes Within 1MB Upstream	Genes within 1MB Downstream
4744 (Frontal Cortex)	TMEM160 (1kbp)	ZC3H4 (8kbp)	NPAS1 (4kbp), ARHGAP35 (44kbp), AC008895.1 (99kbp), AP2S1 (199kbp), AC008622.2 (261kbp), SLC1A5 (262kbp), FKRP (289kbp), STRN4 (300kbp), Y_RNA (313kbp), AC008635.1 (317kbp), PRKD2 (331kbp), MIR320E (338kbp), RN7SL364P (359kbp), DACT3-AS1 (374kbp), DACT3 (390kbp), AC093503.2 (409kbp), GNG8 (413kbp), PTGIR (423kbp), AC093503.1 (438kbp), CALM3 (439kbp), PNMA8B (552kbp), AC011484.1 (555kbp), PNMA8A (575kbp), PNMA8C (618kbp), CCDC8 (636kbp), AC007193.3 (658kbp), PPP5C (659kbp), AC007193.1 (666kbp), HIF3A (705kbp), AC007193.2 (708kbp), RNU6-924P (741kbp), IGFL1 (815kbp), AC006262.3 (817kbp), AC006262.1 (836kbp), IGFL2-AS1 (844kbp) AC007785.1 (851kbp), AC006262.2 (867kbp), IGFL2 (887kbp), IGFL3 (925kbp), IGFL4 (971kbp)	SAE1 (76kbp), BBC3 (161kbp), MIR3190 & MIR3191 (171kbp), AC008532.1 (199kbp), CCDC9 (200kbp), INAFM1 (219kbp), C5AR1 (255kbp), C5AR2 (279kbp), DHX34 (290kbp), MEIS3 (346kbp), SLC8A2 (370kbp), AC073548.2 (414kbp), KPTN (419kbp), NAPA-AS1 (430kbp), NAPA & AC073548.1 (432kbp), ZNF541 (464kbp), RN7SL322P (524kbp), AC010519.1 (550kbp), BICRA (552kbp), BICRA-AS1 (559kbp), AC008985.1 (603kbp), EHD2 (655kbp), NOP53 (686kbp), SNORD23 (698kbp), NOP53-AS1 (700kbp), SELENOW (720kbp), TPRX1 (741kbp), CRX (765kbp), LINC01595 (808kbp), SULT2A1 (812kbp), BSPH1 (911kbp), ELSPBP1 (935kbp), CABP5 (971kbp), PLA2G4C (990kbp)

ERV	Closest Gene Upstream (kbp distance from 5' LTR)	Closest Gene Downstream (kbp Distance from 3'LTR)	Genes Within 1MB Upstream	Genes within 1Mb Downstream
6123 (Frontal Cortex)	ERV transcript spans the entirety of annotated lncRNA AL137800.1		NCF2 (22kbp), SMG7 (61kbp), SMG7-AS1 (140kbp), NMNAT2 (193kbp), AL449223.1 (238kbp), AL354953.1 (353kbp), LAMC2 (367kbp), LAMC1 (468kbp), LAMC-AS1 (469kbp), RNU6-41P (627kbp), SHCBP1L (659kbp), DHX9 (729kbp), AL355999.1 (772kbp), NPL (782kbp), LINC01688 (894kbp), RGS8 (939kbp)	ARPC5 (1kbp), RGL1 (16kbp), AL590422.1 (133kbp), COLGALT2 (309kbp), TSEN15 (429kbp), AL158011.1 (459kbp), Y_RNA (549kbp), AL445228.1 (706kbp), RN7SL654P (711kbp), AL445228.2 (763kbp), C1orf21 (766kbp), AL078645.1 (785kbp), AL713852.1 (985kbp)

As with the previous analysis of the Cerebellum and Frontal cortex this updated analysis with the C9orf72 samples included used the differential expression data from the cellular gene expression analysis to test the nearby genes for enrichment by GSEA. As with the previous dataset only the gene set within 1Mb of ERV3316 was significantly enriched, with an FDR value of 0.22 (Cut-off 0.25). The gene set in this analysis had a slightly lower enrichment score though in addition to the two genes reported in the previous analysis (SLC16A12 with a running ES of 0.1731 and ANKRD22 with a running ES of 0.4694 in this analysis) two more were found to be enriched with the addition of the C9orf72 Samples, Lipase family member J (LIPJ, Running ES: 0.3355) whose gene function is in fat metabolism and Pantothenate kinase 1 (PANK1, Running ES: 0.4719) with a reported gene function of nucleotide binding and pantothenate kinase activity. As with the initial analysis DAVID was used to see if the 4 genes enriched in this 1Mb window had any related gene function but the online tool found no significance between the enriched genes.

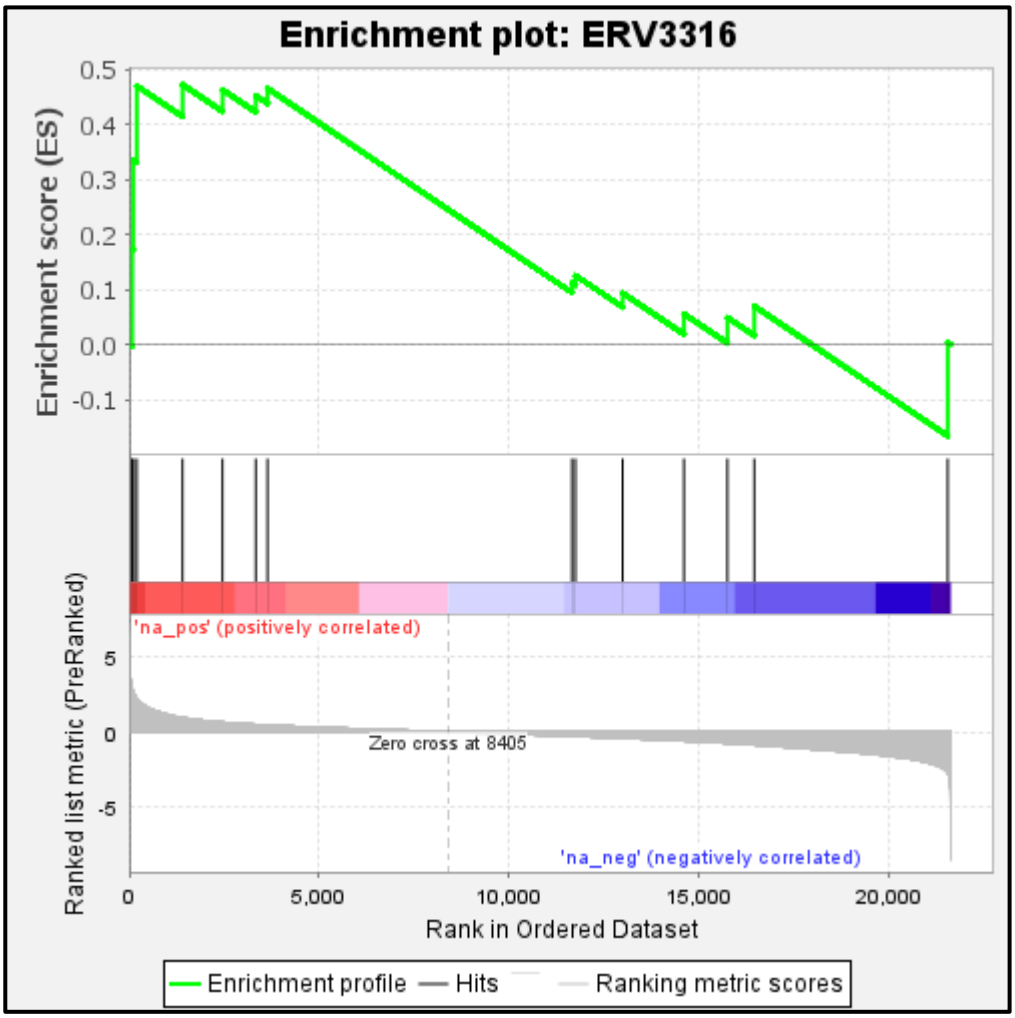


Figure 6.10. Gene Enrichment Plot for ERV3316 (HERV-H) in C9orf72 Frontal Cortex Dataset

The figure above shows the gene enrichment plot for the sole significant enriched gene set in the analysis from the publicly available RNA-Seq data from Prudencio *et.al.* (2017) inclusive of C9orf72 samples. This plot shows the positively correlated genes with the peak to the left of the graph representing the enrichment score (0.47) with the black lines below representing the list of genes within 1Mb of the proviral insertion site. The pre-ranked scores represent the log2fold change values of the individual genes in the full dataset.

As with the previous gene enrichment analysis the set of 4 cellular genes shown to be enriched with the differential expression of HERV-H (3316) were analysed to see whether their differential expression in ALS was significantly regulated. As we can see from Table

6.16 below SLC16A12, PANK1 and LIPJ were significantly upregulated while the remaining gene, ANKRD22 was not significantly expressed.

Table 6.16. DESeq2 Differential Expression Results for ERV3316 Enriched Genes in C9orf72 Inclusive Postmortem Frontal Cortex Tissue Between ALS and Non-ALS Controls.
The table below shows Log2 fold changes in expression for the enriched genes found within 1Mb of the 3316 proviral insertion site. Base mean is the mean counts of the gene from all samples analysed by Deseq2, lfcSE is the standard error of the Log2Fold change, padj refers to the adjusted p-value of the differential expression result, and pvalue is the probability of the log fold change occuring due to random chance.

Gene Symbol	Ensembl Reference	Base Mean	log2 Fold Change	lfcSE	p-value	p-adj
SLC16A12	ENSG00000152779	170.8721	2.826407	0.828082	0.000642	0.00646
LIPJ	ENSG00000204022	320.4053	2.6179	0.939129	0.00531	0.028884
ANKRD22	ENSG00000152766	958.2993	2.207988	0.971572	0.023051	0.089313
PANK1	ENSG00000152782	1131.758	0.928208	0.280591	0.000939	0.008362

To see whether this enriched gene set was significantly co-expressed with the HERV-H locus the normalised counts files for the 4 genes were compared to the normalised counts for the 3316 locus. As we can see from Table 6.17 below none of the genes were significantly co-expressed with the HERV-H locus, despite SLC16A12 shown to have a significant negative correlation with the locus in the smaller sample set.

Table 6.17. R^2 and P-Values for Correlation Analysis Between Differentially Expressed ERV 3316 and Enriched Genes Within 1Mb of the Proviral Insertion site in Frontal Cortex Tissue Samples.
The table below shows the Pearson's r correlation of analysis results for the significantly expressed HERV-H family member when compared to the expression of SLC16A12, LIPJ, ARKD22 and PANK1 between ALS and non-ALS controls and the p-value of the comparisons.

Ensemble ID	Gene Symbol	R2	P-value
ENSG00000152779	SLC16A12	-0.0502	0.8035
ENSG00000204022	LIPJ	0.1054	0.6009
ENSG00000152766	ANKRD22	0.0998	0.6204
ENSG00000152782	PANK1	-0.1336	0.5066

6.2.11 Co-expression Analysis of RNA-seq data to look for relationship between HERV Expression and Transcriptional Regulators *TARDBP* and *BCL11b* in Frontal Cortex and Cerebellum Tissue, Inclusive of C9orf72 Samples.

As with the previous analysis of the Cerebellum and Frontal Cortex significant ERVs we can see if the newly identified differentially expressed ERVs have a relationship in their expression to transcriptional regulators *TARDP* and *BCL11b*. As previously mentioned in section 6.2.8 these have been included alongside other gene targets in a previous chapter so we can see if there is a link once the C9orf72 samples have been included.

This co-expression data for the significant ERV in cerebellum tissue compared to *TARDBP* and *BCL11b* has been displayed in Table 6.18 for the differentially expressed HERV-K3 family member and the ERVs identified in the Frontal Cortex tissue in Table 6.19 below.

As we can see from this data while there does appear to be a positive relationship between the expression of *TARDBP* & ERVID 5387 and a negative relationship when analysed against *BCL11b* these co-expression results are not significant. Interestingly, while there appears to be a positive relationship by R^2 value in the cerebellum for ERVID 5387 there is a negative correlation in the frontal cortex samples. As neither of these are significant however there does not appear to be any meaningful conclusion to be drawn from the analysis.

Table 6.18. Co-expression Analysis Results Comparing ERVID 5387, *TARDBP* and *BCL11b* in Cerebellum Tissue Between ALS and Non-ALS Controls.

The table below shows the Pearson's r correlation of ERVID 5387 (R^2) when compared to the expression of *TARDBP* and *BCL11b* between ALS and non-ALS controls and the p-value of the comparisons.

Ensembl Gene ID	Gene	R^2	P-Value
ENSG00000120948	*TARDBP*	0.1559	0.4469
ENSG00000127152	*BCL11b*	-0.0201	0.9222

Table 6.19. R^2 and P-Values for Correlation Analysis Between Differentially Expressed ERVs and Retroviral Transcriptional Modifiers TDP-43 and BCL11b in Frontal Cortex Tissue Samples.

The table below shows the Pearson's r correlation of analysis results for significantly expressed ERV family members when compared to the expression of TARDBP and BCL11b between ALS and non-ALS controls and the p-value of the comparisons.

		TDP-43		BCL11b	
ERVMap ID	HERV	R2	P-value	R2	P-value
5387	HERV-K3	-0.1314	0.1391	-0.0047	0.9813
3316	HERV-H	-0.0880	0.6624	0.0196	0.9227
ERVW-13	HERV17	-0.0878	0.6634	-0.0764	0.7049
5481	HERV-K3	-0.1068	0.5959	-0.0500	0.8043
1023	HERV-H	0.2922	0.1391	0.2148	0.2820
2710	HERV-E	-0.1283	0.5236	0.1950	0.3297
4744	HERV-H	-0.2719	0.1701	0.1091	0.5879
6123	HERV-H	0.0935	0.6428	-0.0151	0.9405

The next stage was to observe whether the addition of the C9orf72 samples had an effect on the significance of *TARDBP* or *BCL11b* expression in the expanded dataset. Table 6.20 below shows the DESeq2 differential expression data for TARDBP and BCL11b in Cerebellum tissue samples. As we can see the expression of these genes is not significant by either p-value or adjusted p-value cut-offs ($p<0.01$ and adjusted $p<0.05$ respectively). Additionally, there was also no significant differential expression of these two genes within the frontal cortex tissue sample dataset (Table 6.21).

Table 6.20. DESeq2 Differential Expression Results for *TARDBP* and *BCL11b* in Cerebellum Tissue Between ALS and Non-ALS Controls.

The table below shows Log2 fold changes in expression of *TARDBP* and *BCL11b* between ALS and non-ALS controls.

Ensembl Gene ID	Gene	baseMean	log2FoldChange	lfcSE	pvalue	padj
ENSG00000120948	*TARDBP*	7482.399	0.410129	0.34595	0.23581	0.62697
ENSG00000127152	*BCL11b*	34.5542	-0.28493	1.16105	0.80614	0.85881

Table 6.21. DESeq2 Differential Expression Results for *TARDBP* and *BCL11b* in Frontal Cortex Tissue Between ALS and Non-ALS Controls.

The table below shows Log2 fold changes in expression of *TARDBP* and *BCL11b* between ALS and non-ALS controls.

Ensembl Gene ID	Gene	baseMean	log2FoldChange	lfcSE	pvalue	padj
ENSG00000120948	*TARDBP*	3378.826	0.226173	0.171616	0.187536	0.23337
ENSG00000127152	*BCL11b*	833.5142	0.028449	0.242678	0.906679	0.95102

6.2.12 Analysis of HERV-H and HERV-K3 Open Reading Frame for Intact functional Proteins.

As new ERVs have been identified in the updated analysis, inclusive of the C9orf72 samples, the newly identified ERVs were analysed to check if any open reading frames were present and able to code for functional proteins similar to the consensus sequence of the individual ERV. ERV 3316 has been analysed in section 6.2.8 this will not be analysed again here due to the sequence information obtained from the UCSC Genome Browser being the same. Obtaining ORF information from ERVs identified as being differentially expressed in this analysis follows the same process as section 6.2.8. Briefly the ORFs were identified in UGene analysis software, the amino acid sequence copied into NCBI's protein BLAST tool and if a identifiable human protein was revealed the amino acid sequence was entered into ExPASy SWISS-MODEL online tool for structure prediction. Once a structure was predicted for the individual ORF this was downloaded and compared to its consensus sequence counterpart via model alignment using PyMOL.

Of the ERVs identified as being differentially expressed in the Cerebellum and Frontal Cortex samples only ERVs 1023 (HERV-H) and 5481 (HERV-K3) had open reading frames which translated to amino acid sequences identified as human and being part of an identified HERV protein sequence. The other ERVs analysed, while having ORFs of various lengths, did not have any amino acid sequences identified as human or relating to an ERV. Analysis of ERV 1023 identified 2 ORFs which could be translated into identifiable HERV-H proteins, one for the RT region of the ERV and another for the RNaseH region, though both SMART and HMMER failed to identify the RT section of 1023 as RT and instead identified a potential transmembrane region. As the consensus sequence for HERV-H does not contain an ORF for the RT protein the RNaseH identified was aligned against the consensus sequence RNaseH for analysis (Figure 6.10). This protein model was given a QMEAN score

of -6.98 which would normally indicate a poor fit for the amino acid sequence, however the unified QMEANDisCo score (a combined neural learning multi-template version of QMEAN which combines the single score with DisCo analysis (Studer *et al.*, 2020)) was 0.52 indicating that it was a better fit that the QMEAN alone indicates (QMEANDisCo scores above 0.6 are considered a close fit). This model also covered more amino acid residues than other models which implies a closer fit. As we can see from the aligned model given in Figure 6.11 while the sequence from the UCSC Genome Browser 1023 HERV-H RNaseH region is larger than the fragment from the consensus sequence and aligns very well to the model, with a RMSD score of 0.626. HMMER was the only protein identification tool able to identify the 1024 RNaseH sequence at an E-Value of 4.0×10^{-28} as SMART only identified a low complexity region within the amino acid sequence.

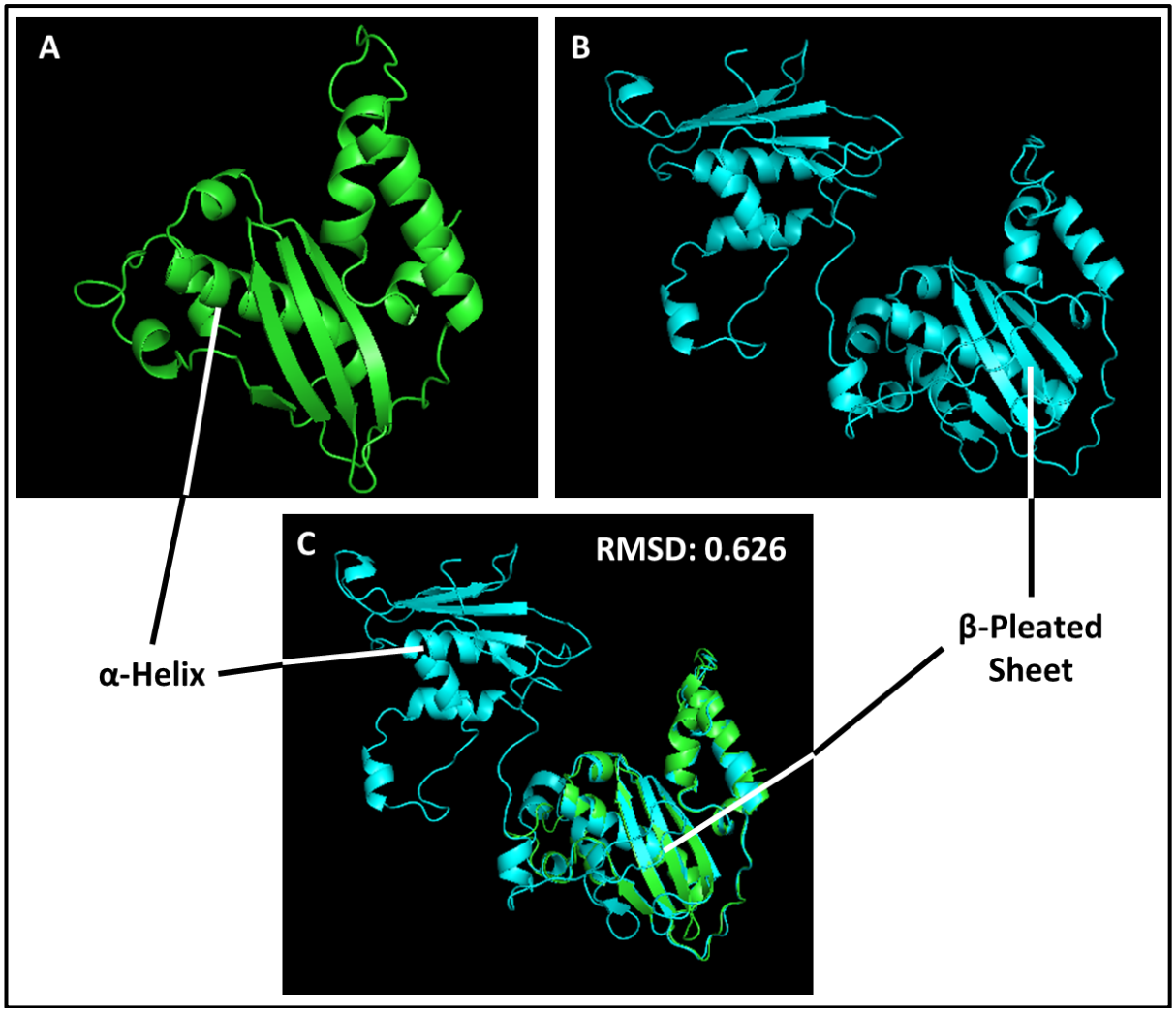


Figure 6.11 SWISS-Model 3D Protein Models for *RNaseH* Open Reading Frames Identified in HERV-H Consensus Sequence and ERVMap 1023 HERV-H sequence.
The figure above shows the SWISS-Model 3D protein models for translated HERV-H consensus sequence *RNaseH* region (A) and the ERVMap 1023 HERV-H *RNaseH* fragment (B). The 3D model for the ERVMap 1023 sequence (blue) was aligned against the *RNaseH* consensus region (green) and a Root-mean-square deviation of atomic positions (RMSD) score given for the alignment (C).

In Figures 6.12 and 6.13 below we see the alignments for the positively identified *pol* and *env* ORFs in the sequence obtained for ERV 5481 (HERV-K3) from the frontal cortex sample data. Neither alignment, despite being identified as subsets of the HERV-K3 *pol* region (Figure 6.12) and *env* region (Figure 6.13, showed similarity to their counterpart consensus sequence regions. The QMEAN data for the *env* fragment was favourable, with a value very close to 0 at -0.4 (and QMEANDisCo confirming this with a value of 0.68), indicating the model was very good for the sequence inputted. The difference in the alignment score of 5.072 then may be due to the fragment ORF for the consensus sequence not covering the same region or variations in the sequence alignment which dfam compiled the consensus

sequence on. The same can be seen for the *pol* fragment for ERV 5481 which had an initial QMEAN score of -2.44 indicating a very good match to the sequence input while the alignment score of 6.544 shows a very poor similarity between the fragments. The only region identified in the 5481 *pol* region using the SMART and HMMER tools was the N-terminal region at the beginning of the open reading frame sequence. Neither tool was able to identify the fragment as reverse transcriptase or belonging to the *pol* region. In the *env* fragment both SMART (E-value N/A) and HMMER (E-Value $1.9x10^{-20}$) were able to identify a transmembrane region for the *env* protein model. This fits with the function of an *env* protein as these are found surrounding the capsid of HERVs and other viruses and protects the viral genome from degradation.

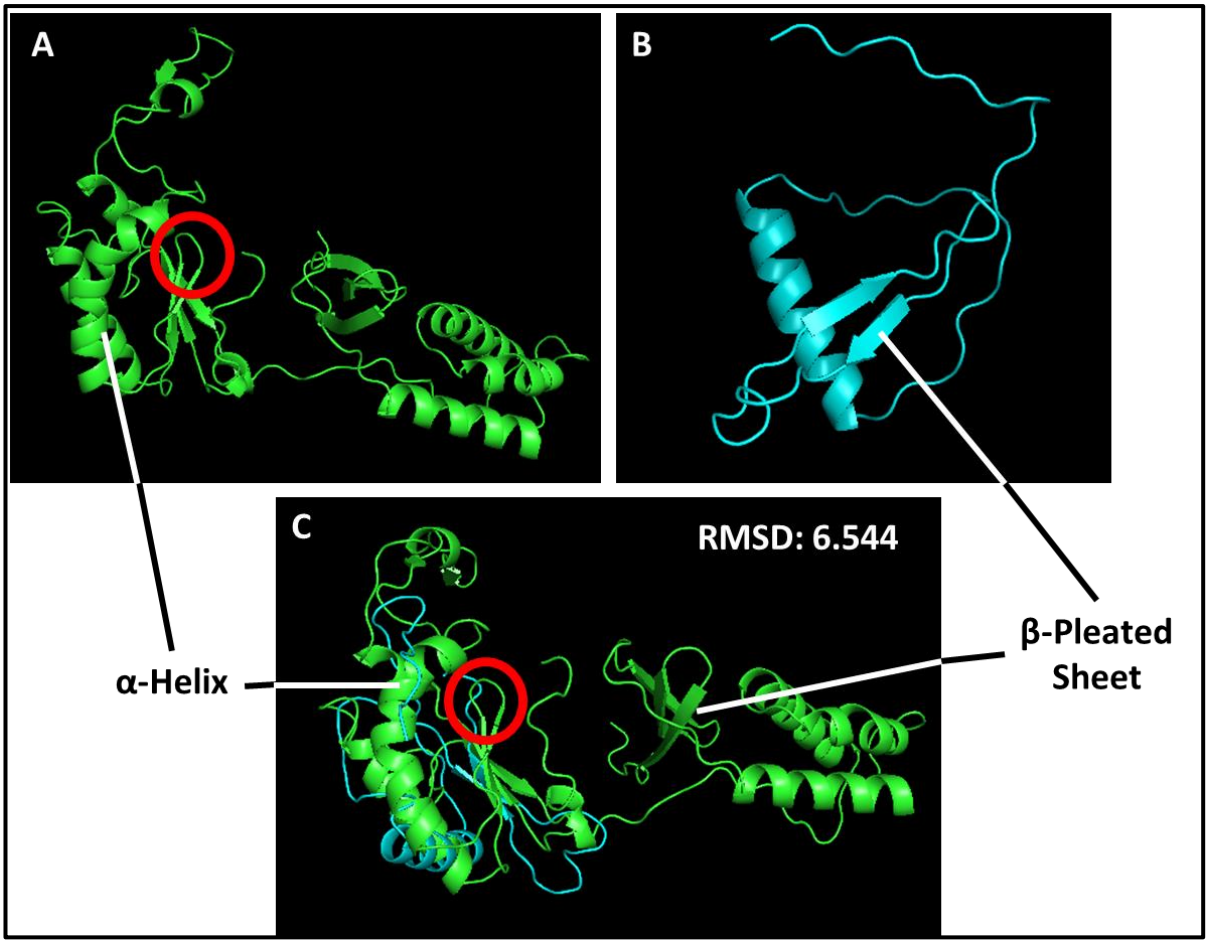


Figure 6.12. SWISS-Model 3D Protein Models for *pol* Open Reading Frames Identified in HERV-K3 Consensus Sequence and ERVMap 5481 HERV-K3 sequence.

The figure above shows the SWISS-Model 3D protein models for translated HERV-K3 consensus sequence *pol* region (A) and the ERVMap 5481 HERV-K3 *pol* fragment (B). The 3D model for the ERVMap 5481 sequence (blue) was aligned against the *pol* consensus region (green) and a Root-mean-square deviation of atomic positions (RMSD) score given for the alignment (C). The red circled region in each of the images identifies the reverse transcriptase active site in each alignment.

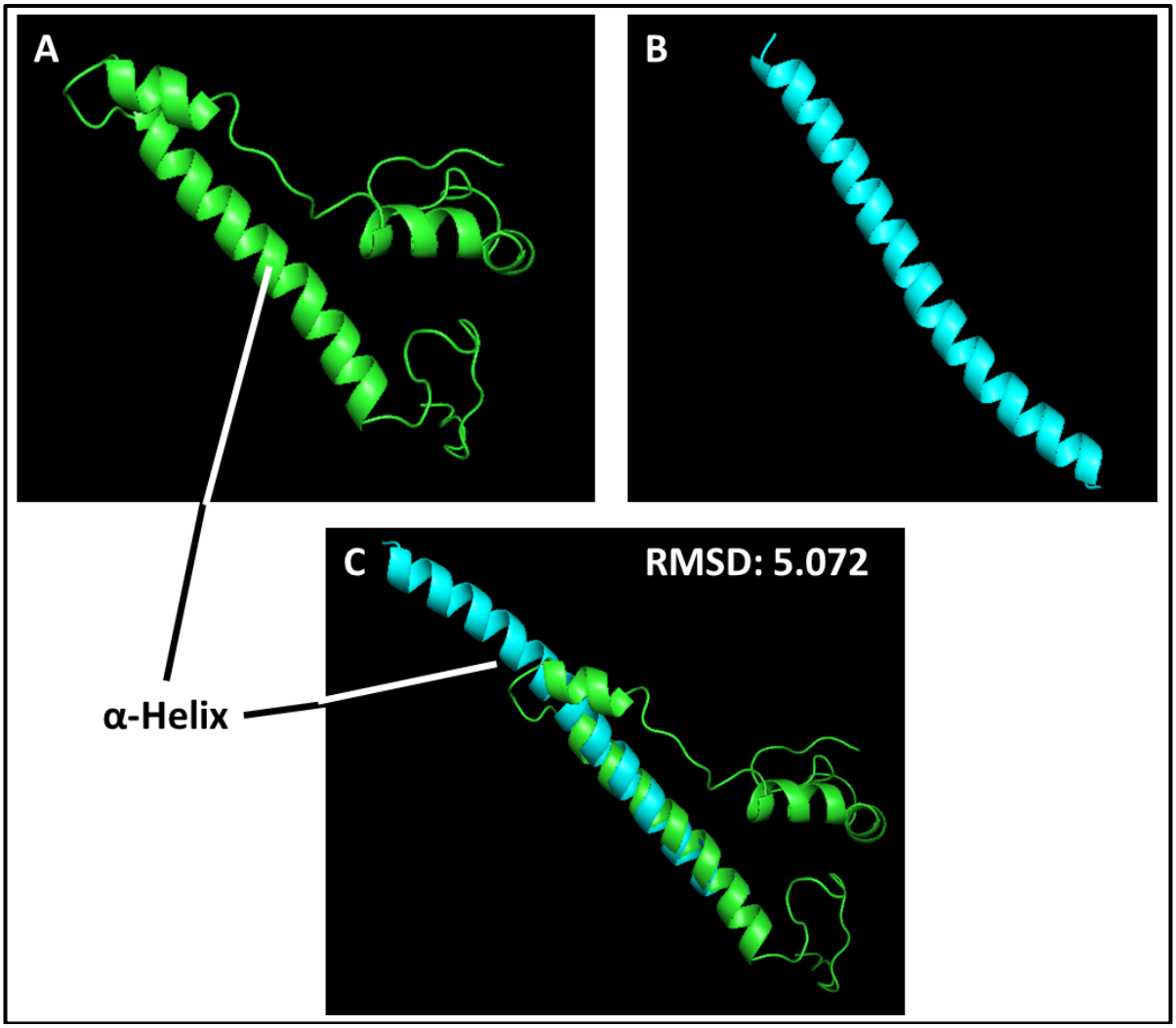


Figure 6.13. SWISS-Model 3D Protein Models for *env* Open Reading Frames Identified in HERV-K3 Consensus Sequence and ERVMap 5481 HERV-K3 sequence.

The figure above shows the SWISS-Model 3D protein models for translated HERV-K3 consensus sequence *pol* region (A) and the ERVMap 5481 HERV-K3 *env* fragment (B). The 3D model for the ERVMap 5481 sequence (blue) was aligned against the *pol* consensus region (green) and a Root-mean-square deviation of atomic positions (RMSD) score given for the alignment (C).

6.2.13 Investigation of Differentially Expressed ERVs Identified in Cerebellum and Frontal Cortex Regions for Nearby LTR Promotor Sequences

The ERVs shown to be differentially expressed in the Cerebellum and Frontal Cortex were analysed for the presence of flanking LTR sequences after assessment of protein fragment presence. These LTR sequences were obtained from the annotated region on the individual ERV as highlighted by the UCSC genome browser and their sequence interrogated using a combination of LTR promotor sequences obtained from the literature (Messeguer *et al.*, 2002; Benachenhou *et al.*, 2013; Manghera and Douville, 2013). The identified LTR sequences were then recorded in Table 6.22 Below. The only ERV which does not feature in Table 6.22 is ERV 2210 which did not have any annotated LTR regions at the 5' or 3' terminal regions of the proviral insert. Details of the transcription factors found in the table below can be read in Section 6.2.8. None of these transcription factors has been highlighted in the literature to be upregulated in ALS cases though multiple transcription factors have been found to be differentially expressed in cancers.

Table 6.22 Promotor Sequences Appearing in LTR Regions Flanking Differentially Expressed HERVs from Cerebellum and Frontal Cortex Tissue Samples.
The table below displays information about LTR sequences found in the ERVMap Region associated with differentially expressed endogenous retrovirus family members identified by DESeq2.

ERV	LTR's Present	LTR Type	5' Promoters (Amount if > 1)	3' Promoters (Amount if > 1)
5387 (Frontal Cortex & Cerebellum)	5'	LTR3A	GC Box, TATA Box (2), Polyadenylation Signal, SIR Signal (2), Interferon Regulatory Factor (2), Lymphoid Enhancer-binding Factor 1 (2), XBOX binding protein, Integrase Promotor Sequence (20)	No LTR in this region
3316 (Frontal Cortex)	5', 3'	LTR7	YY1 (4), GC Box (2), TATA Box (3), Polyadenylation Signal, SIR Signal (3), Interferon Regulatory Factor (3), Integrase Promotor Sequence (15)	YY1, GC Box (2), TATA Box (2), Polyadenylation Signal (1), SIR Signal (3), Interferon Regulatory Factor (5), T-Cell Factor 1, Upstream Transcription Factor, XBOX binding protein, Integrase Promotor Sequence (12)
ERVW-13 (Frontal Cortex)	5', 3'	LTR17	GC Box (5), TATA Box (2), Polyadenylation Signal, SIR Signal (3), E-twenty six (2), Interferon Regulatory Factor (4), Nuclear Factor of Activated T cells, Polyomavirus Enhancer Activator 3 (5), Integrase Promotor Sequence (17)	TATA Box , SIR Signal (4), E-twenty six (11), Interferon Regulatory Factor (2), Myc Associated Zinc finger protein (2), Polyomavirus Enhancer Activator 3 (11), Vitamin D Receptor (2), XBOX binding protein, Integrase Promotor Sequence (10)
5481 (Frontal Cortex)	3'	LTR3A	No LTR in this region	GC Box (3), TATA Box (4), Polyadenylation Signal, SIR Signal (3), E-twenty six (2), Polyomavirus Enhancer Activator 3 (2), Vitamin D Receptor, Integrase Promotor Sequence (20)
1023 (Frontal Cortex)	5', 3'	LTR7	GC Box (3), TATA Box (2), SIR Signal, E-twenty six, Interferon Regulatory Factor (2), Polyomavirus Enhancer Activator 3, Integrase Promotor Sequence (15)	TATA Box (2), SIR Signal (2), Interferon Regulatory Factor (2), Integrase Promotor Sequence (17)
4744 (Frontal Cortex)	5', 3'	LTR7B	YY1 (1), GC Box (2), TATA Box, SIR Signal (2), E-twenty six (7), Ikarose-1, Interferon Regulatory Factor, Nuclear Factor of Activated T cells, Polyomavirus Enhancer Activator 3 (7), Protein 53, Integrase Promotor Sequence (8)	GC Box (2), TATA Box, SIR Signal (2), E-twenty six (7), Ikarose-1, Interferon Regulatory Factor, Nuclear Factor of Activated T cells, Polyomavirus Enhancer Activator 3 (7), Protein 53, Integrase Promotor Sequence (11)

Table 6.22 (Continued) Promotor Sequences Appearing in LTR Regions Flanking Differentially Expressed HERVs from Cerebellum and Frontal Cortex Tissue Samples.

ERV	LTR's Present	LTR Type	5' Promoters (Amount if > 1)	3' Promoters (Amount if > 1)
6123 (Frontal Cortex)	5', 3'	5'LTR7 3'LTR7Y	GC Box (2), TATA Box (2), Polyadenylation Signal, SIR Signal (2), Cellular Myelocytomatosis virus protein, Interferon Regulatory Factor (2), Upstream Transcription Factor 1, Integrase Promotor Sequence (9)	GC Box (5), TATA Box (2), Polyadenylation Signal, SIR Signal (2), Interferon Regulatory Factor (3), , Integrase Promotor Sequence (9)

6.2.14 Analysis of RNA-Seq Dataset Obtained from New York Genome Center (NYGC) Covering Lateral and Medial Motor Cortex Regions.

As the postmortem brain tissue samples provided by Kings College did not reveal any significantly expressed ERV family members in the primary motor cortex further publicly available datasets for the central nervous system were sourced for analysis. The following analysis looks at RNA-Seq datasets provided by the New York Genome Center (NYGC) covering what has been annotated as the Lateral and Medial Motor Cortex regions of the brain. Cervical and Lumbar regions of the spinal cord were also analysed but failed to identify any differentially expressed ERVs. Of these Lateral and Medial Motor Cortex regions the Lateral Motor Cortex samples numbered n=39 ALS and n=6 non-ALS controls (Totalling 45 samples for the region) and the Medial Motor Cortex samples numbered n=34 ALS and n=6 non-ALS controls (Totalling 40 samples for the region). According to the metadata file provided by the NYGC, RNA was purified from the samples and sequencing libraries built using the Automated KAPA HyperPrep RNA library preparation kit (Roche, Switzerland). Following library preparation sequencing was performed on Illumina's NovaSeq platform.

Opposed to the original RNA seq analysis on a subset of Kings College samples this particular analysis revealed a single ERV in each region of the motor cortex (Tables 6.23 & 6.24 below). These ERVs were identified as HERV17 (ERVID 3443, HERV-W) in the medial motor cortex and HERV-H (ERVID 1351) in the Lateral motor cortex. These were the only ERVs in their respective datasets to be identified as being significant by our **adjusted p-value** cut-off of 0.05. The ERV identified in the Medial Motor cortex however, has an extreme log2fold change (20.88), which translates to approximately 1,000,000 times the expression in ALS than controls. This is likely due to a lack of coverage in the counts assigned to the locus across all samples in the counts matrix. On closer inspection of the matrix 18 samples have 0 counts for this ERV. All but one of the control samples have 0 counts in the ERV counts matrix which gives a potential explanation for the high log2fold change seen in the DESeq2 analysis. In an analysis of whether counts aligning to the sex chromosomes were driving any significance in differential expression; when these counts were removed from the DESeq2 analysis the significant ERVs identified in the analysis did not change (Data not shown).

Table 6.23. DESeq2 Differential Expression Results for Statistically Significant Endogenous Retrovirus in Medial Motor Cortex Tissue Between ALS and Non-ALS Controls.

The table below shows Log2 fold changes in expression of the single significant ERV identified from an Adjusted P-value cut-off of 0.05 with p-values below 0.01. Base mean is the mean counts of the gene from all samples analysed by Deseq2, lfcSE is the standard error of the Log2Fold change, padj refers to the adjusted p-value of the differential expression result, and this pvalue is the probability of the log fold change occuring due to random chance.

ERVmap ID	Chromosome base pair Location (GRCh38 Assembly)	Chr Locus ID	UCSC Identified ERV, UCSC ERV Family (base pair length)	Base Mean	log2 Fold Change	lfcSE	p-value	p-adj
3443	chr11:38,601,579-38,612,336	11p12	HERV17, ERV1 (10,758bp)	5.306136	20.88336	1.936599	4.12E-27	6.75E-24

Table 6.24. DESeq2 Differential Expression Results for Statistically Significant Endogenous Retrovirus in Lateral Motor Cortex Tissue Between ALS and Non-ALS Controls.

The table below shows Log2 fold changes in expression of the single significant ERV identified from p-values below 0.01. Base mean is the mean counts of the gene from all samples analysed by Deseq2, lfcSE is the standard error of the Log2Fold change, padj refers to the adjusted p-value of the differential expression result, and this pvalue is the probability of the log fold change occuring due to random chance.

ERVmap ID	Chromosome base pair Location (GRCh38 Assembly)	Chr Locus ID	UCSC Identified ERV, UCSC ERV Family (bp size)	Base Mean	log2 Fold Change	lfcSE	p-value	p-adj
1351	chr4:27,974,870-27,981,380	4p15.1	HERV-H, ERV1 (6,511bp)	20.28117	-2.16677	0.493813	1.14E-05	0.018717

Quality control for the dataset was performed for both tissue types following the template set out in sections 6.2.1-6.2.3. Figures detailing the quality control aspects are shown in Supplementary Figures S235-S2 and show no distinctive variation which would indicate an effect on the differential expression results. The PCA plots for the expression data also shows no distinctive grouping based on disease status. Though do show a difference based on sex of patient (Supplementary Figure S243).

Table 6.25 below shows the genes that appear within 1MB of the proviral insertion point for ERV 3443 and ERV 1351. These ERVs appear in relatively sparse regions of their chromosomes with very few annotated coding regions identified in the UCSC genome browser. Of the annotations listed in the table only a single protein coding gene has been identified, Stromal Interaction Molecule 2 (STIM2), which appears 947kbp upstream of the ERV 1351 provirus. The normal function of STIM2 involves the regulation of calcium concentrations in the cytosol and endoplasmic reticulum. It is also involved in the activation of Ca(2+) associated plasma membrane channels (GeneCards, 2021). While there is a disease associated with this gene it is related to a disorder of the immune system, specifically T Cell and Nk Cell Immunodeficiency. The rest of the genes listed in the table, aside from miRNA 4275 (MIR4275) are either small novel transcripts not associated with any known gene or long non-coding RNA molecules with no diseases associated with them. Unfortunately, none of the genes within 1Mb of these insertion sites were found to be significantly enriched by GSE analysis.

Table 6.25. Annotated Genes within 1MB Up/Downstream of Proviral Insertion Sites for Significant Differentially Expressed ERVs in Medial and Lateral Motor Cortex Publicly Available RNA-Seq Data

The data given in the table below uses the UCSC genome browser to track annotated genes within 1MB up/downstream of the proviral insertion site. The table shows the annotation for the gene and its distance from either the 5' end of the provirus for upstream genes or the 3' end of the provirus for genes appearing downstream of the insertion site.

ERV	Closest Gene Upstream (kbp distance from 5' LTR)	Closest Gene Downstream (kbp distance from 3' LTR)	Genes Within 1MB Upstream	Genes Within 1MB Downstream
3443	AC021713.1 (100kbp)	LINC02759 (9kbp)	AC103798.1 (586kbp), LINC02760 (645kbp), AC061997.1 (895kbp)	LINC01493 (38kbp), AC021723.2 (415kbp), RNU6-99P (651kbp)
1351	Overlaps end of AC007106.1		AC007106.2 (32kbp), LINC02261 (689kbp), AC024132.3 (704kbp), Y_RNA (748kbp), AC024132.1 (752kbp), AC024132.2 (831kbp), STIM2 (947kbp)	AC093791.2 (245kbp), AC093791.1 (380kbp), AC097480.1 (453kbp), AC097480.2 (599kbp), RN7SL101P (728kbp), MIR4275 (838kbp)

6.2.15 Co-expression Analysis of RNA-seq data to look for relationship between HERV Expression and Transcriptional Regulators TARDBP and BCL11b in Lateral and Medial Motor Cortex Tissue Supplied by NYGC

Mirroring the previous analyses detailed above in this chapter the next stage of interrogation for the NYGC datasets is to look to see if the transcriptional regulators TARDBP and BCL11b have an observable effect on the expression of the differentially expressed ERVs identified in the Lateral and Medial Motor Cortex regions. As with the previous analyses however there does not appear to be any significant relationship as both of these particular genes have p-values outside of the 0.05 cut-off for significance (Tables 6.26 & 6.27).

While these values are outside of the cut-off there are 2 values which are close to the 0.05 value. There are close to significant negative R^2 values for BCL11b in the medial motor cortex (Table 6.26) for ERV 3443 and TARDBP in the lateral motor cortex for ERV 1351 (Table 6.27).

Table 6.26. Co-expression Analysis Results Comparing ERVID , *TARDBP* and *BCL11b* in Medial Motor Cortex Tissue Between ALS and Non-ALS Controls.

The table below shows the Pearson's r correlation of ERVID 3443 (R^2) when compared to the expression of *TARDBP* and *BCL11b* between ALS and non-ALS controls and the p-value of the comparisons.

Ensembl Gene ID	Gene	R^2	P-Value
ENSG00000120948	*TARDBP*	0.1632	0.3144
ENSG00000127152	*BCL11b*	-0.3032	0.0572

Table 6.27. R^2 and P-Values for Correlation Analysis Between Differentially Expressed ERVs and Retroviral Transcriptional Modifiers TDP-43 and BCL11b in Lateral Motor Cortex Tissue Samples.

The table below shows the Pearson's r correlation of ERVID 1351 (R2) when compared to the expression of TARDBP and BCL11b between ALS and non-ALS controls and the p-value of the comparisons.

Ensembl Gene ID	Gene	R^2	P-Value
ENSG00000120948	*TARDBP*	-0.2840	0.0587
ENSG00000127152	*BCL11b*	-0.0789	0.6064

In order to investigate the cellular genes further to see if they are differentially expressed between ALS and controls the cellular gene expression was also analysed by DESeq2. As we

can see in Tables 6.28 and 6.29 below these genes are not shown to be significantly expressed by the Adjusted P-Value cut-off of 0.05

Table 6.28. DESeq2 Differential Expression Results for *TARDBP* and *BCL11b* in Medial Motor Cortex Tissue Between ALS and Non-ALS Controls.

The table below shows Log2 fold changes in expression of *TARDBP* and *BCL11b* between ALS and non-ALS controls.

Ensembl Gene ID	Gene	baseMean	log2FoldChange	lfcSE	pvalue	padj
ENSG00000120948	*TARDBP*	1491.002	0.2480	0.1436	0.0842	0.8845
ENSG00000127152	*BCL11b*	1069.63	-0.2562	0.1892	0.1757	0.8848

Table 6.29. DESeq2 Differential Expression Results for *TARDBP* and *BCL11b* in Lateral Motor Cortex Tissue Between ALS and Non-ALS Controls.

The table below shows Log2 fold changes in expression of *TARDBP* and *BCL11b* between ALS and non-ALS controls.

Ensembl Gene ID	Gene	baseMean	log2FoldChange	lfcSE	pvalue	padj
ENSG00000120948	*TARDBP*	1464.727	0.3069	0.1171	0.0087	0.2947
ENSG00000127152	*BCL11b*	1076.885	-0.1820	0.1711	0.2876	0.6395

6.2.16 Analysis of Open Reading Frames for Intact Protein Fragments in Differentially Expressed ERVs Identified in NYGC datasets.

As before the next stage in the analyses of the NYGC Motor Cortex regions is to see if the identified significant proviruses have any open reading frames (ORFs) which could code for functional proteins. This follows the same procedure detailed in previous sections, using the UGene analysis software to identify ORFs, confirming whether these code for known human proteins and utilising the ExPASy SWISS-MODEL online software tool to create predicted protein structures for the amino acid sequences. These structures are aligned to their matching consensus sequence proteins to see if there is an observable match.

Both of the significant ERVs showed a single ORF each which coded for a retroviral protein. ERV 3443 showed a small fragment of a *gag* subunit (Figure 6.13B), shown in the predicted model as a dimer. The protein model for this *gag* fragment was given a QMEAN score of -1.14 and a QMEANDisCo score of 0.78 indicating that the algorithm has found a very high similarity to a model in the database, in this case the murine leukaemia virus capsid C-terminal domain which matches the *gag* fragment. The murine leukaemia virus result is especially interesting in this case as endogenous retroviruses are given HML groups which stand for Human Murine Leukaemia Virus-like. As the fragment is very small (only 61 amino acids) the dimer identified is only a very small part of the larger Homo-18-mer structure seen in Figure 6.14A. Due to this disparity in size the alignment of the structures is very hard to visualise in the alignment shown in Figure 6.14C, for this reason the alignment location has been identified with a red circle. The low RMSD score for this match provides some confidence that this fragment structure does belong to the larger Homo-18-mer capsid protein structure. While SMART was unable to find any related protein domains in the amino acid sequence HMMER was more successful. HMMER was able to show a *gag* sub-domain which showed similarity to the P30 core domain though no human sequences were identified.

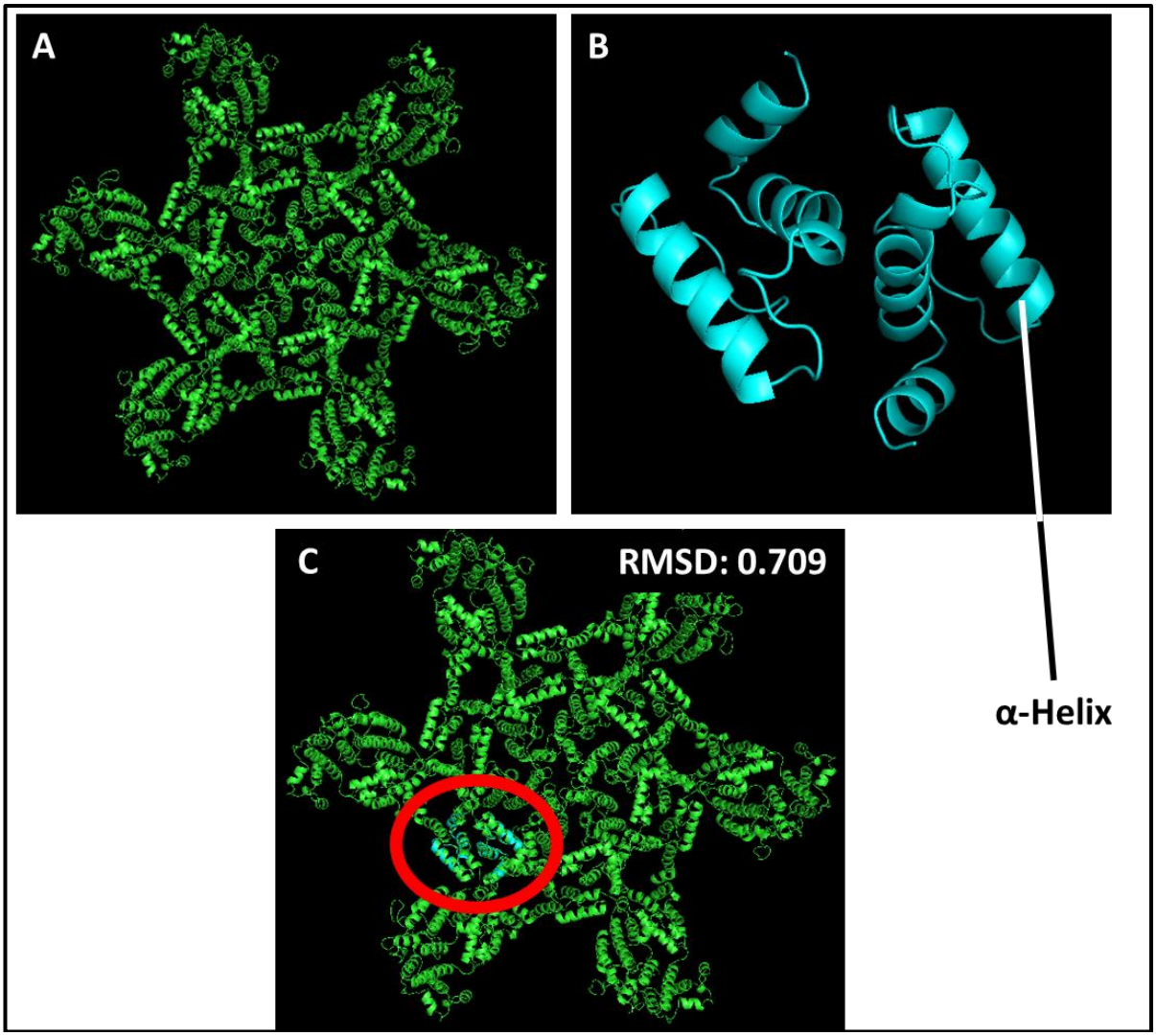


Figure 6.14 SWISS-Model 3D Protein Models for *Gag* Open Reading Frames Identified in HERV17 (HERV-W) Consensus Sequence and ERVMap 3443 HERV17 (HERV-W) sequence. The figure above shows the SWISS-Model 3D protein models for translated HERV17 (HERV-W) consensus sequence *gag* region (A) and the ERVMap 3443 HERV-W *gag* fragment (B). The 3D model for the ERVMap 3443 sequence (blue) was aligned against the *gag* consensus region (green) and a Root-mean-square deviation of atomic positions (RMSD) score given for the alignment (C). The red circled region shown in C is the location of the alignment.

In Figure 6.15B below we can see predicted structure of the *RNaseH* fragment from ERVID 1351 (HERV-H) next to it's counterpart consensus sequence in Figure 6.15A. This fragment scored much lower on the QMEAN and QMEANDisCo algorithms used on the SWISS-MODEL web tool. The poor QMEAN score of -8.56 and QMEANDisCo score of 0.34 indicate this is a poor fit for the amino acid sequence despite being the closes match they could find. The RMSD score of the alignment is interesting however, the low RMSD score would normally indicate a good alignment between the two protein structures. As we can see from Figure

6.15C despite the favourable RMSD score the two molecules are not aligned at all. This is due to the alignment in PYMOL being built on a single alpha-helix structure which appears to be the same between both molecules. In the HMMER and SMART analysis of the protein sequence only HMMER was able to identify the sequence as belonging to RNaseH with SMART identifying only low complexity regions. HMMER also annotated a transmembrane region towards the end of the sequence, past the RNaseH domain with the human sequence scoring an E-value of 2.3x10-10.

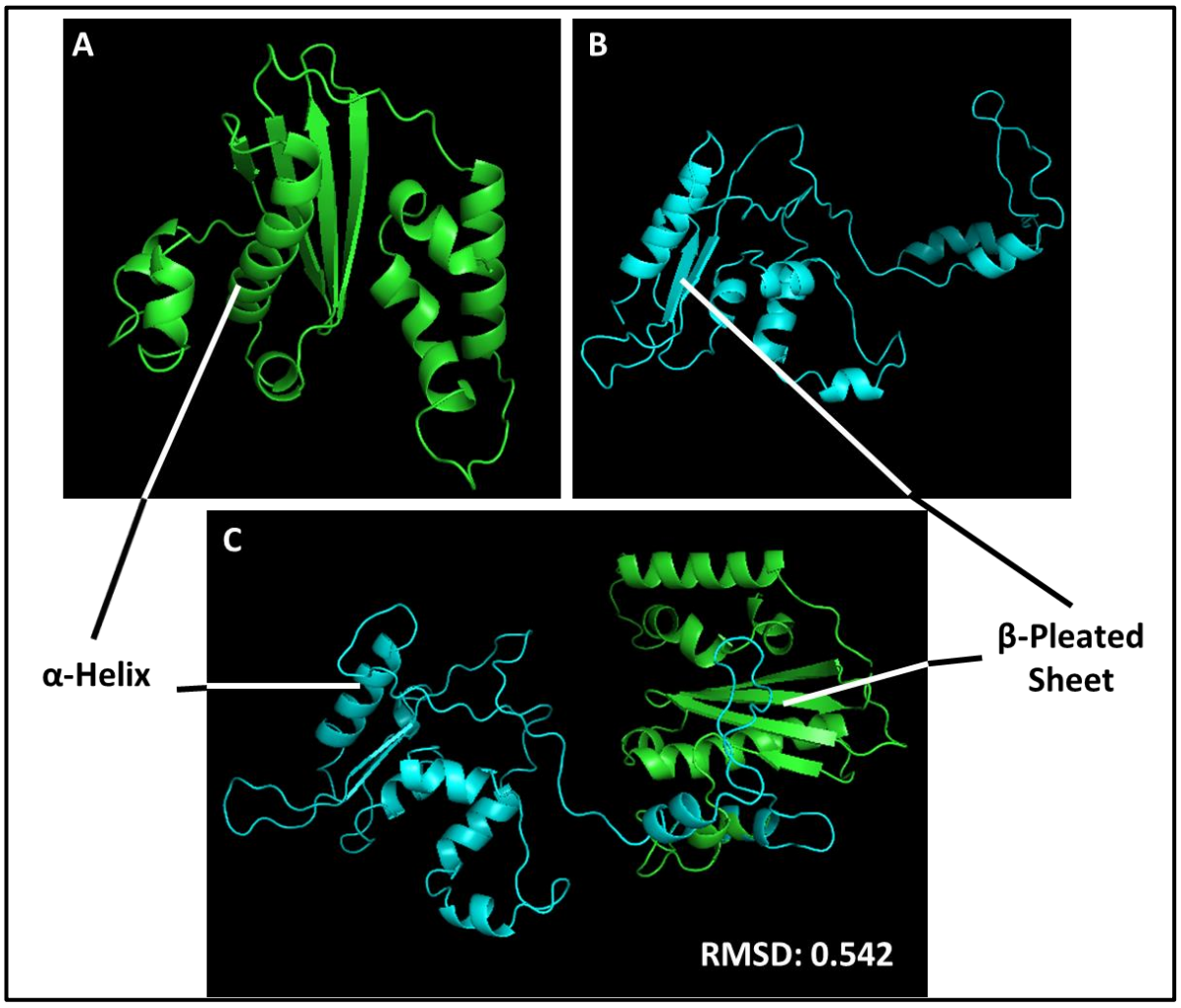


Figure 6.15 SWISS-Model 3D Protein Models for *pol* Open Reading Frames Identified in HERV-H Consensus Sequence and ERVMap 1351 HERV-H sequence.
The figure above shows the SWISS-Model 3D protein models for translated HERV-K3 consensus sequence *RNaseH* region (A) and the ERVMap 1351 HERV-H *RNaseH* fragment (B). The 3D model for the ERVMap 1351 sequence (blue) was aligned against the *pol* consensus region (green) and a Root-mean-square deviation of atomic positions (RMSD) score given for the alignment (C).

6.2.17 Investigation of Differentially Expressed ERVs Identified in Lateral and Medial Motor Cortex Regions for Nearby LTR Promotor Sequences

The proviral sequences identified in the Lateral and Medial Motor Cortex regions had their flanking LTR regions identified on UCSC and downloaded so that they could be investigated for intact promotor sequences for gene transcription. As with previous sections the LTR list was curated from the literature and variations in sequences noted for easy searching of the LTR fasta sequence (Messeguer *et al.*, 2002; Benachenhou *et al.*, 2013; Manghera and Douville, 2013). Those sequences that were found to be present in the relevant LTR regions were recorded in Table 6.30 below. Details of the individual LTR promotor sequences can be found in Section 6.2.8 though as with the C9orf72 analysis none of the promotor sequences identified have been linked to ALS gene expression.

Table 6.30 Promotor Sequences Appearing in LTR Regions Flanking Differentially Expressed HERVs from Medial and Lateral Motor Cortex Tissue Samples.
The table below displays information about LTR sequences found in the ERVMap Region associated with differentially expressed endogenous retrovirus family members identified by DESeq2.

ERV	LTR's Present	LTR Type	5' Promoters (Amount if > 1)	3' Promoters (Amount if > 1)
3443	3'	LTR17		GC Box (3), TATA Box (3), SIR Signal (4), CCAAT-Enhancer Binding Protein, E-twenty six (4), GATA binding protein, Interferon Regulatory Factor (2), Myc Associated Zinc finger protein, Polyomavirus Enhancer Activator 3 (4), Protein 53, Vitamin D Receptor (2), X-box binding protein, Integrase Promoter Sequences (23)
1351	5'. 3'	LTR7	GC Box (4), TATA Box (2), Poly-A Signal, SIR Signal (4), Interferon Regulatory Factor (4), Integrase Promoter Sequences (19)	GC Box (2), TATA Box, Poly-A Signal, SIR Signal (3), Interferon Regulatory Factor (2), Integrase Promoter Sequences (20)

6.2.18 Analysis of Differentially Expressed ERVs found in Publicly Available RNA-Seq Data by RT-qPCR of Premotor Cortex brain Tissue Samples.

RNA Sequencing (RNA-Seq) remains a powerful tool for analysing large amounts of publicly available datasets for multiple tissue regions (McDermaid *et al.*, 2019). This tool, combined with the analysis pipeline provided by ERVMap (Tokuyama *et al.*, 2018) has allowed for the analysis of transcriptome-wide Endogenous Retroviral expression in multiple publicly available datasets described in this chapter. The ERVs identified earlier in both Prudencio *et.al.* (2017) and New York Genomic Centre (NYGC) datasets are HERV-H (Lateral Motor Cortex in NYGC data and Frontal Cortex for Prudencio *et.al*) and HERV-W (Medial Motor Cortex in NYGC data and Frontal Cortex in Prudencio *et.al*) while other candidates are HERV-K22 (HML-5) and HERV-K3 (HML-6).

HERV-K22 (HML-5) appears to have no recent papers detailing function or potential pathogenic involvement in disease. The most recent publication detailing research on the proviral sequence was published by Lavie et al., (2004) which found that only 9 of the 139 proviral insertions of HERV-K22 contained close to full length sequences, the rest only being truncated versions of the provirus. HERV-K3 (HML-6) has had more recent attention, and has been shown to have a similar distribution across the genome to HERV-K22 (HML-5) though its exact function in the genome has yet to be discovered (Pisano et al., 2019). Unlike HERV-K22 however HERV-H is well characterised in the literature and is one of the more recently incorporated endogenous retroviruses in our genome, existing in multiple transcripts across the human genome (Lu *et al.*, 2014; Gemmell, Hein and Katzourakis, 2019; Carter *et al.*, 2021; Sexton, Tillett and Han, 2021). In healthy tissue HERV-H has been found to be highly expressed in embryonic stem cells (ESCs) and induced pluripotent stem cells (iPSCs) and studies involving the suppression of its transcription have found a loss of pluripotency in ESCs and a loss of efficiency in reprogramming iPSCs though the precise role of HERV-H in these cells has yet to be characterised (Lu *et al.*, 2014; Gemmell, Hein and Katzourakis, 2019; Carter *et al.*, 2021; Sexton, Tillett and Han, 2021). While this family of ERVs has not been observed in ALS to date it has been shown to be differentially expressed in both Multiple Sclerosis (MS) and Cervical Adenocarcinoma (Brudek *et al.*, 2009; Cohen, Lock and Mager, 2009; Byun *et al.*, 2021).

6.2.19 Determining HERV-K22 *pol* and HERV-H *env* Differential Expression in 54 ALS and 36 Non-ALS Control Postmortem Premotor Cortex Tissue Samples.

Primers for HERV-K22 *pol* and HERV-H *env* were designed following a similar methodology to Chapter 4.0. Briefly, this involved utilising multiple sequence alignment of ERV sequences obtained from the UCSC genome browser (University of California, Santa Cruz Genomics Institute, USA) matching loci identified in CNS and Peripheral Blood Mononuclear Cell (PBMC) RNA-Seq datasets. These were aligned in MEGAX (Molecular Evolutionary Genetics Analysis Software Version 10.1.8) alongside consensus sequences obtained from dfam.org (Institute for Systems Biology, USA) and Lavie et.al. (2004) for HERV-K22 and primer sequences were derived for genomic regions based on sequence homology between the aligned sequences.

Once primers had been selected for HERV-K22 *pol* and HERV-H *env* sequences these were processed through the quality control procedures detailed extensively in chapter 3. Briefly these showed the primers were specific to the target HERV sequences and had primer efficiency scores within the 90-110% optimal range (Supplementary Figures S248-252, Supplementary Tables 19-20).

With the HERV-K22 and HERV-H RT-qPCR assays showing good SD values for the duplicate reactions of patient samples though one sample had to be removed from the dataset due to abnormal Ct values across all assays. The differential expression of the transcripts were calculated by the 2^-ΔΔCt and Pfaffl methods. The information for the calculations performed on Microsoft Excel and verified by GraphPad Prism are summarised in Table 6.33 and Figure 6.61 below. As we can see from the results the difference in expression of HERV-K22 *pol* transcripts is not significant in premotor cortex tissue samples when measured against GAPDH, XPNPEP1 or a geometric mean of the two when measured using the 2^-ΔΔCt method and not significant when using the Pfaffl method of measuring differential gene expression

HERV-H *env* does show a significant reduction in expression in ALS when measured against the XPNPEP1 reference gene in the 2^-ΔΔCt method and is on the cusp of significance for the Geometric mean and Pfaffl measurements. This reduction in expression was also seen in the earlier Frontal Cortex RNA-Seq results (Tables 6.4 and 6.14) where the HERV-H provirus showed a log-2 fold reduction of -1.58 (-1.48 when C9orf72 samples are included) in expression compared to controls. As this was a significant result for the differential expression of HERV-H transcript the 2^-ΔΔCt data for XPNPEP1 was tested by binary logistic regression to see wether the difference in Disease and control samples were significantly related to the differential expression or whether another variable could be driving the change (Table 6.32). As we can see in the table while the significance score of the 2^-ΔΔCt score is low the most significant factor appears to be RIN (highlighted yellow).

Table 6.31. Geometric Mean and relative expression of HERV-K22 pol and HERV-H env transcripts in ALS and non-ALS cases, Normalised to GAPDH or XPNPEP1 Reference Genes.

The table displays the geometric means of the 2^-ΔΔCt differential expression values from n=54 ALS and n=36 non-ALS control samples for HERV-K22 *pol* and HERV-H env gene targets used in the RT-qPCR expression assay.

	GAPDH		XPNPEP1		GeoMean		Pfaffl	
	HERV-K22 *pol*	HERV-H *env*	HERV-K22 *pol*	HERV-H *env*	HERV-K22 *pol*	HERV-H *env*	HERV-K22 *pol*	HERV-H *env*
ALS	1.006	0.883	0.937	0.823	0.970	0.852	1.021	1.174
Control	1	1	1	1	1	1	1.075	0.899
P-Value	0.6865	0.1921	0.2850	0.0392	0.8201	0.0593	0.8586	0.0926
Significant	No	No	No	Yes	No	No	No	No

Table 6.32 Binary Logistic Regression Analysis of HERV-H *env* 2^-ΔΔCt Using a Single Reference Gene, XPNPEP1

The combined table below shows the R^2 model summaries for the binary regression followed by the p-value significance (Sig.) that each variable is related to the difference between ALS and Non-ALS control samples.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	109.117[a]	.125	.169

Variables in the Equation

		B	S.E.	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Sex(1)	-.324	.466	1	.487	.723	.290	1.803
	Age	.028	.020	1	.173	1.028	.988	1.070
	PMD	-.006	.010	1	.595	.994	.974	1.015
	RIN	-.689	.322	1	.032	.502	.267	.944
	XPNPEP1	.609	.518	1	.239	1.839	.667	5.073
	Constant	1.750	2.741	1	.523	5.757		

a. Variable(s) entered on step 1: Sex, Age, PMD, RIN, XPNPEP1.

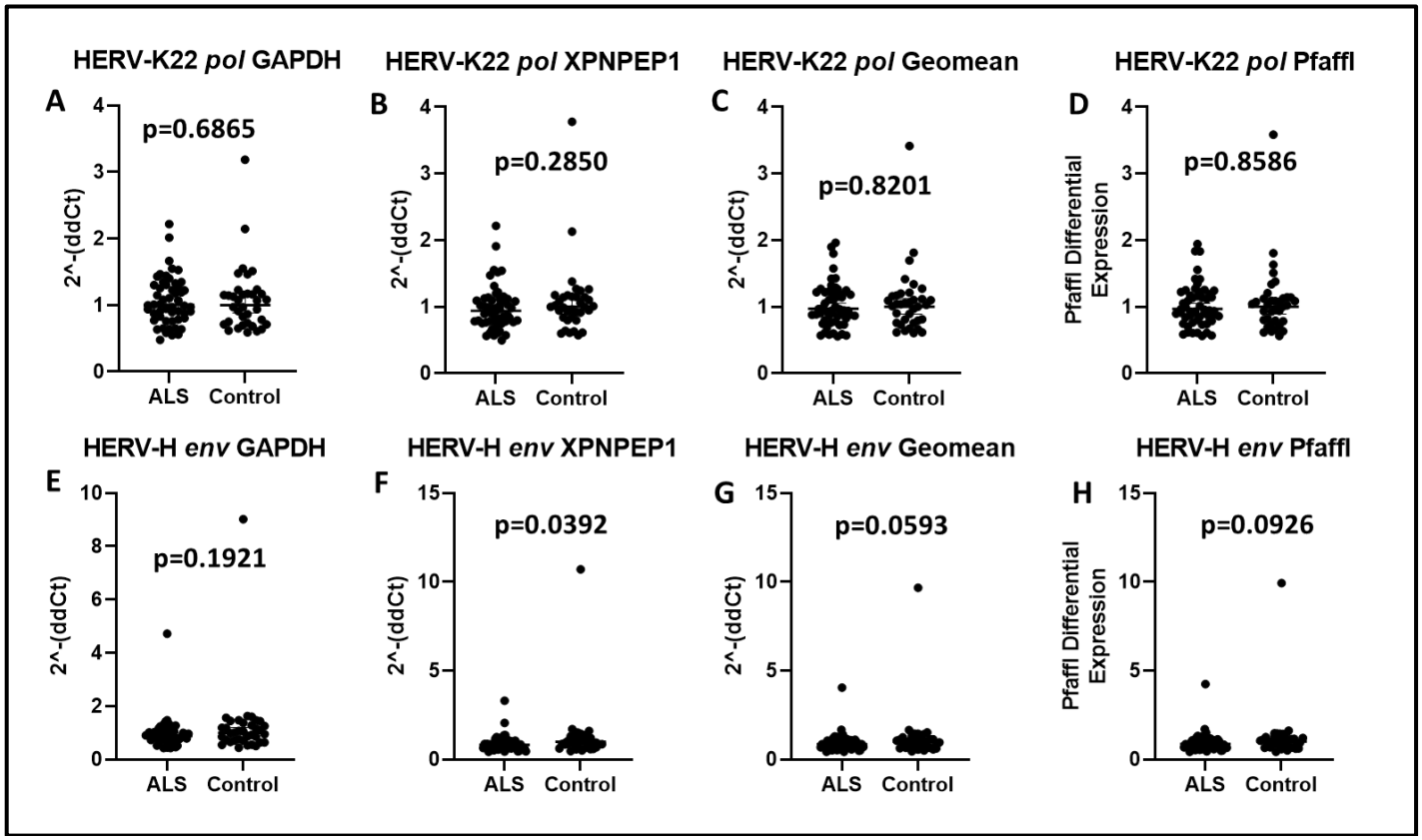


Figure 6.15. Column Graphs Showing Comparison of n=54 ALS and n=36 Non-ALS Control Samples Obtained from Postmortem Primary Motor Cortex Tissue Samples.

The Figure above shows the plotted 2^-(ΔΔCt) (A-C, E-G) and Pfaffl method (D, H) values for A-D) HERV-H *env* and E-H) HERV-K22 *pol* expression in n=54 ALS and n=36 non-ALS control postmortem primary motor cortex samples . The thick line in the middle of each group represents the geometric mean with the lines above and below representing the 95% confidence interval of the geometric mean. Also included in each graph is the Mann-Whitney U test p-value for the analysis.

6.3 Discussion

RNA-Seq is a powerful transcription quantification tool that provides an accurate way to measure differential expression of genes within a sample set. With its ability to discover novel splice junctions, long non-coding RNA and modifiers of transcription like micro-RNAs (miRNA) small interfering RNAs (siRNA) and endogenous retroviruses (ERVs) it is a useful tool to look for individual ERVs that may be differentially expressed within a sample set.

The research work conducted in the chapter above reveals 2 important results, the differentially expressed HERV-K22 and HERV-H results from the Prudencio *et al.* (2017) paper, found in the Cerebellum and Frontal Cortex samples respectively. The RNA-Seq results for the small subset of primary motor cortex samples from the Kings College sample set did not yield any significant ERV differentially expressed in the sample set. The exciting results in the Prudencio et al., (2017) dataset were the matches to consensus amino acid sequence structures where a similarity to the *pol* region was found for fragments originating from both the HERV-K22 and HERV-H proviral sequences. While there was also a novel HERV-W locus identified in the medial motor cortex sample set the log2fold change was 20.88, representing over 1 million fold difference in ALS due to controls (Table 6.23). The reason for this can be seen in the counts file which showed no counts aligning to the control tissue and a variation of 0-5 counts in the ALS samples. This has resulted in what is potentially a false positive in the DESeq2 analysis of the tissue type. However, the HERV-H locus being expressed in the lateral motor cortex, with a log2fold change of -2.17 is in line with what has been observed in the frontal cortex RNA-Seq dataset (Tables 6.4 & 6.14).

In our primary motor cortex samples, we did not see significant differential expression for those ERVs identified by the modified ERVmap expression pipeline by our adjusted p-value cut-off of $p<0.05$. When sequencing depth is above 2 million reads per sample (RNA-seq KCL data was 30 million paired-end, with a target depth of 60 million reads per sample) the largest factor affecting differential expression results was detailed by Baccarella *et al.* (2018), the amount of samples included in the study, with the greatest variation in data seen when samples were below 7 per group. The Kings College dataset gives a good example of reaching the minimum required sequencing depth, as detailed in the paper by Chen and Li (2018) for their ERVCaller protocol the minimum depth for RNA-Seq should be set to 30M as this will detect 98% of ERV sequences. As our sample set was n=11 for ALS

and n=14 for non-ALS controls this is very close to their lower limit of samples, indicating that our low sample size for comparison could be a significant factor in the inability to detect significant differential expression of HERVs in the primary motor cortex. This was confirmed by the research team at Kings College who independently confirmed that the low sample size of our analysed cohort was the reason behind the lack of detection of the HERV-K3 provirus at locus 3p21.31c as they performed the same analysis on a larger dataset.

A potential issue affecting the detection of significant differential expressed ERVs is the use of short (150bp) read sequencing. ERVs exist in multiple copies in the human genome, as their genomes are often only separated by a few base pairs this can create ambiguities in some analyses which can result in the data being biased to some regions where the final read length after quality control can map to multiple areas in the genome (Treangen and Salzberg, 2012). ERVmap accounts for this by providing an annotation file for 3220 individual ERVs, most of which will not be included in a standard .gtf file for the human genome and includes a "soft clipping" step in the alignment of data to the reference genome to preferentially filter for ERVs (Tokuyama *et al.*, 2018b). Long read sequencing methods, provided by 3rd generation sequencing platforms, could be a solution to this potential problem. When dealing with repetitive sequences that appear in ERVs, long read platforms allow for sequence reads that completely span low complexity regions, allowing mapping to specific chromosomes within the genome (Pollard *et al.*, 2018). Long read sequencing has also been proven as a useful tool in the analysis of viral genomes, allowing for better categorisation of their transcriptome, including detection of RNA editing and modification during transcription (Boldogkői *et al.*, 2019). As we can see from the ERVS identified by unadjusted p-value cut-off of $p<0.01$ HERV-H, HERV-L, MER57, HERV9, HERV4 and HERVS71 also appear to have been differentially expressed. The detection of these novel HERVs being dysregulated in ALS patients, while not significant in this sample set, shows the advantage of RNA-Seq in the interrogation of HERV expression.

The dispersion estimates plots (Figure 6.2, Supplementary Figures S210, S211, S223, S224, S237 & S238) in our data concurs with expected spread of ERV expression related to the mean of counts recorded for ERVs when compared to example plots available in the DESeq2 guidelines available on bioconductor. This gives us a good estimation for the

expression strength of ERVs in the study and while it does not show deviations of individual genes from the trend it provides a good overview of the dataset as a whole (Love, Huber and Anders, 2014). The distribution of p-values within our sample set however is not ideal, the ideal distribution for p-value results being a left leaving conservative distribution. Other histogram shapes give us alternate interpretations of the data with the common distribution of p-values for an RNA-Seq data set is binomial (ours shows a uniform distribution with very slight conservative lean). Uniform distribution in the p-value histogram is likely an artefact of even significant and non-significant samples across the dataset. These are of course interpretations based on uncorrected p-value graphs, when looking at adjusted p-values we should be seeing mostly right-leaning conservative distributions as the number of significant samples has been reduced. However, there is a clear pattern seen in Supplementary Figure S216 for the cerebellum HERV-K22 family member with reads evenly mapping across the entire genome, with an identifiable open reading frame for the RT section of *pol* though this does not code for the full protein (Figure 6.8). Functional proteins would be expected to have more reads mapping consistently to those regions without breaks in the sequence (Conesa *et al.*, 2016). However, inconsistent read mapping to the locations could either be an effect of the non-significant (as identified by the adjusted p-value of the individual ERV results) nature of these identified ERVs or the low number of reads identified as mapping to ERV loci (Table 6.1).

While there does not appear to be significant differential expression of ERVs in our Primary Motor Cortex dataset between ALS and non-ALS controls, the principal component analysis of our samples normalised counts revealed one interesting variation. Figure 6.5 displays the grouping of samples for disease status and sex (Figure 6.5A) and for RNA integrity (Figure 6.5B); these graphs show a trend in the separated groups for differing RIN values. The separate grouping of RIN values as seen by the PCA plot in Figure 6.5B is an interesting result as previous papers have shown no significant differences in HERV expression data in other brain regions in relation to differing RIN values (Sonntag *et al.*, 2016; Mayer *et al.*, 2018; Garson *et al.*, 2019). These studies primarily focused on RT-qPCR rather than RNA-Seq data so there should be some similarity between the expression data in brain tissue samples. This grouping of RIN values has been seen in RNA-Seq data focusing on gene expression in other tissues however, with placental and PBMCs showing differences in

samples on PCA analysis plots (Gallego Romero *et al.*, 2014; Reiman *et al.*, 2017). However, while the RIN values in brain tissue have been seen as having little significance in previous studies this could still have an impact on the RNA-Seq platforms ability to estimate the lower read counts. The slight separate grouping of males and females in the PCA plot is also interesting as differential HERV expression between genders has been observed in HERV-W in another neurodegenerative disease, Multiple Sclerosis (MS) (Garcia-Montojo *et al.*, 2013; García-Montojo *et al.*, 2014). While HERV-W expression was not significant by either p-value or q-value cut-off (data not shown), this does provide some evidence of differential expression of HERVs between different genders. However, there has not been evidence of differential expression in other HERV families by RT-qPCR (Chapter 4.0 included) or RNA-Seq in ALS (Tam *et al.*, 2019) apart from HERV-K (HML-2) by Li et al, 2015 by RT-qPCR. The sparse clustering of ALS samples compared to controls in the data can also be observed in Figure 6.4 heatmap. While some ALS samples do show similarity in hierarchical clustering there is little overall similarity between the expression within either the ALS or non-ALS control groups. This lack of differentiation in the clustering method may be due to the normalised counts data used to build the heatmap for significant genes identified by unadjusted p-value. This lack of effective cut-off for data means this particular graph is more effective in observing an overall trend in differences between ALS and non-ALS controls. As there is not much variation within the samples normalised counts the gene expression patterns for the genes appear to be more similar than would otherwise be noted in genes with higher mapped reads to the genomic locus.

Another interesting result in the data is the 4 samples with a significant observed “shift” in the distribution of ERV count data identified in Figure 6.6, A132_14, A218_09, A292_09 and A251_09. While these samples are a mix of both ALS and non-ALS controls and male & female samples; when we look at the PCA plots we see these samples grouped together. This seems to indicate that the gene expression in these samples is somewhat similar to each other, though they are still separated in the dendrogram of samples shown on the top of Figure 6.6. This inconsistency in ERV expression data can also be seen in Supplementary Figure S219 where the ALS samples shown their medians shifted from each other as well as the non-ALS control group. While the shifting in data for Primary Motor Cortex samples are limited to those 4 mixed samples the variability in the Cerebellum seems to be spread

across multiple patient samples. Unfortunately there was no information from the Nature paper (Prudencio *et al.*, 2017) detailing this experiment in regards to PMD, Age or Sex of the patients so this variability has not been investigated in the originating study (Prudencio *et al.*, 2017). No similar grouping of samples can be seen for the cerebellum samples in Supplementary Figure S217, indicating the difference in expression between these samples may be due to other factors. Ideally this data when normalised should show similar distributions of counts between samples, two potential explanations for the differences in these samples are either the total RNA expression in the samples is different to the others and (less likely in this instance) or the library size is different (Steinbaugh *et al.*, 2018).

The PCA groupings due to sex in Figures 6.5A and Supplementarty Figure S243 seem to indicate a difference in endogenous retrovirus expression due to sex. A previous paper by Garcia-Montojo *et al.*, (2013) also showed a significant difference in expression between female and male patients in Peripheral Blood Mononuclear Cells (PBMCs) with females showing higher expression compared to their male counterparts in the study. This difference was also stated in a paper by the same lead author the following year in relation to the expression of HERV-W showing sex related differences in expression (García-Montojo *et al.*, 2014). This sex related difference is particularly interesting as the ERV identified in the Medial Motor Cortex has been identified as HERV17, the common UCSC annotated form of HERV-W. In the 2014 paper however they also noted that the disease in question for both papers, Multiple Sclerosis, had higher occurrences in females compared to males and this presents a possible answer for overall gene expression differences observed. It can also be noted that the results have not been seen repeated in recent papers looking at HERV expression in neurological conditions which have specifically referenced García-Montojo's work (Dolei *et al.*, 2019; Morris *et al.*, 2019). One paper not referencing García-Montojo's work is from Perron *et al.* (2009), an earlier paper which reviews the HERV-W transcription in MS and references the loci for the HERV on the X chromosome as the driving force for the difference in sex in the disease. This difference in expression relating to sex in the initial analysis of Prudencio *et.al.* (2017) is likely due to reads aligning to ERVs on the sex chromosomes despite their lack of significance.

While our previous RT-qPCR expression study looked into quantifying relative expression of HERV-K *gag-pol-env* transcripts as well as HERV-W env transcripts in ALS compared to

controls (Chapter 4.0) it is unlikely that the three HERV-K sequences identified by our adjusted p-value cut-off (0.5), HERV-K3 (HML-6) (Tables 6.13 & 6.14), HERV-K22 (HML-5) and HERV-K9 (HML-3) (Table 6.3), would have been detected using the HERV-K (HML-2) specific primer sets we used in the RT-qPCR assay this was confirmed by aligning the RT-qPCR primer sequences to the identified ERV nucleotide sequences (data not shown). Our RT-qPCR primer validation process included aligning individual sequences against a multiple sequence alignment of full length HERV-K *gag-pol-env* sequences obtained from the NCBI database; primers were selected based on their ability to capture multiple HERV-K (HML2) family members so targeted conserved sequences within the *gag, pol and env* protein coding regions. Our initial investigation of HERV-K/W expression with RT-qPCR was focused on the premotor region of the motor cortex while the RNA-Seq data from matched samples was selected from the primary motor cortex. The paper Douville *et al.*, (2011) already identified that HERV-K expression can vary significantly between differing regions of the brain, though their study looked at the prefrontal, motor, occipital and sensory cortices.

Our initial plan with the primary motor cortex data was to look for co-expression of HERV-K with *TARDBP* and *BCL11b*, previously identified as transcriptional regulators of HERV-K expression in ALS (Li *et al.*, 2015; Lennon *et al.*, 2016; Douville and Nath, 2017; Ochoa Thomas *et al.*, 2020). As no HERV-K (HML-6) locus was deemed to be differentially expressed in the Primary Motor Cortex sample set we performed RNA-seq analysis on, then looking at a correlation with *TARDBP* and *BCL11b* would not be meaningful in this dataset. This correlates with what was observed in Chapter 4 where no observable correlation was seen with HERV-K transcripts and these gene sets. While some evidence was seen that showed a significant correlation with HERV-K transcripts in this chapter there was no significant difference in expression of *TARDBP* and *BCL11b* when measured by DESeq2 (Tables 6.8-6.11).

Additional analysis of the significant ERVs identified in the Cerebellum and Frontal Cortex involved looking at open reading frames with the potential to code for viral proteins and looking at the LTR regions for promotors which could potentially affect the expression of nearby genes. The alignment of the significant open reading frame for the *pol* region of the HERV-K22 family member (2152) against the consensus sequence (Figure 6.8C) shows the

fragment region starting in the methionine section of the functional motif and codes for the remaining RT section, RNaseH and Integrase region of the polyprotein while the fragment from the HERV-H family member only aligns to the RNaseH section of the polyprotein. RNaseH is an integral protein of the *pol* polyprotein region as this dissociates the bond between the cDNA produced from the viral RNA genome prior to the insertion of the cDNA into the host genome through the involvement of integrase. This methionine appears to be the first start codon in the reading frame so effects such as stop codon read-through are unlikely to produce a full pol polyprotein (Li and Zhang, 2019). The disruption in the RT active site at this point also prevents the function of the full polyprotein due to DNA and RNA molecules being unable to combine (Hu and Hughes, 2012).

The next stage of the analysis was to look for any differences in the differentially expressed genes when the C9orf72 positive ALS samples from the Prudencio *et.al.* dataset were included in the analysis. This would give us an indication on whether the mutation has an effect on the reported differentially expressed ERVs. As we can see from Tables 6.4 and 6.14 the only similar ERV to be differentially expressed occurs in the Frontal Cortex, ERV 3316 (HERV-H). Of the newly identified ERVs a single HERV-K3 family member is especially interesting (ERV 5387) as this is shown to be upregulated in both the Cerebellum and Frontal Cortex tissue regions. HERV-K is of particular interest in ALS as a study by Douville and Nath (2017) shows that the expression of HERV-K in HIV+ patients can have a protective effect by activating the inflammatory mechanisms in neuronal tissue. If we look at Table 6.13 though we can see that while some of the genes within 1MB of the proviral insertion site can be involved in neurological conditions none are related to ALS. The single ERV within the newly identified differentially expressed ERVs which has a gene within 1MB of the insertion site shown to be involved in ALS is ERV 4744 with the NOP53 Ribosome Biogenesis Factor (NOP53) gene. This gene is listed on GeneCards as having an involvement with ALS/Parkinsons but the exact mechanism of its involvement is unclear.

There are multiple transcription factors evident in both the original and C9orf72 analysis of the Prudencio *et.al.* dataset. While these have been listed in some detail in section 6.2.8. One particular promotor sequence of note is the TATA box, which is the binding site for the TATA box binding protein, acting as a transcriptional initiator for RNA polymerases I, II and III (Tjitro et al., 2019), which could potentially be involved in the transcription of the proviral

sequence. The apparent lack of other promotor/enhancer sequences in addition to those listed could be explained by only 50% of HERV LTR sequences retaining promotor sequences due to evolution based silencing methods (Buzdin et al., 2006). In order to identify whether any of these LTR sequences have other promotor sequences not identified by the sequences obtained from Manghera and Douville (2013) further testing for promotor regions for the nearby genes using methods such as 5'RACE and other promotor identifying techniques could be used.

While the HERV-K22 provirus identified in the Cerebellum by RNA seq was not able to be confirmed by RT-qPCR in our ALS and non-ALS premotor cortex sample set the HERV-H that was downregulated in the Frontal Cortex of ALS patients in the Prudencio *et al.* (2017) RNA seq dataset was also shown by RT-qPCR to be significantly downregulated in our premotor cortex samples when normalised against XPNPEP1. This represents the first instance that HERV-H has been found to be differentially expressed in ALS and represents a novel HERV family locus to be investigated further in ALS as a potential biomarker of the disease that is lacking to-date. It does not represent the first time it has been found in neurological diseases, as HERV-H env transcripts have been found upregulated on the surface of MS patient B-cells and monocytes (Rydbirk *et al.*, 2016). The exact mechanism of why HERV-H is downregulated in ALS is unclear as this is the first instance of it being discovered as being differentially regulated in ALS. HERV-H is better categorised in Cancer however, with multiple cancers showing upregulated expression compared to controls (Yi, Kim and Kim, 2006; Byun *et al.*, 2021; Manca *et al.*, 2022). In Cancer the HERV-H provirus *env* transcripts have been found to have an immune system modulating effect, specifically reducing inflammation and suppressing the immune system, some cancer cells expressing *env* proteins to escape detection by the body's immune system (Mangeney *et al.*, 2001; Gröger and Cynis, 2018). Whether this is the reason that the transcript is downregulated in ALS pathology or some other biological pathway will need to be investigated further. The data shown in Table 6.32 seems to indicate that this significant difference in expression between ALS and non-ALS controls could be driven by the change in RIN value. While this is a possibility the binary regression analysis presented in the table could also be showing that there is a significant difference in the spread of RIN values between the ALS and Control sample sets independent of the slight change in 2^-ΔΔCt expression values. As with the

potential effect of HERV-H this change should be investigated further. What is also intriguing is that the ORF we found to be intact in the 3316 sequence was the RNaseH section of *pol* and not *env*, while the ORFs in the *env* region may have been unidentifiable by BLAST maybe some fragment of the *env* protein is being expressed.

7.0 Determining the Differential Expression of Human Endogenous Retroviruses (ERV) Transcripts in ALS derived Peripheral Blood Mononuclear Cells using RNA-seq analysis on publicly available data.

7.1 Introduction

While analysis of ERV expression in post-mortem premotor and primary motor cortex tissue either by RT-qPCR and RNA-seq analysis respectively, allows for the determination of the relative amount of ERV expression of effected regions of the brain in ALS affected patients it is not without its limitations. Various factors can affect the integrity of mRNA transcripts, such as post-mortem delay and pH of the brain at time of death, with Bioanalyser derived RIN values being reported as a poor indicator of quality of total RNA transcripts recovered from post-mortem brain tissue (Durrenberger *et al.*, 2010; Sonntag *et al.*, 2016). In addition, the expression of Human Endogenous Retroviruses (HERVs) can vary due to sex, and with increasing age of the patient, meaning the snapshot of expression in post-mortem tissue is a good research tool but the use of biopsy brain tissue material is not feasible for use in a diagnostic setting (Rebollo, Romanish and Mager, 2012; Balestrieri *et al.*, 2015; Nevalainen *et al.*, 2018). Peripheral blood mononuclear cells (PBMCs), cells in circulating blood which contain a round nucleus and consist of lymphocytes (T cell, B cells, NK cells) and monocytes, and represent a better avenue for detection of novel biomarkers for disease monitoring in which sample collection is less invasive than sourcing CSF or brain biopsy material and a significant percentage of gene expression is mirrored in this sample type (Liew *et al.*, 2006; Kleiveland, 2015). Another significant advantage in measuring the expression of HERVs in ALS derived PBMCs is the lack in variation of gene expression levels in comparison to brain tissue where there is variability in gene expression depending on which region of the brain is sourced (Douville *et al.*, 2011). As HERVs are currently being researched as a novel biomarker for sporadic ALS based on researching findings by Li et al, 2015, on post-mortem brain tissue, then being able to detect differential expression of HERVs in PBMCs could be used for early disease diagnosis as well as measuring treatment response as antiretrovirals are currently being used in phase 1 clinical trials to determine efficacy and toxicity in ALS patients (Nardo *et al.*, 2011; Wildschutte *et al.*, 2016; Küry *et al.*, 2018; Dolei *et al.*, 2019; Gold *et al.*, 2019). The advantage of using PBMCs as a biomarker for HERVs over other blood components can be

seen in the paper by Bhardwaj *et al.*, 2014 where they were able to detect significant changes in HERV expression in PBMCs but not in blood plasma.

As PBMCs represent an alternative sample source to premotor cortical brain tissue or tissue taken from other regions of the brain to determine if HERVs are differentially expressed in sporadic ALS, a different set of reference genes from those we used on post-mortem brain tissue using RT-qPCR will need to be validated for accurate normalisation of gene expression data. In order to determine the appropriate stably expressed reference genes for PBMCs a new set of candidate genes have been selected including Glyceraldehyde 3-phosphate dehydrogenase **(GAPDH)**, used in the previous validation of post-mortem premotor cortex tissue (Thesis section 3.0), due to its high expression in most tissue types (Eisenberg and Levanon, 2013). Also included in the candidate set of reference genes which were identified in a recent paper, include Glucuronidase Beta, a protein coding gene **(GUSB)**, Ribosomal protein S17 **(RPS17)** and Beta-Actin, a cytoskeletal protein **(ACTB)** (Usarek *et al.*, 2017). Of these candidate reference genes identified in the study by Usarek et al 2017, RPS17 and GUSB, have been identified as the most stable across both ALS and control samples by Bestkeeper, Normfinder and geNorm reference gene selection method. Also included in the validation of reference genes is Ubiquitin C **(UBC)**, which has been shown to be stably expressed in PBMCs in multiple sclerosis, another neurodegenerative disorder (Oturai *et al.*, 2016). As with the previous validation of reference genes, geNorm (Vandesompele *et al.*, 2002), NormFinder (Andersen, Jensen and Ørntoft, 2004) , BestKeeper (Pfaffl *et al.*, 2004) and ΔCt (Silver *et al.*, 2006) mathematical models will be used to determine the appropriate reference genes for use in PBMCs. As with the work on postmortem premotor cortex tissue samples we tested in our RT-qPCR assay to measure relative HERV expression in ALS (Chapter 4.0) the amplification efficiency for HERV-K gene targets and selected reference genes will be undertaken to determine assay performance characteristics, required for determining differential expression of putative HERVs in ALS and non-ALS derived PBMCs using the ΔΔCt methodology (Livak and Schmittgen, 2001; Taylor *et al.*, 2019).

In addition, we will look to see if this expression is mirrored in publicly available RNA-Seq data using the modified RNA-Seq pipeline demonstrated in chapter 6.0 and could possibly identify novel HERVs that are differentially expressed in ALS which can be tested and

confirmed by RT-qPCR as RNA-seq analysis will allow us to undertake a broad screening of all HERV families in comparison to targeted RT-qPCR methodology.

7.2 Results

7.2.1. Differential Expression of Endogenous Retroviruses (ERVs) in PBMCs from Publicly available RNA-Seq Data

RNA-Sequencing data was obtained from the NCBI Sequence Read Archive (SRA) using the SRP123453 (2017) and SRP149638 (2018) datasets for ALS and Non-ALS controls as utilised by the paper by Zucca *et al.* (2019). As we are only interested in sporadic ALS cases, those listed as mutated ALS in the dataset were excluded from the modified ERVMap pipeline used in chapter 6.0 and in materials & methods section 2.2.15. As with the previous RNA-seq analysis as described in chapter 6.0, ERV members with low counts in multiple samples were filtered out during the analysis process. Table 7.2 displays data obtained from the DESeq2 differential expression and has been filtered to show those Endogenous Retroviruses (ERVs) whose Benjamini-Hochberg procedure adjusted p-value (q-value) was reported below the cut-off of 0.05, the adjusted p-value cut-off was selected due to its use in the literature (Picardi *et al.*, 2012; Kiskinis *et al.*, 2014; D'Erchia *et al.*, 2017; Rotem *et al.*, 2017; Kovanda *et al.*, 2018). Additionally, those unadjusted p-values identified by the $q<0.05$ cut-off were uniformly under 0.01. This q-value helps prevent the reporting of false positives by multiple comparison testing of the reported p-value of individual ERVs.

The UCSC genome browser was then used to identify the individual ERVs from their ERVmap bed file ID and the chromosome location of these HERVs recorded (Table 7.1.). Most regions identified by the ERVMap bed file which have reads mapped appear to be close to the full length ERV genome, allowing for the possibility of intact *gag-pol-env* coding regions. The dominant HERV family which is both up and downregulated is ERV1; nomenclature differences may be responsible for this prevalence as HERV-E/HERV-H/HERV17(HERV-W) are generally considered different families rather than falling under ERV1. The second most predominant ERV family is HERV-K with results showing 4 distinct family members HERV-K3 (HML-6), HERV-K9 (HML-3), HERV-K13 (HML-2) and HERV-K22 (HML-5) and HML groups identified from the paper by Subramanian *et.al.* (2011). While some ERVs showing significant differential expression in sporadic ALS are close to full length

all excluding one (1115) they do not contain intact open reading frames for viral proteins (*gag, pol* and *env*), this was determined by using the ExPASy translate tool (Supplementary Figures S118-S168). This was confirmed by running the nucleotide sequences through the NCBI nucleotide-protein BLAST search tool, these observations showed some similarities to HERV protein sequences though these results were either non-human or had very low similarity to functional proteins (Data not shown).

Table 7.1. DESeq2 Statistically Significant Differential Expression Results for Endogenous Retroviruses in Peripheral Blood Mononuclear Cells between n=15 ALS and n=7 Non-ALS Controls.

The table below shows Log2 fold changes in expression of significant ERVs identified from an Adjusted P-value cut-off of 0.05 with p-values below 0.01. Base mean is the mean counts of the gene from all samples analysed by Deseq2, lfcSE is the standard error of the Log2Fold change and padj refers to the p-value of the differential expression result adjusted by Benjamini-Hochberg correction to decrease the false discovery rate.

ERVmap ID	Chromosome base pair Location	Chr Locus ID	UCSC Identified ERV, UCSC ERV Family (Sequence Length in bp)	Base Mean	log2 Fold Change	lfcSE	p-value	p-adj
1728	chr4:184,736,174-184,745,240	4q35.1	HERV9NC, ERV1 (9,067bp)	39.6792	-5.9481	0.9671	7.73E-10	5.00E-07
570	chr2:35,434,617-35,443,665	2p22.3	HERV-H, ERV1 (9,049bp)	237.2980	1.5324	0.3099	7.60E-07	0.0002
2307	chr6:118,617,101-118,626,810	6q22.31	HERV30, ERV1 (9,710 bp)	274.2526	1.6855	0.3443	9.83E-07	0.0002
1549	chr4:106,110,153-106,115,653	4q24	HERV-K22, ERVK (5,501bp)	39.2671	2.0573	0.4260	1.37E-06	0.0002
4152	chr14:52,280,716-52,286,570	14q22.1	PRIMA4, ERV1 (5,855bp)	168.4801	2.9331	0.6368	4.10E-06	0.0005
2049	chr5:175,991,929-176,006,067	5q35.2	HERV-L, ERVL (14,139bp)	30.3235	4.2896	0.9325	4.23E-06	0.0005
2305	chr6:118,579,851-118,590,655	6q22.31	HUERS-P3, ERV1 (10,805bp)	232.0420	1.3973	0.3144	8.83E-06	0.0008
4757	chr19:51,804,687-51,811,880	19q13.41	HERV-K3, ERVK (7,194 bp)	32.6903	-4.8731	1.1045	1.02E-05	0.0008
2916	chr8:102,979,372-102,991,718	8q22.3	MER57A, ERV1 (12,347bp)	2276.0453	-5.2151	1.1910	1.19E-05	0.0009
4060	chr13:111,189,506-111,199,087	13q34	HERV9, ERV1 (9,582bp)	230.9310	-1.5318	0.3661	2.86E-05	0.0014
2306	chr6:118,574,055-118,578,291	6q22.31	HERVIP10B3, ERV1 (4,237bp)	100.6859	1.5571	0.3712	2.73E-05	0.0014
1797	chr5:32,548,454-32,560,424	5p13.3	HERV9, ERV1 (11,971bp)	77.9762	2.0002	0.4776	2.81E-05	0.0014
5359	chrX:47,866,274-47,875,865	Xp11.23	MER89, ERV1 (9,592bp)	19.6257	2.2734	0.5440	2.93E-05	0.0014
1143	chr3:143,797,678-143,805,775	3q24	HERV-H, ERV1 (8,098bp)	9.0072	4.2258	0.9997	2.37E-05	0.0014
2334	chr6:130,192,280-130,199,095	6q23.1	HERV-H, ERV1 (6,816bp)	394.2870	2.1846	0.5400	5.22E-05	0.0022
857	chr3:5,140,822-5,149,818	3p26.1	HERVP71A, ERV1 (8,997bp)	609.0807	-1.4828	0.3691	5.88E-05	0.0024
1739	chr4:189,925,280-189,933,307	4q35.2	HERV-H, ERV1 (8,028bp)	26.6740	3.0567	0.7729	7.65E-05	0.0029
4352	chr16:20,674,232-20,681,995	16p12.3	HERV15, ERV1 (7,764bp)	103.7242	-2.2839	0.5958	0.0001	0.0045
6078	chr1:158,947,554-158,959,388	1q23.1	HERV-K22, ERVK (11,835bp)	452.6244	1.7128	0.4509	0.0001	0.0047
4340	chr16:10,419,373-10,424,167	16p13.13	HERV-K9, ERVK (4,795bp)	67.0883	1.7392	0.4573	0.0001	0.0047

Table 7.1. (Continued) DESeq2 Statistically Significant Differential Expression Results for Endogenous Retroviruses in Peripheral Blood Mononuclear Cells between n=15 ALS and n=7 Non-ALS Controls.

The table below shows Log2 fold changes in expression of significant ERVs identified from an Adjusted P-value cut-off of 0.05 with p-values below 0.01. Base mean is the mean counts of the gene from all samples analysed by Deseq2, lfcSE is the standard error of the Log2Fold change and padj refers to the p-value of the differential expression result adjusted by Benjamini-Hochberg correction to decrease the false discovery rate.

ERVmap ID	Chromosome base pair Location	Chr Locus ID	UCSC Identified ERV, UCSC ERV Family (Sequence Length in bp)	Base Mean	log2 Fold Change	lfcSE	p-value	p-adj
ERVW-8	chr1:88,924,527-88,931,881	1p22.2	HERV17, ERV1 (7,355bp)	21.6786	1.8531	0.4916	0.0002	0.0050
3547	chr11:74,869,433-74,884,487	11q13.4	HUERS-P3, ERV1 (15,055bp)	183.0292	1.2273	0.3298	0.0002	0.0058
2621	chr7:124,957,310-124,967,789	7q31.33	HERV-H, ERV1 (10,480bp)	12.2666	2.2968	0.6307	0.0003	0.0076
W-92	chr4:152,841,827-152,842,330	4q31.3	HERV17, ERV1 (504bp)	5.0116	4.6403	1.2933	0.0003	0.0090
4249	chr15:34,154,888-34,161,933	15q14	HERVIP10F, ERV1 (7,046bp)	83.7398	1.2902	0.3703	0.0005	0.0128
1115	chr3:129,901,264-129,907,244	3q22.1	HERV-K9, ERVK (5,981bp)	8.1298	3.1679	0.9235	0.0006	0.0150
5446	chrX:65,437,710-65,442,396	Xq12	HERV-K9, ERVK (4,687bp)	14.9788	-1.6829	0.5109	0.0010	0.0213
5361	chrX:49,139,778-49,148,918	Xp11.23	HERV-E, ERV1 (9,141bp)	36.4247	1.9040	0.5755	0.0009	0.0213
4656	chr19:21,430,786-21,441,082	19p12	HERV3, ERV1 (10,297bp)	25.0329	2.2188	0.6754	0.0010	0.0213
3200	chr10:18,570,091-18,577,466	10p12.31	HERVIP1-F, ERV1 (7,376bp)	74.8154	2.2345	0.6801	0.0010	0.0213
2458	chr7:33,158,829-33,174,607	7p14.3	MER57A, ERV1 (15,779bp)	20.4487	3.1097	0.9430	0.0010	0.0213
5633	chrX:108,139,440-108,148,433	Xq22.3	HERV-L, ERVL (8,994bp)	62.1782	1.2280	0.3772	0.0011	0.0228
5947	chr1:84,185,856-84,193,548	1p31.1	MER101, ERV1 (7,693bp)	198.7230	1.4349	0.4443	0.0012	0.0243
765	chr2:186,431,385-186,438,154	2q32.1	HERVL18, ERVL (6,770bp)	66.3523	1.3542	0.4205	0.0013	0.0243
3409	chr11:17,370,179-17,379,293	11p15.1	HUERS-P3, ERV1 (9,115bp)	104.4376	-1.9734	0.6223	0.0015	0.0269
3776	chr12:42,440,798-42,451,511	12q12	HERV-K22, ERVK (10,714bp)	236.3573	0.6462	0.2041	0.0015	0.0269
3388	chr11:4,495,394-4,498,723	11p15.4	HERV-K13, ERVK (3,330bp)	10.9237	2.4675	0.7762	0.0015	0.0269
1379	chr4:39,539,258-39,544,115	4p14	HERV-K3, ERVK (4,858bp)	9.7132	2.2500	0.7144	0.0016	0.0278
2724	chr8:9,058,880-9,065,201	8p23.1	HERV-H, ERV1 (6,322bp)	7.8843	2.4954	0.7962	0.0017	0.0286
909	chr3:32,458,603-32,467,714	3p22.3	HERV-H, ERV1 (9,089bp)	188.5672	1.4021	0.4510	0.0019	0.0304

Table 7.1. (Continued) DESeq2 Statistically Significant Differential Expression Results for Endogenous Retroviruses in Peripheral Blood Mononuclear Cells between n=15 ALS and n=7 Non-ALS Controls.

The table below shows Log2 fold changes in expression of significant ERVs identified from an Adjusted P-value cutoff of 0.05 with p-values below 0.01. Base mean is the mean counts of the gene from all samples analysed by Deseq2, lfcSE is the standard error of the Log2Fold change and padj refers to the p-value of the differential expression result adjusted by Benjamini-Hochberg correction to decrease the false discovery rate.

ERVmap ID	Chromosome base pair Location	Chr Locus ID	UCSC Identified ERV, UCSC ERV Family (Sequence Length in bp)	Base Mean	log2 Fold Change	lfcSE	p-value	p-adj
K-46	chr20:34,126,942-34,136,578	20q11.22	HERV-K, ERVK (9,637bp)	16.6921	-2.8862	0.9335	0.0020	0.0314
3704	chr12:10,360,360-10,370,314	12p13.2	HERV15, ERV1 (9,955bp)	148.0559	1.3340	0.4337	0.0021	0.0323
4861	chr21:13,889,757-13,900,479	21q11.2	HERV-L, MaLR (10,723bp)	244.1969	3.9934	1.3044	0.0022	0.0331
4444	chr17:35,500,761-35,508,355	17q12	Harlequin, ERV1 (7,595bp)	1111.8052	1.2323	0.4051	0.0023	0.0345
3167	chr9:131,680,372-131,686,728	9q34.13	HERV-K9, ERVK (6,357bp)	499.7927	-1.1823	0.3911	0.0025	0.0359
4180	chr14:70,536,592-70,544,307	14q24.2	LTR19, ERV1 (7,716bp)	41.4278	1.1829	0.3957	0.0028	0.0393
3606	chr11:95,934,453-95,943,860	11q21	HERV-K22, ERVK (9,408bp)	19.4266	-1.8524	0.6221	0.0029	0.0399
673	chr2:127,369,874-127,377,853	2q14.3	HERV-K22, ERVK (7,980 bp)	836.5657	-1.2260	0.4179	0.0033	0.0420
4678	chr19:23,340,354-23,353,938	19p12	HERV-H, ERV1 (13,585 bp)	419.8686	0.7405	0.2533	0.0035	0.0420
3866	chr12:88,173,234-88,183,818	12q21.32	HERV-K22, ERVK (10,585bp)	76.0556	1.2634	0.4328	0.0035	0.0420
2360	chr6:143,077,707-143,086,385	6q24.2	HERV-17, ERV1 (8,679bp)	9.5500	2.1800	0.7483	0.0036	0.0420
1679	chr4:163,093,169-163,103,788	4q32.2	HERV-H, ERV1 (10,620bp)	13.4325	2.3570	0.8006	0.0032	0.0420
1643	chr4:143,177,945-143,180,459	4q31.21	HERVP71A, ERV1 (2,515bp)	11.3830	2.7891	0.9454	0.0032	0.0420
2223	chr6:80,109,511-80,113,639	6q14.1	HERV4_I, ERV1 (4,129bp)	5.2879	2.8457	0.9755	0.0035	0.0420
6195	chr1:221,947,780-221,958,635	1q41	HERV-H, ERV1 (10,856bp)	5.6702	3.1732	1.0848	0.0034	0.0420

The majority of ERVs identified as being differentially expressed in ALS vs controls PBMCs appeared in the introns of genes which have not been identified in the literature with an involvement in neurological conditions (n=36 ERVs). Of those ERVs from the HERV-K family, 6 appear in the introns of other genes, predominantly HERV-K22, aside from ERV 3606 which is 10kbps (kilo base pairs) downstream of MTMR2 (Myotubularin Related Protein 2) a tooth disease related protein. ERV 1549 (HERV-K22) appears in the intron of TBCK (TBC1 Domain Containing Kinase) which is associated with psychomotor retardation, cerebellar atrophy, developmental delay, and seizures. The only HERV-K family member (HERV-K9, ERV 1115) to have a single open reading frame for its protease gene with all other genes encoding for additional viral proteins are incomplete due to the presence of stop codons. The remaining ERVs that are significantly differentially expressed in ALS are either up or downstream of genes which are unrelated to neurological conditions as reported to-date in the literature. While none of these ERVs have been associated with ALS pathology, two ERVs have been shown to be within an intron of or close to genes related to other neurological conditions. For example, ERV 5359 (MER89, ERV1) appears within an intron of ZNF81 (Zinc Finger Protein 81, involved in Non-Syndromic X-Linked Intellectual Disability) and ERV 5361 (HERV-E) which appears 13kbp upstream of MAGIX (MAGI Family Member, X-Linked, involved in Martin-Probst Type X-Linked Mental Retardation).

Quality control for this dataset was performed as with Chapter 6 Section 6.2.1-6.2.3, showing values within acceptable ranges for the control figures (Supplementary Figures S253-S272). The exception to this is with the box plots (Supplementary Figures S271-S272) where overall the density distributions of normalised counts are not identical but still not very different, especially when compared to the ERV counts. However, 4 samples (SRR7251667-SRR7251670) are shown to be significantly different compared to the rest of the data, all these samples are controls and can be seen to form a wide separated group in the PCA plot in Supplementary Figure S260.

.

7.2.2 Analysis of RNA-seq data to measure TDP-43 and BCL11b Gene Expression in ALS derived PBMCs compared with controls and if there is any correlation with HERVs that are differentially expressed in ALS.

As TDP-43 expression has been identified as being correlated with ERV expression in previous studies due to its role as a transcriptional regulator and its interaction with the viral replication cycle as a DNA binding agent prior to the formation of the viral capsid we performed a co-expression analysis to see if the expression of TDP-43 correlates with ERV expression within our RNA-Seq data set. Additionally, as BCL11b has been shown to have a role in the suppression of retroviral transcription in the CNS during HIV infection we have also included it in our analysis to see if it regulates ERV expression in ALS. As PBMC's have been shown to be an effective mirror of gene expression in Multiple Sclerosis and have been identified as an important biomarker in ALS this analysis was performed on the publicly available RNA-seq data obtained from ALS and non-ALS PBMC dataset (Achiron and Gurevich, 2006; Nardo *et al.*, 2011). The dataset was calculated within Rstudio using jaffelab to regress out the variables in the dataset and plot a correlation matrix for p-values and R^2 linear regression statistics.

This correlation matrix has been plotted as a correlogram in Figure 7.1 below, the data has been hierarchically clustered as with the heatmap shown in Supplementary Figure S259 to cluster genes with similar R^2 values together for easy potential visualisation of patterns. If TDP-43 had a direct positive correlation with ERV expression, increasing ERV expression with increasing expression of TDP-43 between ALS and non-ALS controls, there would be a consistent blue colour as this represents a positive correlation. With BCL11b there is a weak correlation (negative correlation) denoted by red colours across the row, however it has several statistically insignificant (shown in white) correlations with ERV members. This means that a solid inference of a suppressive relationship between BCL11b and endogenous retroviruses cannot be confidently confirmed.

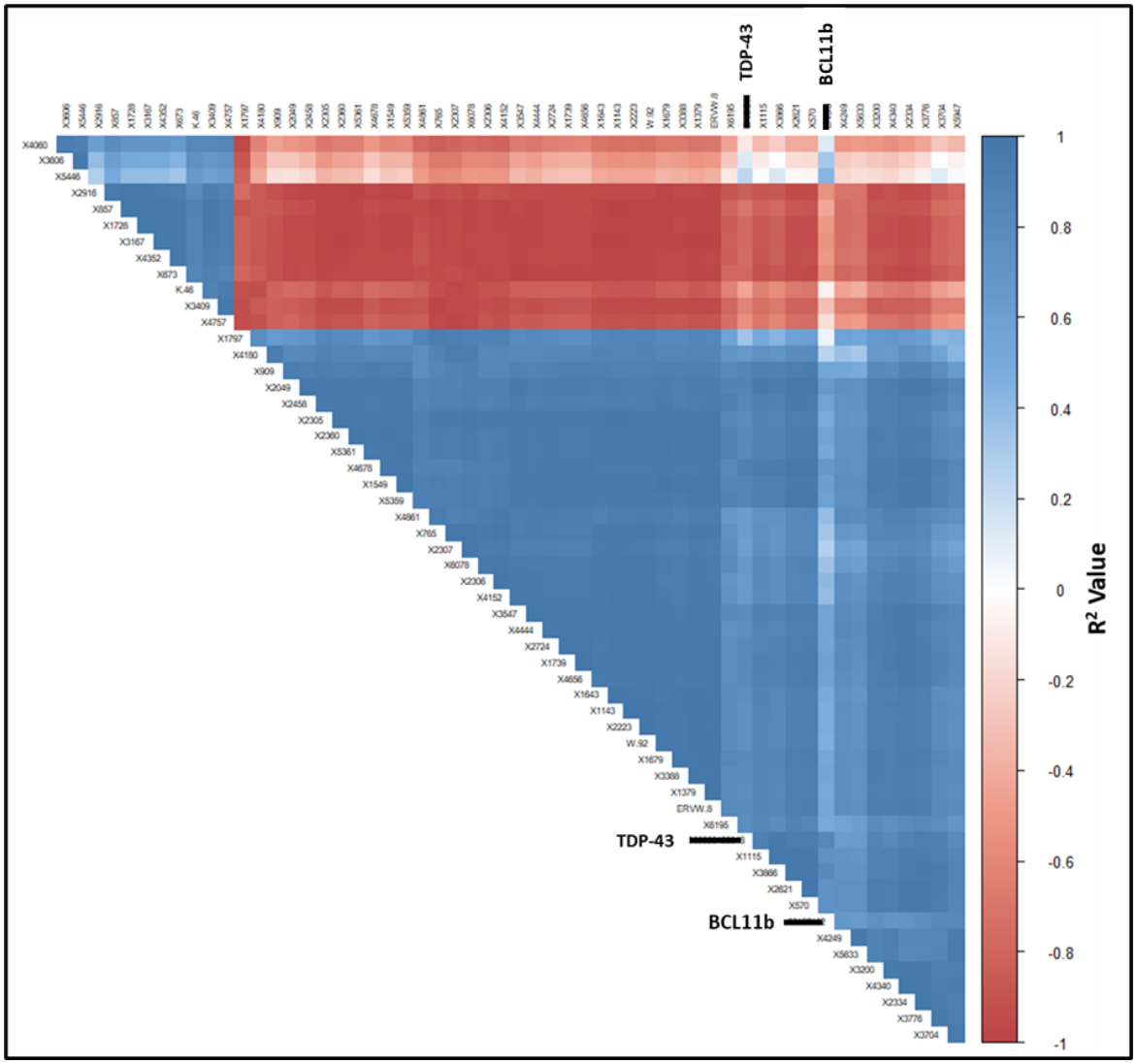


Figure 7.1. Correlogram of R2 Values for Correlations Between Statistically Significant ERVs and Transcriptional Regulators TDP-43 and BCL11b.

The figure above plots R2 values for co-expression correlations between statistically significant ERVs and transcriptional regulatory proteins TDP-43 and BCL11b. The strength of Blue or Red in the correlogram indicates the closer to 1 or -1 respectively the R2 value lies. Blank squares indicate those co-expression correlations which fall outside of a $p<0.01$ p-value cut-off for the RStudio program.

While Figure 7.1 provides an efficient method of looking at the correlation of all significant HERVs with TDP-43 and BCL11b it does not provide specific information on HERV families of interest ie: as we had looked previously at HERV-K and HERV-W correlation with TDP-43 and BCL11b in chapter 4.0 by RT-qPCR. As can be seen in Table 7.2, significantly expressed HERVs for both HERV-K and HERV-W families have been extracted and correlated with TDP-43 and BCL11b and the R^2 and p-values generated by the correlation analysis are shown.

Table 7.2. R^2 and P-Values for Correlation Analysis Between Statistically Significant HERV-K Family Members and Retroviral Transcriptional Modifiers TDP-43 and BCL11b.
The table below shows the correlation analysis results for significantly expressed HERV-W and HERV-K family members isolated from the Correlogram displayed in Figure 7.12 Those p-values outside of the cut-off of 0.05 are coloured in red.

		TDP-43		BCL11b	
ERVMap ID	HERV	R2	P-value	R2	P-value
673	HERV-K22	-0.43085	0.045304	-0.38301	0.078509
1115	HERV-K9	0.591986	0.003703	0.418535	0.052553
1379	HERV-K3	0.499029	0.018065	0.337674	0.124311
1549	HERV-K22	0.759579	0.0000413	0.435582	0.042736
2360	HERV-W	0.592743	0.003648	0.451861	0.034758
3167	HERV-K9	-0.5364	0.010067	-0.32748	0.136823
3388	HERV-K13	0.59984	0.003168	0.424933	0.048684
3606	HERV-K22	0.558118	0.006949	0.305038	0.167464
3776	HERV-K22	0.804955	0.00000625	0.605269	0.002838
3866	HERV-K22	0.84146	0.000000925	0.600871	0.003103
4340	HERV-K9	0.529835	0.011208	0.548496	0.008213
4757	HERV-K3	-0.0554	0.806571	0.314953	0.153389
5446	HERV-K9	0.557186	0.007063	0.497617	0.018446
6078	HERV-K22	0.312005	0.157483	0.03687	0.8706
K-46	HERV-K46	0.082522	0.715044	0.301515	0.172673
W-92	HERV-W	0.474321	0.025727	0.186974	0.404744

As we can see in Table 7.2 several of these correlations fall outside of the correlogram p-value cut-off of 0.05 with 3 HERV-K22 loci showing a significant positive association with both of these cellular genes. A single HERV-K22 locus also showed a significant negative correlation with TDP-43 in contrast to the rest of the HERV-K22 loci differentially expressed in this PBMC dataset (ERVID 673). In addition, more HERV-K families showed a significant positive correlation with TDP-43 expression than BCL11b and a single HERV-W was also detected as significantly correlated with TDP-43. To investigate whether TDP-43 and BCL11b are significantly differentially expressed in ALS compared with non-ALS PBMC samples, in which DESeq2 was used to generate differential expression data as shown in Table 7.3. The data in Table 7.3 shows that neither BCL11b or TDP-43 were significantly expressed in ALS by either p-value or adjusted p-value and does not support findings reported in the literature that observe overexpression of TDP-43 in ALS brain tissue and could be used as a potential biomarker for ALS.

Table 7.3. DESeq2 Differential Expression Results for TDP-43 and BCL11b in Peripheral Blood Mononuclear Cells Between ALS and Non-ALS Controls.
The table below shows Log2 fold changes in expression of TDP-43 and BCL11b between ALS and non-ALS controls.

Ensembl Gene ID	Gene	baseMean	log2FoldChange	lfcSE	pvalue	padj
ENSG00000120948	TDP-43	496.44668	0.037451361	0.253573	0.882584	0.937241
ENSG00000127152	BCL11b	2048.6354	0.028898121	0.193889	0.881519	0.879876

7.2.3. Analysis of HERV-K22 & HERV-H Open Reading Frames for Intact Protein Fragments.

The ERV sequences for differentially expressed ERVs belonging to HERV-K22 (such as ERVID 6078 mentioned earlier in this chapter) and HERV-H family members were identified in the ERVmap.bed reference file and sequences for the region obtained from UCSC genome browser. These families were focused on as they represent the most common significant differentially expressed ERVs in the results and were found to be significantly expressed in Chapter 6 in brain tissue regions. As PBMCs are supposed to act as an effective mirror for gene expression in other tissues these were deemed to be good candidates for potential novel biomarkers so further analysis of potentially expressed proteins was needed. Consensus sequences for the internal region of the identified HERV-K22 sequences were downloaded from DFam and a consensus sequence which included 5' and 3' LTRs was constructed from the information given in Lavie *et.al.* (2004). The HERV-H family consensus sequence was also downloaded from Dfam though this only covered the internal region of the ERV. Each of the ERVMap identified sequences were aligned separately against the consensus sequences to look for regions of high similarity. These were initially aligned in MegaX using the ClustalW alignment algorithm and analysed within UGene.

The amino acid sequences from the HERV-K22 consensus sequence genomic regions for *gag*, *prot*, *pol* and *env* sequences were initially identified using UGene's ORF identifier tool. The amino acid sequences for the open reading frames were copied and entered into the SWISS-Model tool to obtain 3D structure models to align with ORFs identified from ERVMap regions. Figure 7.2 shows the protein models obtained from the HERV-K22 consensus sequence genomic regions which were used as a basis to look for similar

conserved sequences in the differentially expressed HERV-K22 family members identified in the DESeq2 differential expression analysis. Similarly Figure 7.3 shows the 3D model for the translated protein open reading frame from the HERV-H consensus sequence.

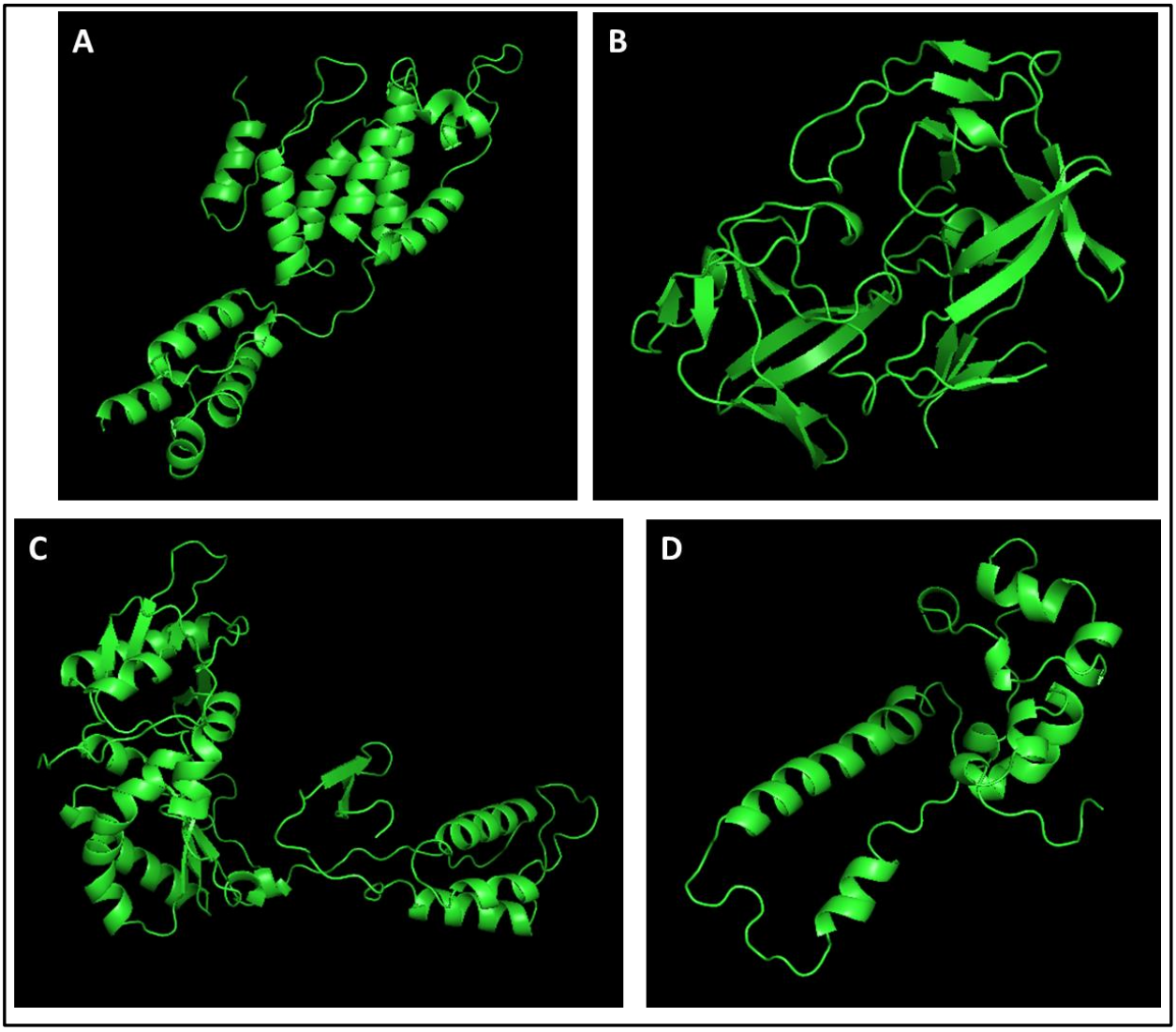


Figure 7.2 SWISS-Model 3D Protein Models for Open Reading Frames Identified in HERV-K22 Consensus Sequence

The figure above shows the SWISS-Model 3D protein models for translated HERV-K22 consensus sequence protein regions. The 3D models in the figure above represent A) *gag*, B) *prot*, C) *pol* and D) *env* genomic regions. Protein models from A) *gag* and D) *env* represent only part of the full sequence region as SWISS-Model was only able to identify structure in specific regions of the protein.

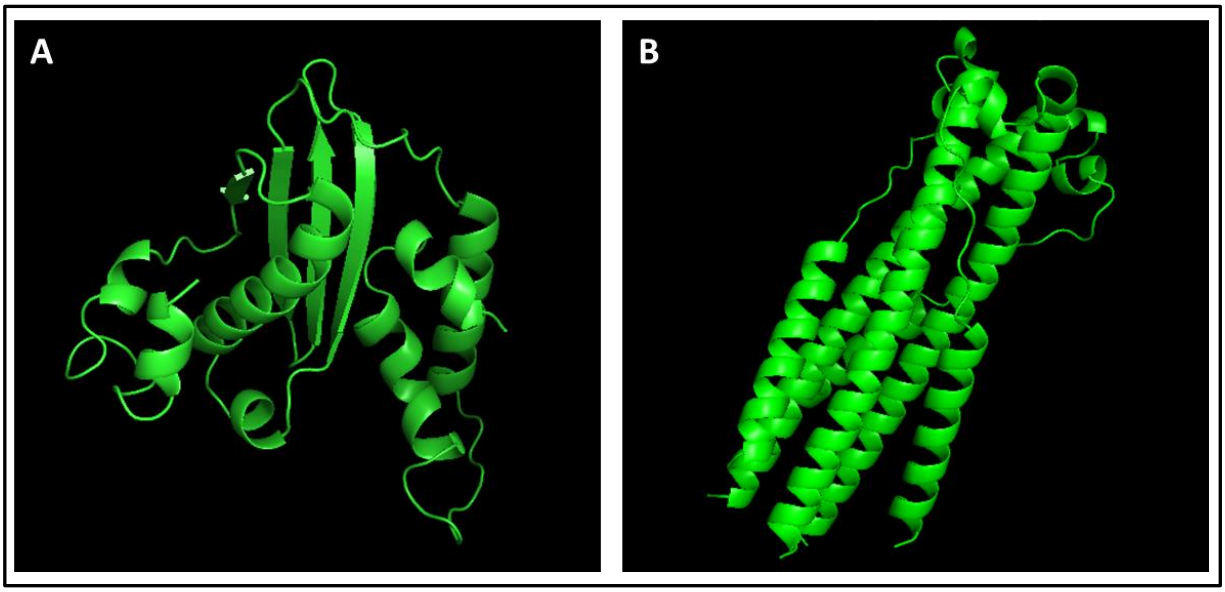


Figure 7.3 SWISS-Model 3D Protein Models for Open Reading Frames Identified in HERV-H Consensus Sequence

The figure above shows the SWISS-Model 3D protein models for translated HERV-H consensus sequence protein regions. The 3D models in the figure above represent A) RNaseH and B) *env* genomic regions. Protein models from B) *env* represent a single protein sequence which appears as a trimer from the full sequence region as SWISS-Model was only able to identify structure in a specific region of the protein.

Open reading frames (ORFs) from ERVMap regions were chosen for analysis based on the presence of known functional motifs and the length of the individual ORF. Within HERV-K22 this was the motif for the *pol* reverse transcriptase functional site, shown as FMDD in the consensus sequence. Other open reading frames identified by the UGene open reading frame tool were assessed for similarity to known HERV proteins by NCBI's BLASTp tool but did not return any results relating to human sequences. An example of an identified reading frame is given in Figure 7.4 which shows the ERVMap 673 HERV-K22 sequence which showed the highest similarity to the consensus sequence. Highlighted in this figure is the open reading frame (Red box) and functional motif (Green box) for the reverse transcriptase section of the *pol* polyprotein. For HERV-H the regions identified in the consensus sequence only found ORFs for the RNaseH and *env* regions but due to high variability within the HERV-H sequences identified by ERVMap only a single HERV-H member, 1143, had open reading frames similar to these proteins or ORFs related to human sequences.

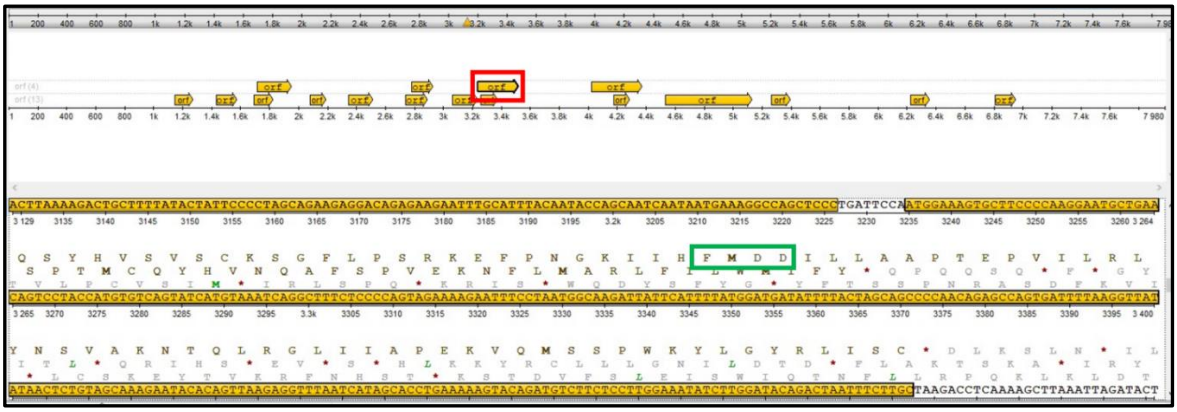


Figure 7.4 Open Reading Frames for ERVMap ID 673 HERV-K22 Sequence
The figure above shows the open reading frames (ORF) identified by UGene for the ERVMap ID 673 HERV-K22 nucleotide sequence. The ORF identified by the red box is shown in the lower half of the image with the reverse transcriptase functional motif highlighted by the green box.

These open reading frames were isolated from each of the HERV-K22 and HERV-H sequence members and entered into the SWISS-Model web tool provided by ExPASy to generate a 3D model for each of the differentially expressed HERV-K Sequences. Figure 7.5 shows the alignments for 4 of the HERV-K22 sequences, ERVMap ID 3866 did not have any open reading frames conforming to known human HERV proteins and the open reading frame for ID 6078 did not align significantly to the consensus sequence (RMSD 23.680). The alignment score for each of the pol fragments is given in the top right of each alignment image in RMSD which stands for Root-mean-square deviation of atomic positions. This is a calculation of the average distance between atoms in a superimposed image. As we can see from the HERV-K22 alignments the ERVMap 673 HERV-K22 is the closest to the consensus sequence, which mirrors its nucleotide sequences high similarity to the HERV-K22 consensus.

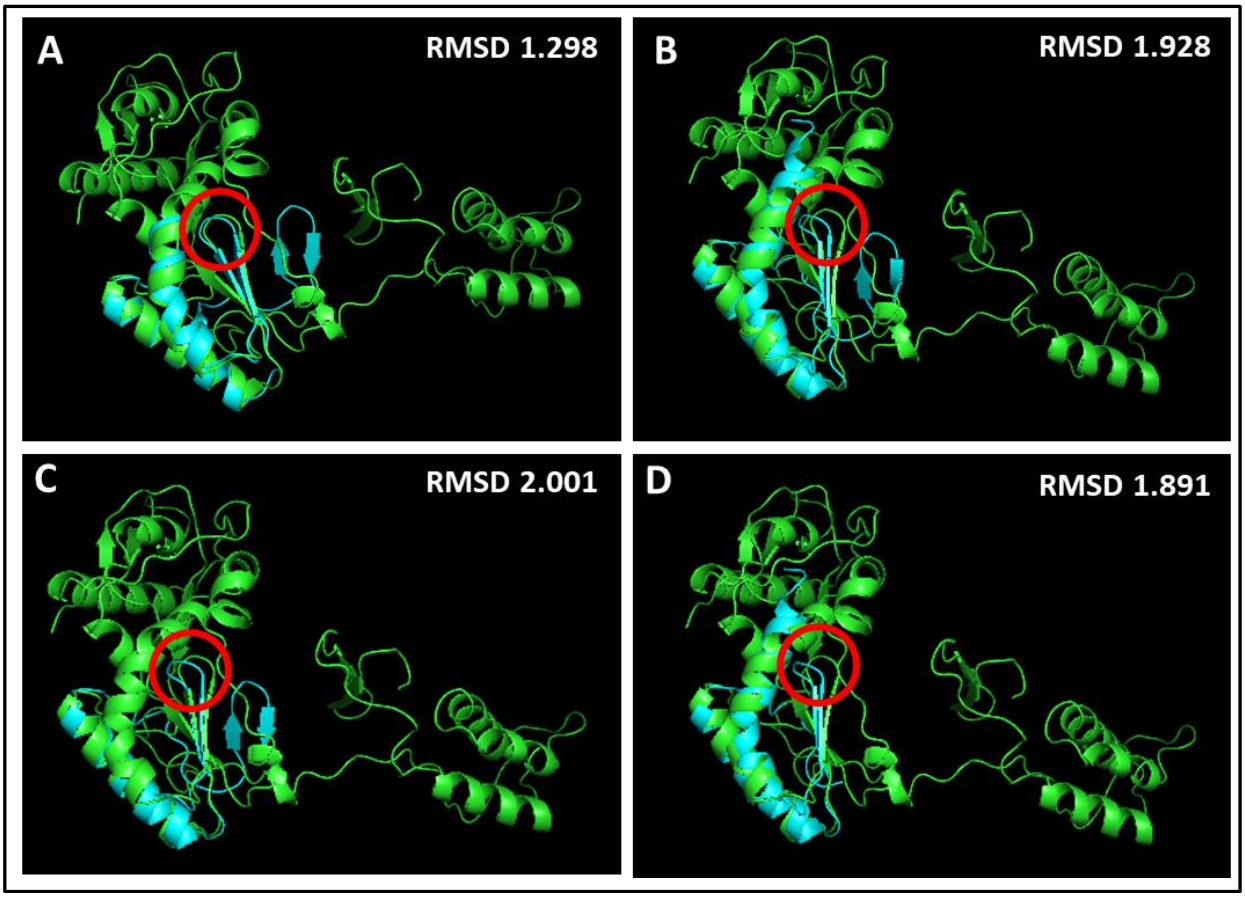


Figure 7.5 3D Model Protein Alignments for the *pol* Region ORFs for HERV-K22 Sequences That Were Differentially Expressed in ALS.
The figure above shows the PyMOL output for the alignment of HERV-K22 *pol* open reading frame fragments for significantly expressed ERVMap regions. The 3D model for a ERVMap sequence (blue) was aligned against the *pol* consensus region (green) and a Root-mean-square deviation of atomic positions (RMSD) score given for each alignment. The ERVMap ID's for the images in the figure above are A) 673, B) 1549, C) 3606 and D) 3776. The red circled region in each of the images identifies the reverse transcriptase active site in each alignment.

Only a single HERV-H family member showed open reading frames for either of the consensus sequence regions, ERVMap ID 1143. Unlike the HERV-K22 regions however there were 2 open reading frames which could be identified as being similar to the consensus sequence, one for the RNaseH region and one for the *env* protein region. Figure 7.6 below shows the alignments for these HERV-H 3D models, with the RMSD score showing that these sequences align very closely to their consensus sequences. It should be noted that the MER57A family member highlighted earlier in this chapter was also analysed for significant open reading frames and while a transmembrane *env* protein motif was found similar to HERV-H the ORF from the UCSC sequence identified as an RT molecule which was not present in the consensus.

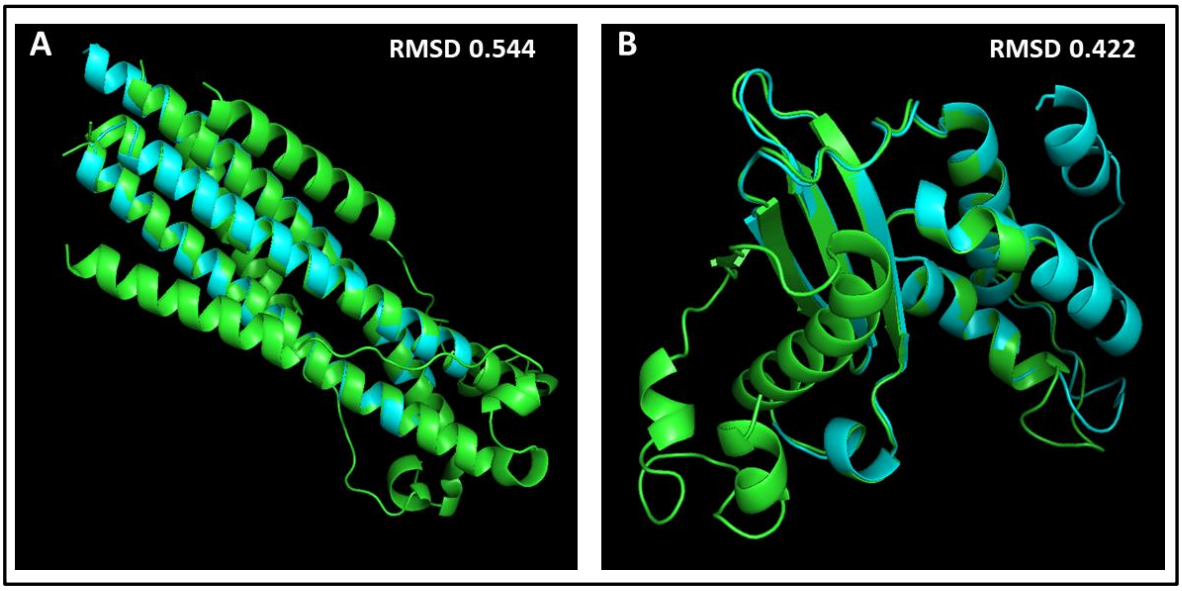


Figure 7.6. 3D Model Protein Alignments for the *env* and RNaseH Region ORFs in Significant HERV-H Sequence 1143.
The figure above shows the PyMOL output for the alignment of HERV-H A) *env* and B) RNaseH open reading frame fragments for ERVMap ID 1143. The 3D model for the ERVMap sequence (blue) was aligned against the consensus region (green) for each of the protein regions and a Root-mean-square deviation of atomic positions (RMSD) score given for each alignment.

7.2.4. Investigation of HERV-K and HERV-H Regions for Nearby LTR Promotor Sequences

As HERV-K sequences are the ones predominantly found to be differentially expressed in ALS, and with the significant result for HERV-H seen in the previous chapter, the differentially expressed HERVs from these families seen in PBMCs were analysed to see whether they were flanked by LTR sequences. The flanking location and annotation of those LTRs was recorded in Table 7.4 and the sequence downloaded from the UCSC genome browser for downstream analysis. These sequences were analysed for sequence motifs relating to promotor/enhancer sequences common to LTR regions. The paper by Manghera and Douville (2013) provided a detailed list of multiple generic and hormone specific sequences found in a multiple sequence alignment of full length (5'LTR-gag-prot-pol-env-3'LTR) HERV-K sequences, including HERV-K115 (The Genbank sequence used in Chapter 3.0), of which the generic transcriptional promoters were recorded for use in the analysis. An additional well-known promotor sequence known as a TATA Box was also recorded for use in the analysis. A single Hormone specific Androgen sequence was

included in the analysis as a there has been a significant difference in expression based on patient sex in brain tissue RNA-Seq data as shown in Chapter 6.0.

The presence of the promotor/enhancer sequences within the LTR sequences was identified by searching for known motifs identified in Manghera and Douville (2013) and recording similar sequences in Table 7.4. As we can see from Table 7.4 the majority of sequences significantly expressed in peripheral blood mononuclear cells with LTRs were flanked on both ends of the HERV sequence.

Table 7.4 Promotor Sequences Appearing in LTR Regions Flanking Significantly Expressed HERVs in ALS Derived PBMCs .

The table below displays information about LTR sequences found in the ERVMap Region associated with significantly expressed endogenous retrovirus family members identified by DESeq2.

ERV (Tissue)	LTR's Present	LTR Type	5' Promoters (Amount if > 1)	3' Promoters (Amount if > 1)
570, HERV-H	5', 3'	LTR7	YY1, GC Box (2)	YY1, GC Box
909, HERV-H	5', 3'	LTR7	GC Box (2)	YY1, GC Box (2)
1143, HERV-H	5', 3'	LTR7	GC Box (2, Both C repeat), TATA Box	GC Box (2), TATA Box
1679, HERV-H	5'	LTR7B	GC Box (C repeat)	N/A
1739, HERV-H	5', 3'	LTR7	TATA Box	GC Box, TATA Box
2334, HERV-H	5', 3'	LTR7	GC Box (2, Both C repeat), TATA Box	GC Box (3, 2*C repeat), TATA Box
2621, HERV-H	5'	LTR7B	GC Box	N/A
2724, HERV-H	5', 3'	LTR7	YY1, GC Box	YY1
6195, HERV-H	5', 3'	LTR7C	GC Box (C repeat)	TATA Box, GC Box (C repeat)
4757, HERV-K3	5', 3'	LTR3A		
1379, HERV-K3	5', 3'	LTR3B		
K-46, HERV-K	5', 3'	LTR5B	YY1, GC Box (1 + 1 C repeat)	GA Box, GC Box (2 + 1 C Repeat), YY1
6078, HERV-K22	5', 3'	LTR22B1		
673, HERV-K22	5', 3'	LTR22C0	YY1, GC Box	YY1
1549, HERV-K22	3'	LTR22C0	N/A	
3776, HERV-K22	5', 3'	LTR22E		GC Box (C repeat)
3866, HERV-K22	5'	LTR22E		N/A
5446, HERV-K9	5'	MER9a1	GC Box (2)	N/A
3167, HERV-K9	5', 3'	MER9a2		GC Box
1115, HERV-K9	5', 3'	MER9a3	GC Box	
4340, HERV-K9	3'	MER9a3	N/A	

The most frequently occuring promotor sequence "family" was the GC box (classical sequence GGGCGG), which canonically appears upstream of the TATA box promotor sequence. This concurrent appearance with the TATA box was seen in individual LTR regions but was not in the overall trend of the data. The C rich GC box sequence CCCCC is also commonly represented in Table 7.4, with 2 of the differentially expressed HERV-H regions (ERVID: 1143 and 2334) having a duplicate of this sequence close to the initial insert (Figure 7.7).

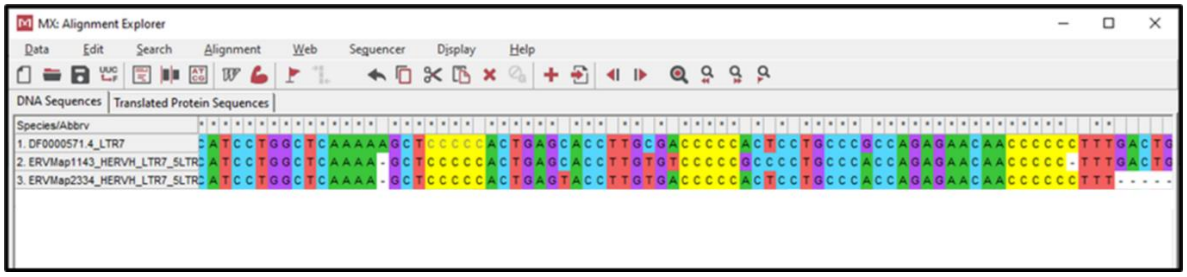


Figure 7.7 Multiple Sequence Alignment of 5'LTRs for HERV-H sequences with multiple occurrences of C-Rich GC Box

The figure above shows an example alignment of HERV-H LTR7 from the 5' end of 3 significantly expressed HERV-H sequences aligned against their consensus sequence obtained from DFam. As we can see in the image there are several example of the C-Rich GC box.

7.3 Discussion

Amyotrophic Lateral Sclerosis (ALS) is a fatal neurodegenerative disease which results in progressive paralysis and death, normally within 5 years of diagnosis (Dadon-Nachum, Melamed and Offen, 2011). While Peripheral Blood Mononuclear Cells (PBMCs) are not currently considered to be involved in disease progression they are a useful method of monitoring gene expression in the disease due to previous identification effectively mirroring gene expression in Multiple Sclerosis (Achiron and Gurevich, 2006). This may be unique to multiple sclerosis however as a more recent paper has shown weak correlation to brain tissue expression (Cai *et al.*, 2010). If a novel biomarker in the form of a Human Endogenous Retrovirus (HERV) was elucidated, then PBMCs could be useful in terms of a sample source for early diagnosis of sporadic ALS, alongside prediction of treatment response in the advent of HIV-1 ART. (Swindell *et al.*, 2019).

As displayed in table 7.1 n=55 Endogenous Retroviruses (ERVs) were identified to be differentially expressed in ALS in publicly available RNA-seq data performed on PBMCs derived from ALS and non-ALS controls. The majority of ERVs discovered are upregulated, though there are examples of an ERV family being both up and down regulated on the same genome ie: HERV-K is both up and downregulated on Chromosome 11, though it should be noted these are separate family members. In the ERV1 UCSC family there are also examples of HERVs being differentially expressed in the same genetic locus, in 19p12 both ERV 4656 (HERV3, ERV1), 3.2kbp downstream of ZNF493 (Zinc Finger Protein 493) and ERV 4678 (HERV-H, ERV1) located within intron of ZNF91 (Zinc Finger Protein 91) are significantly upregulated though neither are near/within genes that are associated with ALS or other neurological conditions that have been reported to-date. . There are 3 upregulated ERVs on 6q22.31 all relatively close to each other as also noted in their ERVMap designation, ERV 2307 (HERV30, ERV1), ERV 2306 (HERVIP10B3, ERV1) and ERV 2305 (HUERS-P3, ERV1), all identified within introns of CEP85L (Centrosomal Protein 85 Like), which has been associated with breast cancer but not with neurological conditions. Finally Xp11.23 has 2 upregulated ERVs, such as ERV 5359 (MER89, ERV1), within intron of ZNF81 (Zinc Finger Protein 81), and mutations in this gene cause an X-linked form of intellectual disability (MRX45) and secondly, ERV 5361 (HERV-E, ERV1) 13kbp upstream of MAGIX (MAGI Family Member, X-Linked), Diseases associated with MAGIX include Mental Retardation; interestingly this locus holds the only 2 ERVs which are close to neurological condition determinant genes identified so far.

All ERVMap identified ERV regions were additionally searched for open reading frames within their genetic code to look for intact *gag-pol-env* regions along with other potential protein coding frames. Only 2 ERVs showed identifiable protein coding open reading frames, W-92 (HERV17, ERV1) showed an open reading frame for Syncytin-1 (Supplementary Figure S118) a known neurotoxic protein active in multiple sclerosis (Grandi and Tramontano, 2017, 2018). The other ERV shown to have an open reading frame was ERV 1115 (HERV-K9, Supplementary Figure S123) which, as previously mentioned, contains a single open reading frame for its protease gene. Interestingly, ERV 3200 (HERVIP1-F, Supplementary Figure S134) contains a different open reading frame in each of its reading frames, none of which code for a viral protein. NCBI nucleotide and nucleotide

to protein BLAST searches were also performed on all the ERVs identified by ERVMap that passed the adjusted p-value cut-off of 0.05, the majority of search results indicated non-human ERV matches as the highest similarity to the genetic code for the region if an ERV was detected at all, with the exception of W-92. This backs up the ExPASy translation of the ERVMap identified regions as few open reading frames were discovered, indicating lack of functional proteins produced.

Similar to chapter 6.0 our dispersion plot (Supplementary Figure S254) shows data conforming to the expected spread of ERV expression recorded in our DESeq2 differential expression analysis. As before, this provides a good overview of our dataset as there are only a few outliers from the ERVs that are shrunk towards the mean (Love, Huber and Anders, 2014). While the distribution of p-values for ERVs within the dataset initially appears to have a left leaning conservative distribution when corrected for multiple testing the adjusted p-value shows a largely uniform distribution (Supplementary Figures S255 & S256). Looking at the coverage distribution plots in Supplementary Figures S255 and S256 we see a graphical representation of the reads aligning to the example HERV loci and the depth of those reads represented by the heights of the individual peaks. In these two figures we have opposite ends of the scale for the significant differentially expressed ERV regions, one with noticeable patterns in coverage across the genome and the other with no discernible similarity between either ALS or non-ALS controls. One potential reason for the differential expression observed in the MER57A genomic region is the lack of any mapped reads to two of the control samples, despite there being a low number of reads aligning to the region. A more obvious difference is seen in Supplementary Figure S257 where the coverage of the region in ALS samples are more even, the lower read count in the control samples gives a clear difference between the expression levels between the two groups. The uniformity of coverage across a coding region represents the length of the gene transcribed from the RNA, while the genomic region for ERV 6078 has good coverage the intermittent gaps seems to confirm the reading frame translations shown in the supplementary material in which there are no intact reading frames for functional proteins such as ERV envelope (Conesa *et al.*, 2016).

While the data displayed in the heatmap (Supplementary Figure S259) and PCA plots (Supplementary Figures S260 & S261) primarily show there is a significant difference in

expression profiles between ALS and non-ALS controls there was no similar divide between male/female samples (Data not shown), this suggests gender did not influence the expression data in the publicly available RNA-Seq data set. There was however a definite grouping of patient samples from the lower age range separated from the larger cluster of the other age ranges in the PCA plot (SRR7251667-SRR7251670, Supplementary Figure S261). These samples obtained from patients with a younger age range in Supplementary Figure S261 also appeared to be the control samples for the dataset as seen in Supplementary Figure S260, so it is unclear whether age or their nature as controls was the cause of the separate grouping of the samples on PCA plot and requires further investigation. While there is no specific data on RIN values within the sample set as all reported as being above 8.0 in the original paper (Zucca *et al.*, 2019) then we were unable to determine if there was a correlation of ERV expression with RINs. From the previous RNA-Seq data chapter we described (chapter 6.0) and the RT-qPCR data (chapter 4.0), we reported that RINs ≤ 7.0 could potentially affect measurement of differential gene expression of HERVs in samples with low RNA values. However, it has been reported that RINs are an ineffective measure of how intact total RNA in the brain especially, as the tissue has overall lower recorded RINs than other tissue types (Koppelkamm *et al.*, 2011). A distinct advantage of using PBMCs is that it alleviates the inherent problems associated with postmortem delay and RIN values that is a problem when working with post-mortem brain tissue. However a study looking at RNA-seq from various tissues showed that a lower RIN did not have an effect on the quality of RNA reads from individual samples (Suntsova *et al.*, 2019).

As tar DNA binding protein 43 (TDP-43) has been reported as being differentially upregulated in ALS patients and could serve as a potential biomarker of the disease, we performed an analysis of its potential co-expression with those ERVs which were significantly dysregulated in our sample set as well as seeing if we could confirm research findings showing that it is upregulated in ALS compared with controls (Cohen, Lee and Trojanowski, 2011; Li *et al.*, 2015; Manghera, Ferguson-Parry and Douville, 2016; Douville and Nath, 2017; Prudencio *et al.*, 2017). In addition, a regulator of viral senescence in HIV infections of the spinal cord, B-Cell CLL/Lymphoma 11B (BCL11b) was included in our analysis, as it has been theorised to have a potential role in suppression of ERV expression

(Cismasiu *et al.*, 2008; Desplats *et al.*, 2013; Lennon *et al.*, 2016, 2017). The correlogram displayed in Figure 7.1 provides a visual key for the expression analysis of these genes compared to significantly expressed ERVs, while there does appear to be some ERVs which have positive correlations with TDP-43 we would expect to see this pattern across all ERVs if there were a definitive link between expression profiles. Additionally, there are several correlations in the sample set which are not statistically significant in both TDP-43 and BCL11b, along with weak significant correlations indicated by pale red/blue squares in the figure. From this co-expression analysis further analysis is warranted into looking at a possible correlation of these two transcriptional regulators and their association with certain ERVs in ALS in a larger PBMC sample set we are going to obtain from King's College London, UK , looking in a cohort of n=40 ALS and n=40 non-ALS controls in more detail. Unfortunately, due to the COVID-19 pandemic all face-to-face meetings at clinics have been put on hold as ALS patients have been classified as high-risk for complications due to the virus. Because of this we have been unable to obtain any PBMC samples for reference gene selection and determination of relative expression in the specimen type by RT-qPCR.

The analysis of open reading frames (ORFs) seen in section 7.2.3 attempts to show whether a functional protein can be produced from these sequences despite their truncation by stop codons introduced during our evolutionary history. As we can see in the 3D models produced from ORFs from HERV-K22 *pol* sequences (Figure 7.2) the one that is readily identified as persevered in those identified by DESeq2 differential expression analysis. More precisely the reverse transcriptase section of the *pol* polyprotein is intact with RNaseH and integrase regions being more effectively disrupted by stop codons. There is a possibility that the full genomic region could be transcribed despite stop codon inclusion however, in the form of an effect known as stop codon read-through. This essentially is the cell "overriding" a stop codon that may have been introduced into a gene through nonsense mutation. Though several genes in the human genome have detectable stop codon read-through it is thought that this is largely due to a non-adaptive molecular error and has not been observed in the literature for HERVs (Li and Zhang, 2019).

While the majority of HERV sequences that exist in the human genome consist of solo LTRs or either have multiple stop codons disrupting their *gag, prot, pol* and *env* genomic regions

or are truncated due to multiple silencing mutations that have occurred over our evolution their LTR's (if present) still have the opportunity to contain sequences for enhancing or initiating transcription (Yu, Zhao and Zhu, 2013). It has been shown in the literature that around 50% of HERV LTR sequences contain promotors or enhances for cellular gene expression (Buzdin *et al.*, 2006). While this promotor activity has been better categorised in Cancer patients the ability to effect transcription of nearby cellular genes in other diseases such as ALS should be taken into consideration. From the data obtained in Table 7.4 we can see that one particular group of promotor sequences is most common in the LTR sequences studied. The GC box is either commonly found downstream of a TATA box or considered a TATA independent promotor sequence for cellular expression (Yang *et al.*, 2007). It is interesting that HERV-H sequences show more intact LTR promotor sequences than the HERV-K LTR sequences analysed. This has been explored in a recent paper by Gemmell, Hein and Katzourakis (2019) which showed that the integrity of these regions is directly related to how well it is expressed in the cell. This would seem to correlate with our data as these LTR regions appear to have a higher number of intact promotors. However, while this data is interesting, specific experimentation to see whether these promotors are active type would be warranted to confirm whether their activity is related to the enhancement of nearby gene expression.

8.0 Discussion

8.1 Introduction

As research into Human Endogenous Retroviruses continues to shed light on transposable elements within the human genome the mystery of their role in human disease is slowly unravelled. For more than a decade now HERVs have been found to contribute to the pathology of many diseases, most often found in Cancer, but have also been found in Diabetes, Autoimmune disorders and more recently neurological conditions like Schizophrenia, Alzheimer's, Multiple Sclerosis and Amyotrophic Lateral Sclerosis (Hohn, Hanke and Bannert, 2013; Douville and Nath, 2014; Mason *et al.*, 2014; Li *et al.*, 2015, 2019, 2022; Hanke, Hohn and Bannert, 2016; Grandi *et al.*, 2017; Arru *et al.*, 2018b; Wang, Huang and Zhu, 2018; Savage *et al.*, 2018). Categorisation of HERV roles in these disorders varies between diseases however, with Cancer cells thought to take advantage of their roles in transcriptional regulation but their specific role in other conditions still not fully explained (Buzdin, Prassolov and Garazha, 2017; Montesion *et al.*, 2017; Byun *et al.*, 2021). The pathogenic role of HERVs in ALS in particular, while their differential expression has been studied, is still a subject of debate. Studies have looked into the neurotoxicity of HERV *env* elements, role in inflammation and their co-expression with TDP-43 as potential pathogenic determinants in ALS but the full picture has yet to be resolved (Oluwole *et al.*, 2007; Lemaitre *et al.*, 2014; Li *et al.*, 2015; Phan *et al.*, 2021; Simula *et al.*, 2021).

The research performed as part of this PhD thesis has focused primarily on differential expression of HERV transcripts within sporadic ALS as this form of the disease accounts for 90% of all ALS cases and its eitology is largely unknown (Ajroud-Driss and Siddique, 2015; Valko and Ciesla, 2019). From the initial paper linking a retroviral source for the disease in the 1970's to more recent papers looking at reverse transcriptase activity in ALS and a key paper published by Li et al., (2015), found HERV-K (HML2) transcripts to be differentially expressed in the motor cortex of ALS cases in an American cohort, and could be a potential biomarker for early disease diagnosis that is lacking to-date (Viola *et al.*, 1975; Norris, 1977; McCormick *et al.*, 2008; Arru *et al.*, 2018b). While RT-qPCR and microarrays were initially used to determine differential gene expression between control and disease states recent advances in the field of next generation sequencing has allowed for the non-specific application of looking at the entire transcriptome of a tissue type or even a singular cell

which enables a broad screening of differential expression of all known HERVs compared with RT-qPCR that is target specific (Bustin *et al.*, 2009; Derveaux, Vandesompele and Hellemans, 2010; Lowe *et al.*, 2017; Mayer *et al.*, 2018). In this section I will be discussing each of the research findings I have obtained based on differential expression of HERV transcripts in sporadic ALS using RT-qPCR methodology as well as more cutting-edge RNA seq analysis that I have performed and looking in different anatomical regions of the brain as well as blood of ALS patients compared with non-ALS cases. In addition, to investigate whether differentially expressed HERVs that have been identified have intact LTRs as well as ORFs that have the ability to form functional proteins which could be the key to understanding the pathology of HERVs in ALS.

8.2 Summary and Discussion of Results

One of the aims of this research study was to independently replicate findings by Li et al., (2015) to see if HERV-K (HML-2) was differentially expressed in the premotor cortex of sporadic ALS patients compared to non-ALS cases using the exact same primers and FAST SYBR green chemistry in our MIQE compliant RT-qPCR assay. Validation of experimental processes is an invaluable start to any research project, choosing the correct reference genes for the tissue type, primers that are specific to the gene of interest and accurately double the amplicon target every round of PCR are vital to the accurate measurement of differential expression by RT-qPCR. An important document to this effect is the minimum information required for publication of RT-qPCR data (MIQE) guidelines which set out a framework for consistent reporting of RT-qPCR data between research papers (Bustin *et al.*, 2009). During the validation of the RT-qPCR assay we identified that GAPDH and XPNPEP1 were the optimal reference genes for use on post-mortem premotor cortex brain tissue samples using several different reference gene selection algorithms in order to select the best reference genes that are stably expressed in disease and non-disease state as well as the correct number of reference genes for normalisation purposes, in which our findings were published in Garson et al., (2019). These reference gene selection methods mostly gave similar results, agreeing that 2 reference genes were the optimal quantity for normalising the RT-qPCR assays, and GAPDH and XPNPEP1 were included in all of our RT-qPCR assays on postmortem brain tissue based on this. However, there was a single outlier in this analysis; the BestKeeper algorithm which showed a completely different ranking to

the other methods used in the analysis. This may be expected though as this method had been shown to be unreliable across different RT-qPCR experiments which were conducted on tissues other than premotor cortex samples (Petriccione *et al.*, 2015).

As the Li *et.al.* (2015) study had a set of primers to target gag, pol and env transcripts of HERV-K (HML2) provirus, I also designed a set of HERV-K pol primers and validated these alongside HERV-W *env*, TDP-43 and BCL11b primers to determine the specificity and efficiency of these primer sets. The new HERV-K (HML2) pol primers were designed in the hope of targeting the section of the provirus that was more integral to retroviral replication, the reverse transcriptase enzyme as the Li *et.al.* pol primers targeted integrase which is responsible for incorporating the retroviral DNA into the host genome following reverse transcription.

Once the primers had been validated for targeting HERV-K (HML2) transcripts, and the RT-qPCR assay was deemed MIQE compliant we sought and obtained ethical approval to obtain fresh frozen premotor cortex brain tissue samples from sporadic ALS and non-ALS cases from the MRC neurodegenerative disease brain bank, which were processed to extract total RNA for analysis. In total RNA from n=20 ALS and n=20 non-ALS controls were extracted from post-mortem premotor cortex brain tissue samples which made up the sample set for the initial RT-qPCR experiments (Chapter 4.0). These samples all had RINs that were above a cut-off of 4 which was deemed acceptable for these brain tissue samples to be used in this assay. While a higher RIN cut-off would normally be used for other tissue types research has shown that RIN is not an ideal measure of RNA integrity for brain tissue samples (Stan *et al.*, 2006; Sonntag *et al.*, 2016).

Initial RT-qPCR analysis of the premotor cortex sample set was performed on a slightly reduced sample set of n=19 ALS and n=20 non-ALS controls as one ALS sample (A331/09) had too low a concentration of RNA to be measured effectively (Table 4.1). However, despite the robust nature of the remaining samples none of the HERV-K (HML2) primer targets used in the assays were shown to be significantly differentially regulated in the premotor cortex (Table 4.2). This was one of the first studies in sporadic ALS to show that HERV-K (HML2) transcripts were not differentially regulated in the premotor cortex, opposite to the data presented earlier by Li *et.al.* (2015). This lack of significance did not change when non-ALS samples that had cancer as co-morbidities were removed from the

dataset or when additional no-cancer control samples were added to replace those lost (Tables 4.4-4.8). These negative results when looking at HERV-K (HML2) gene targets proved to be the first in a series of subsequent negative findings when looking for differential expression of this particular HERV-K family in ALS post-mortem brain tissue Garson et.al. (2019) looked in a different sample set of premotor cortex tissue obtained from sporadic ALS patients from the same UK patient cohort that I had obtained premotor cortex tissue from and also reported no significant difference in HERV-K (HML2) transcript expression compared to control cases. In addition to this, a paper published this year looking into HERV-K (HML2) expression in the premotor cortex obtained from a Japanese ALS patient cohort, also failed to confirm Li et al., 2015, findings of differential expression of HERV-K (HML2) in ALS (Ishihara et al., 2022).

As HERV-K family members from the HML-2 subgroup were not found to be differentially expressed in our UK ALS cohort by RT-qPCR, and we worked closely with our collaborators at Kings College London, to perform RNA seq analysis to undertake a broad screening of all known HERV families that might be differentially expressed in ALS and would be missed using our primer sets in RT-qPCR that targeted only HERV-K (HML2) families. In the study by our collaborators, Jones *et al.* (2021) the research group found a unique HERV-K3 HML6 provirus from locus 3p21.31c to be differentially expressed in ALS vs Controls in both the frontal cortex and the primary motor cortex tissue types. In order to confirm this finding by RT-qPCR the KCL research group provided us with fresh frozen primary motor cortex tissue from n=10 ALS and n=10 non-ALS controls which were preferentially selected as the n=10 ALS samples we received had the most counts for this particular ERV and selected a set of n= 10 controls with the lowest ERV counts to see if we could confirm RNA seq findings by RT-qPCR. After validation of primer sets to target both pol and env transcripts specific to HERV-K3 (HML6) locus in chromosome 3, using both SYBR green and TaqMan/probe approaches and utilising GAPDH and XPNPEP1 as reference genes (as they had been previously validated on post-mortem brain tissue samples) the HERV-K3 *pol* gene target was shown to be significantly upregulated in the primary motor cortex (Table 5.2) compared to controls and we were able to confirm RNA seq findings (Manuscript in preparation). The HERV-K3 sequence at this particular locus, was obtained from the UCSC genome browser, and shown to have open reading frames for both *pol* and *env* genomic

regions. The integrase open reading frame situated within pol, was predicted to have a dimeric structure by SWISS model with an active site in between the protein dimers. It was also shown to have the translated amplicon sequence at the start of the amino acid chain (Figure 5.4A). This shows that the primer targeted a potentially fully transcribed protein in the locus.

Considering that this specific locus of HERV-K3 was elevated in primary motor cortex tissue samples, some of whose patient samples were taken from the same donors as our premotor cortex tissue samples, it was decided that we should test for the expression of this locus on our ALS and non-ALS premotor cortex sample set to see if HERV-K3 expression is differentially expressed in different anatomical regions of the brain. Combining the initial premotor cortex brain tissue samples used in our MIQE compliant RT-qPCR assay we used to measure relative expression of HERV-K (HML2) (chapter 4.0) with those tested by Garson *et.al.* (2019) a larger sample set (n=91) was amalgamated consisting of n=54 ALS and n=37 non-ALS controls. As the copy number of the HERV-K3 transcripts from the 3p21.31c locus appeared to be of low copy number n=8 ALS and n=8 non-ALS controls had to be removed from the subsequent analysis due to poor replicate SD values. Unfortunately, we were unable to confirm a significant upregulation of HERV-K3 pol transcript from the 3p21.31c locus, with both the 2^-ΔΔCt and Pfaffl methods showing a lack of statistical significance (Table 5.3). This lack of significance in HERV-K3 expression could be due to many reasons. Research papers have shown that expression of genes sets can be the same or vary across tissue types which could account for why this HERV-K3 locus could not be verified in the premotor cortex samples as RNA seq analysis was performed on the primary motor cortex (Lederer *et al.*, 2007; Phan *et al.*, 2021). Alternatively the inherent limitations of the RT-qPCR assay to pick up low copy number transcripts as the previous HERV-K3 RT-qPCR assay performed on the primary motor cortex, the ALS samples were preferentially selected for their high HERV-K3 counts based on RNA-Seq data (Bernardo, Ribeiro Pinto and Albano, 2013). Alternative approaches such as digital PCR that can detect small differences in copy number could be performed in future studies.

While expression of HERV-K3 locus in the premotor cortex of ALS patients compared to controls was not significant in the dataset we tested the ability to look at the whole transcriptome via RNA Sequencing showed a more advanced method for looking at HERV

expression (Prudencio *et al.*, 2015, 2017; Melnick *et al.*, 2021). Following the modified ERVMap (Tokuyama *et al.*, 2018b) bioinformatics pipeline used in Jones *et.al.* (2021), on ALS primary motor cortex samples, we decided to test this pipeline on the RNA seq data that Jones et al, 2021 had generated but only for samples in which we had premotor cortex brain tissue samples available Unfortunately, due to the smaller sample size in which RNA seq analysis was performed on we were not able to replicate Jones et al, 2021 findings, and this was also confirmed by our collaborators at KCL when the analysis was repeated. However, it allowed me to become familiar with running the ERVMap pipeline for RNA seq analysis on publicly available datasets.

The modified ERVMap protocol subsequently enabled the analysis of publicly available RNA-Seq datasets from multiple CNS regions. This facilitated the analysis of differentially expressed ERVs between ALS and non-ALS control samples in studies where the authors may have been initially looking purely at gene expression. For RNA-Seq analysis, two publicly available datasets were found with a wide enough selection of patient samples to perform an accurate assessment of HERV expression. RNA seq data was obtained from a research paper by Prudencio *et.al.* (2017) and a wider dataset obtained from the New York Genomic Centre (NYGC) (Baccarella *et al.*, 2018). These datasets represent 3 distinct regions of the brain, the frontal cortex and cerebellum from Prudencio et.al. (2017) and medial & lateral motor cortex regions from the New York Genomic Centre.

Both of these datasets revealed novel HERV loci differentially expressed in ALS vs Controls, though the Prudencio *et.al.* (2017) dataset provided an additional layer of analysis with a subset of ALS samples positive for a mutation in the C9orf72 region of the genome. The cerebellum and frontal cortex samples showed a mostly different list of significant HERVs when the additional samples were added, with a single HERV, HERV-H (10q23.31) showed to be significant in both datasets. Interestingly, while HERV-K22 (HML5) and HERV-H loci were significantly differentially regulated in the smaller sample set for cerebellum and frontal cortex tissue when reads aligning to the sex chromosome were removed the same was not the case for the larger dataset in which no HERV was recognised as being significantly differentially expressed.

The HERVs that were shown to be significant in the first analysis (when sex chromosome counts were present and removed from DESeq2 analysis) were then tested on our larger

cohort of n=91 sample set of ALS and non-ALS premotor cortex tissue samples to see if either HERV was significantly expressed in the tissue type. This analysis showed the first significant result of differentially expressed HERVs in this larger patient cohort as HERV-H env transcript was found to be differentially expressed (downregulated) by RT-qPCR when normalised against XPNEP1 reference genes. (Table 6.27). This also matched the results from the RNA-Seq frontal cortex analysis as this showed a downregulation of the locus in ALS vs Controls. The open reading frame analysis for this HERV-H loci only showed an identifiable reading frame in the RNaseH region however, though the protein was still shown to have a high similarity to the consensus sequence for HERV-H (Figure 6.9).

For the medial and lateral motor cortex analysis in the NYGC dataset a single HERV was found to be differentially regulated in each motor cortex region, with a different HERV-H locus (4p15.1) to the frontal cortex found to be downregulated in the lateral region of the motor cortex (Table 6.19). This would appear to also match our analysis of the premotor cortex sample set and the frontal cortex data, while the downregulation of the locus is slightly higher than the frontal cortex the primer set was designed to capture as many HERV-H loci as possible from the sequences obtained from RNA-Seq data. This additional significant result for HERV-H loci reveal that this is a promising novel biomarker for monitoring ALS disease as this particular HERV has only been seen to be differentially regulated in cancers prior to these analyses (Yi, Kim and Kim, 2006; Golan *et al.*, 2008; Toufaily *et al.*, 2011; Zhang, Liang and Zheng, 2019; Byun *et al.*, 2021; Manca *et al.*, 2022).

While the presence of these loci being confirmed to be differentially expressed in brain tissue regions provides positive evidence of differential HERV expression in ALS compared to controls a novel biomarker detectable in this region alone may only inform future therapy targets and not a potential method of minimally invasive monitoring of the disease. Thankfully peripheral blood mononuclear cells (PBMCs) provide a potential avenue for monitoring ALS biomarkers as this fraction of blood acts as an effective mirror of expression in other tissue types (Achiron and Gurevich, 2006; Mameli *et al.*, 2009; Rollins *et al.*, 2010; Balestrieri *et al.*, 2015; Zucca *et al.*, 2019; Arru *et al.*, 2021). Using data from a publicly available RNA-Seq dataset (Zucca *et al.*, 2019) and applying the modified ERVMap protocol used previously, we were able to identify a total of n=55 significant HERV loci differentially expressed in ALS vs Controls in this cell fraction (Table 7.1). While the latter part of the

analysis in this section primarily deals with HERV-K loci found to be differentially expressed in the sample set it should also be of note that there are several HERV-H loci which are shown to be differentially expressed in the analysis as well. While these HERV-H loci do not appear to match the loci found to be differentially expressed in the frontal cortex or lateral motor cortex tissue regions this does provide some evidence that HERV expression can be seen to be differentially regulated in PBMCs in ALS providing several novel biomarkers for future analysis.

8.3 Conclusion

The research into HERV expression in ALS has an important part to play if they are found to be implicated in ALS for purposes of a novel biomarker in monitoring the progression of the disease as well as the discovery of novel pathways contributing to the pathology of ALS. Current research into the effect of anti-retroviral drugs targeting HERV-K (HML2) expression in ALS patients has yielded some positive results in reducing the expression of these transcripts but any beneficial effect on those suffering from the disease has yet to be observed (Gold *et al.*, 2019; Garcia-Montojo *et al.*, 2021). While the expression of HML2 transcripts have been unable to be confirmed in studies past the initial research paper published by Li *et.al.* (2015) the study by Jones *et.al.* (2021) shows that other HERV-K family members may provide alternative biomarkers for diagnosis of ALS in which the life span is approximately 3-5yrs after initial diagnosis.

As this thesis has shown the differential expression of HERV transcripts is not limited to the HERV-K family in ALS. The research conducted in this thesis on publicly available RNA-Seq datasets has identified HERV-H as a potential avenue for study as a novel biomarker and disease determinant. While this represents a promising starting point for a new research path into HERV expression in the premotor cortex the exact reason why this HERV family is downregulated requires further investigation.

With the avenues for research into novel biomarkers in ALS continuing to be explored the hope is that the body of work reported in this thesis will give future researchers a good starting point for looking at novel transcripts like HERV-H in the brain and to undertake RNA seq analysis in other tissue types such as PBMCs, which is relatively non-invasive compared to obtaining biopsy tissue material. In addition, in this study we found a higher

number of HERVs differentially expressed in PBMCs compared to the CNS and brain and requires confirmation in larger datasets.

8.4 Future Work

The global supply chain crisis and pandemic resulting in laboratory closure have had a number of effects for all researchers since the onset of troubles at the start of 2020. With delays due to worker shortages due to COVID-19 infection and lack of stock to adequately keep up with demand of researchers for reagents and single use plastics, research activities have understandably slowed. With this in mind it can be understood that not all the avenues of potential study have been explored during this thesis and as outlined in chapter 7.0, cancellation of face-to-face clinics during the pandemic has prevented ALS patients attending clinics which has had a severe impact on receiving blood samples for PBMC separation so this part of the study could not be completed. However, due to the pandemic the research study had to switch to part analytical and this allowed me to learn new skills, such as how to run the ERVMap pipeline on publicly available RNA seq data as well as allow me to become familiar with running a collection of other bioinformatic tools

While HERV expression has been adequately explored in the premotor cortex, with exciting findings of a new HERV family being found to be significantly differentially expressed in this tissue type in ALS, analysis of potential HERVs at the protein level requires further investigation. The use of Immunohistochemistry techniques on FFPE tissue sections of the central nervous system would be able to provide evidence of HERV proteins being fully translated with the advantage of also being able to see where the HERV protein localises within the affected tissue type. As the current data coming out of the RNA-Seq analysis only provides predicted open reading frames for viral proteins having confirmation by Immunostaining and other techniques such as ELISA and Western Blotting would confirm whether these open reading frames encode functional proteins.

As the analysis of PBMCs as a potential source of a novel biomarker was only just touched upon with the help of publicly available RNA-Seq data, confirmation of these findings with RT-qPCR on a UK cohort of ALS patients and controls was unable to be performed due to sample availability as face-to-face clinics were closed during the pandemic when we had scheduled the research work to be undertaken. Future work would be to perform pre-

assay validation, confirming a suitable set of reference genes, testing primers for HERV-K3 3p21.31c and other HERV loci shown to be expressed in PBMCs that were revealed during the RNA-Seq analysis outlined in chapter 7.

This research study has provided useful data on potential areas for targeting further research, showing that HERVs can be differentially expressed in different anatomical regions of the brain and blood based on RT-qPCR and RNA seq analysis. If HERVs are found to be differentially expressed in ALS they have the potential to act as biomarker for disease diagnosis and target for treatment that is lacking to-date in the case of ALS.

Supplementary Data.

S1. Supplementary Information for Chapter 3.0

Figure S1. BLAST matches for nucleotide sequences obtained from Sanger Sequencing reactions of A) XPNPEP1, B) GAPDH, C) HERV-K *env*, D) HERV-K *gag*, E) HERV-K *pol*, F) HERV-K *RT* and G) HERV-W *env* amplicons generated by RT-qPCR assay using ALS sample A151/10.

Figure S2. BLAST matches for nucleotide sequences obtained from Sanger Sequencing reactions of A) XPNPEP1, B) GAPDH, C) HERV-K *env*, D) HERV-K *gag*, E) HERV-K *pol*, F) HERV-K *RT* and G) HERV-W *env* amplicons generated by RT-qPCR assay using non-ALS Control sample A292/09.

Figure S3. MEGA7 alignment for Li et.al. (2015) HERV-K gag Forward Primer, showing full-length HERV-K gag-pol-env sequences 1-46, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S4. MEGA7 alignment for Li et.al. (2015) HERV-K *gag* Forward Primer, showing full-length HERV-K gag-pol-env sequences 47-92, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S5. MEGA7 alignment for Li et.al. (2015) HERV-K *gag* Forward Primer, showing full-length HERV-K gag-pol-env sequences 93-116, and HERV-K *gag* fragments from 117-138 obtained from GenBank (NCBI, USA). Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S6. MEGA7 alignment for new HERV-K gag Forward Primer, showing full-length HERV-K gag-pol-env sequences 1-46, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S7. MEGA7 alignment for new HERV-K *gag* Forward Primer, showing full-length HERV-K gag-pol-env sequences 47-92, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S8. MEGA7 alignment for new HERV-K *gag* Forward Primer, showing full-length HERV-K gag-pol-env sequences 93-116, and HERV-K *gag* fragments from 117-138 obtained from GenBank (NCBI, USA). Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S9. MEGA7 alignment for Li et.al. (2015) HERV-K gag Reverse Primer, showing full-length HERV-K gag-pol-env sequences 1-46, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S10. MEGA7 alignment for Li et.al. (2015) HERV-K *gag* Reverse Primer, showing full-length HERV-K gag-pol-env sequences 47-92, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S11. MEGA7 alignment for Li et.al. (2015) HERV-K *gag* Reverse Primer, showing full-length HERV-K gag-pol-env sequences 93-116, and HERV-K *gag* fragments from 117-138 obtained from GenBank (NCBI, USA). Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S12. MEGA7 alignment for new HERV-K gag Reverse Primer, showing full-length HERV-K gag-pol-env sequences 1-46, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S13. MEGA7 alignment for new HERV-K *gag* Reverse Primer, showing full-length HERV-K gag-pol-env sequences 47-92, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S14. MEGA7 alignment for new HERV-K *gag* Reverse Primer, showing full-length HERV-K gag-pol-env sequences 93-116, and HERV-K *gag* fragments from 117-138 obtained from GenBank (NCBI, USA). Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S15. MEGA7 alignment for Li et.al. (2015) HERV-K *pol* Forward Primer, showing full-length HERV-K gag-pol-env sequences 1-46, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

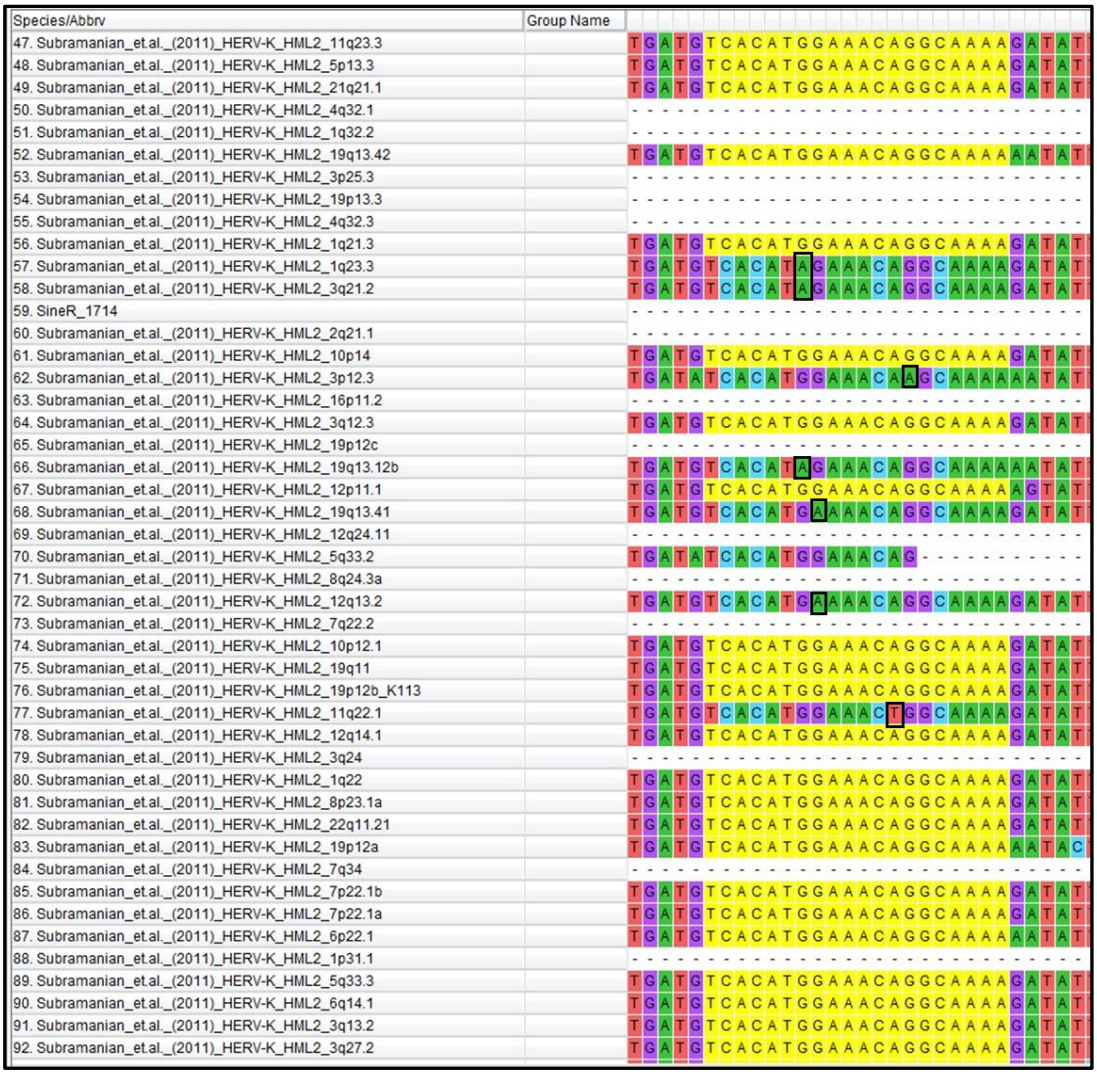


Figure S16. MEGA7 alignment for Li et.al. (2015) HERV-K *pol* Forward Primer, showing full-length HERV-K gag-pol-env sequences 47-92, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

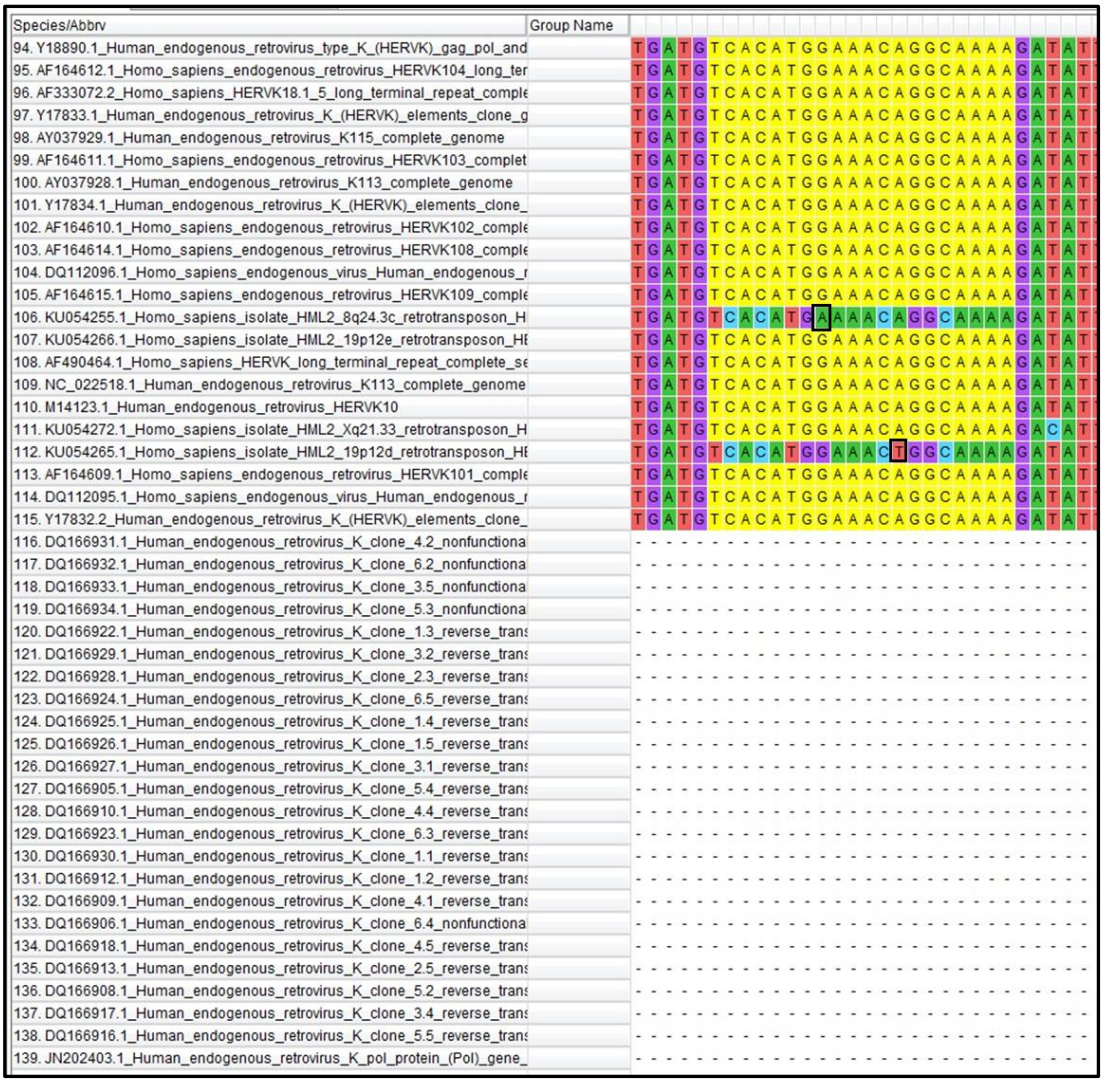


Figure S17. MEGA7 alignment for Li et.al. (2015) HERV-K *pol* Forward Primer, showing full-length HERV-K gag-pol-env sequences 93-116, and HERV-K *pol* fragments from 117-138 obtained from GenBank (NCBI, USA). Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S18. MEGA7 alignment for new HERV-K *RT* Forward Primer, showing full-length HERV-K gag-pol-env sequences 1-46, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S19. MEGA7 alignment for new HERV-K *RT* Forward Primer, showing full-length HERV-K gag-pol-env sequences 47-92, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

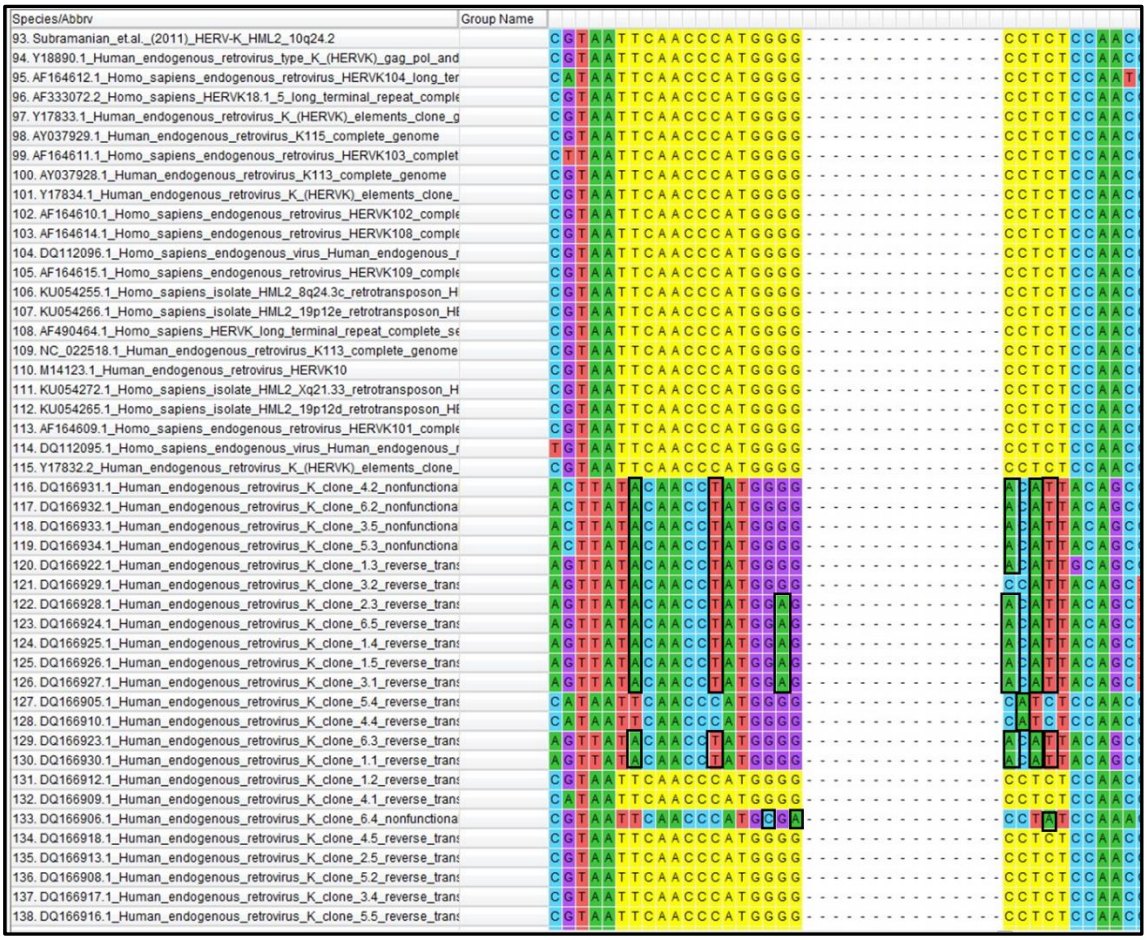


Figure S20. MEGA7 alignment for new HERV-K *RT* Forward Primer, showing full-length HERV-K gag-pol-env sequences 93-116, and HERV-K *pol* fragments from 117-138 obtained from GenBank (NCBI, USA). Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

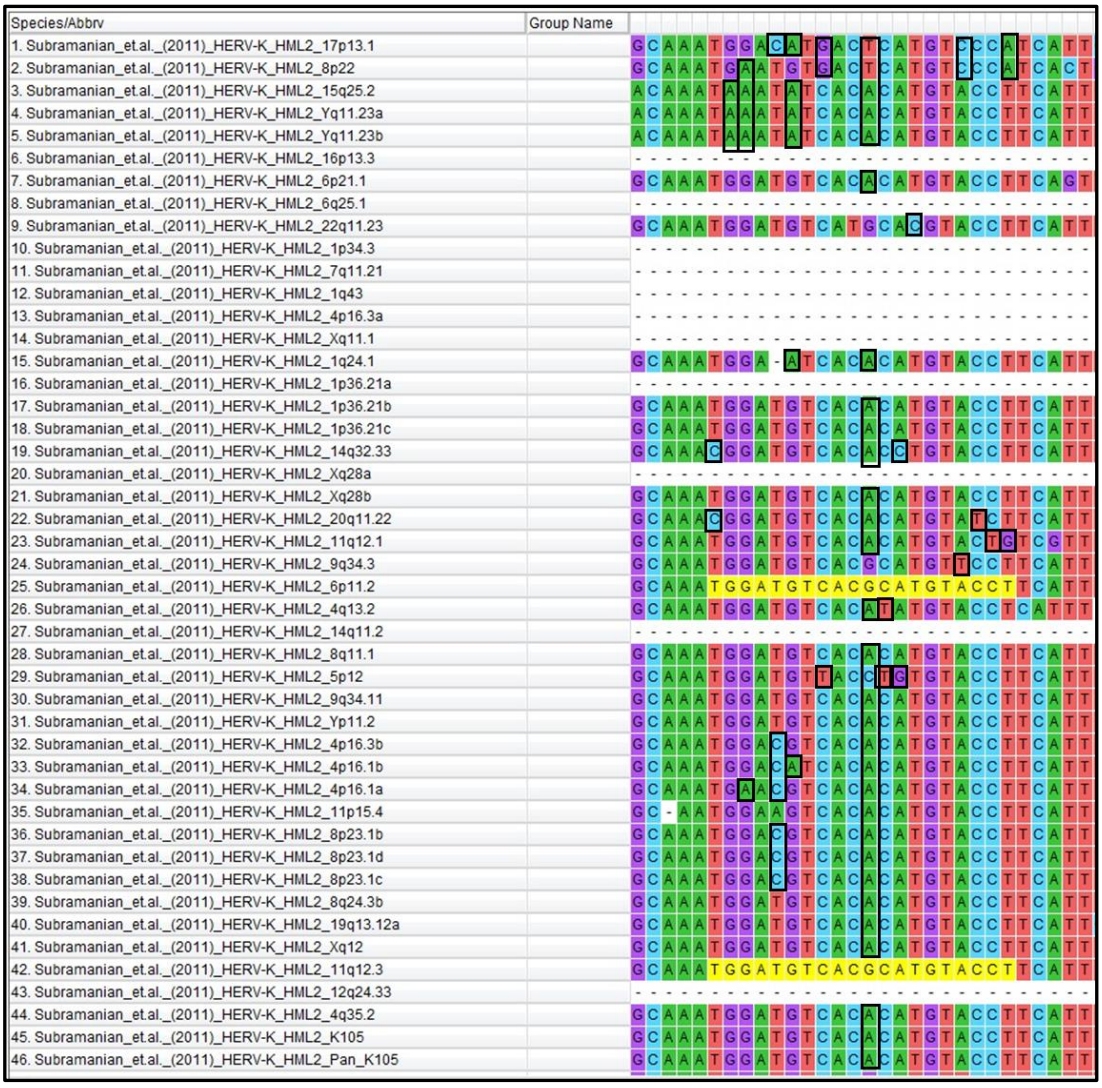


Figure S21. MEGA7 alignment for Li et.al. (2015) HERV-K *pol* Reverse Primer, showing full-length HERV-K gag-pol-env sequences 1-46, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

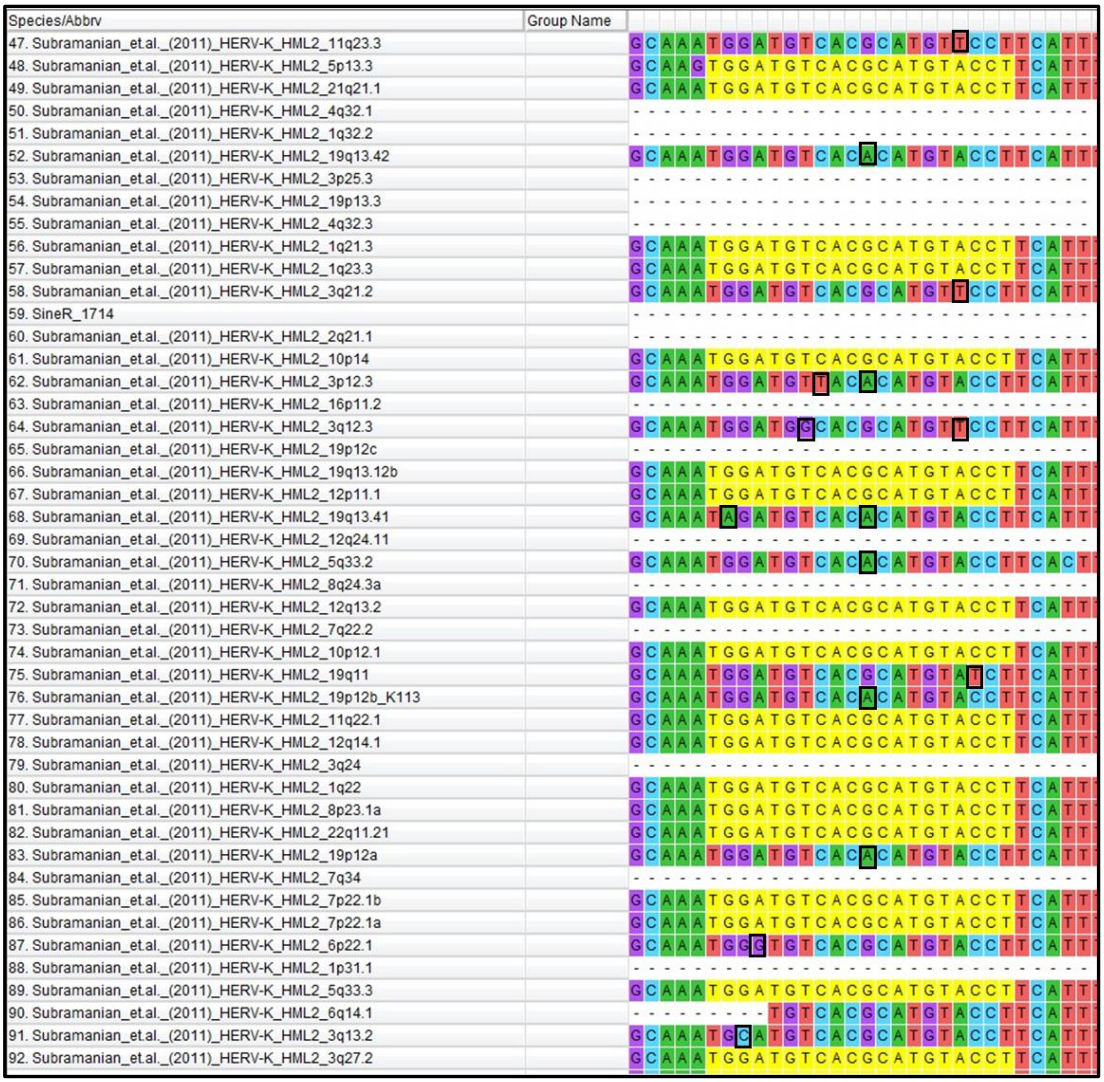


Figure S22. MEGA7 alignment for Li et.al. (2015) HERV-K *pol* Reverse Primer, showing full-length HERV-K gag-pol-env sequences 47-92, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

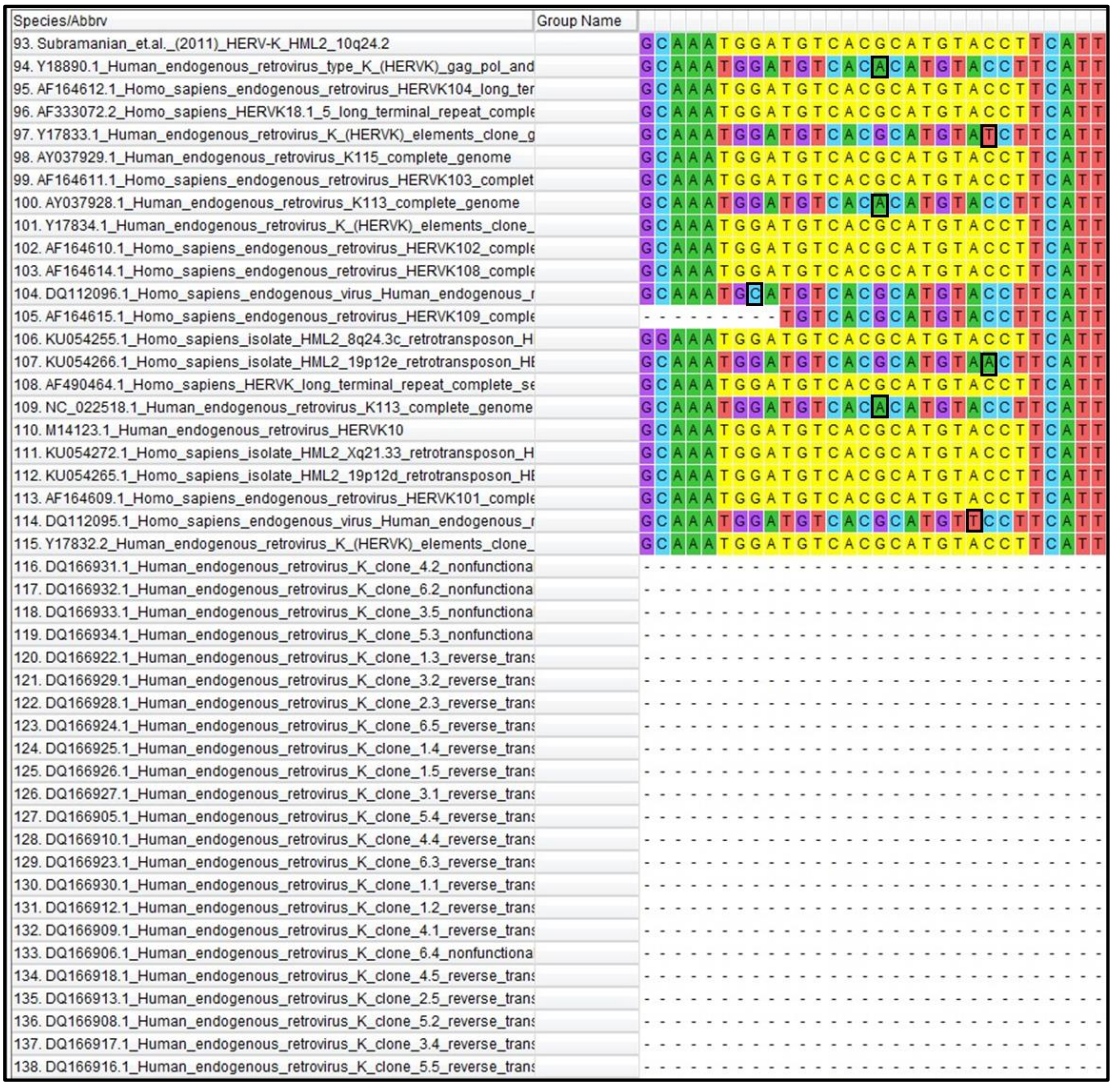


Figure S23. MEGA7 alignment for Li et.al. (2015) HERV-K *pol* Reverse Primer, showing full-length HERV-K gag-pol-env sequences 93-116, and HERV-K *pol* fragments from 117-138 obtained from GenBank (NCBI, USA). Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S24. MEGA7 alignment for new HERV-K *RT* Reverse Primer, showing full-length HERV-K gag-pol-env sequences 1-46, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S25. MEGA7 alignment for new HERV-K *RT* Reverse Primer, showing full-length HERV-K gag-pol-env sequences 47-92, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S26. MEGA7 alignment for new HERV-K *RT* Reverse Primer, showing full-length HERV-K gag-pol-env sequences 93-116, and HERV-K *pol* fragments from 117-138 obtained from GenBank (NCBI, USA). Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S27. MEGA7 alignment for Li et.al. (2015) HERV-K *env* Forward Primer, showing full-length HERV-K gag-pol-env sequences 1-46, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S28. MEGA7 alignment for Li et.al. (2015) HERV-K *env* Forward Primer, showing full-length HERV-K gag-pol-env sequences 47-92, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

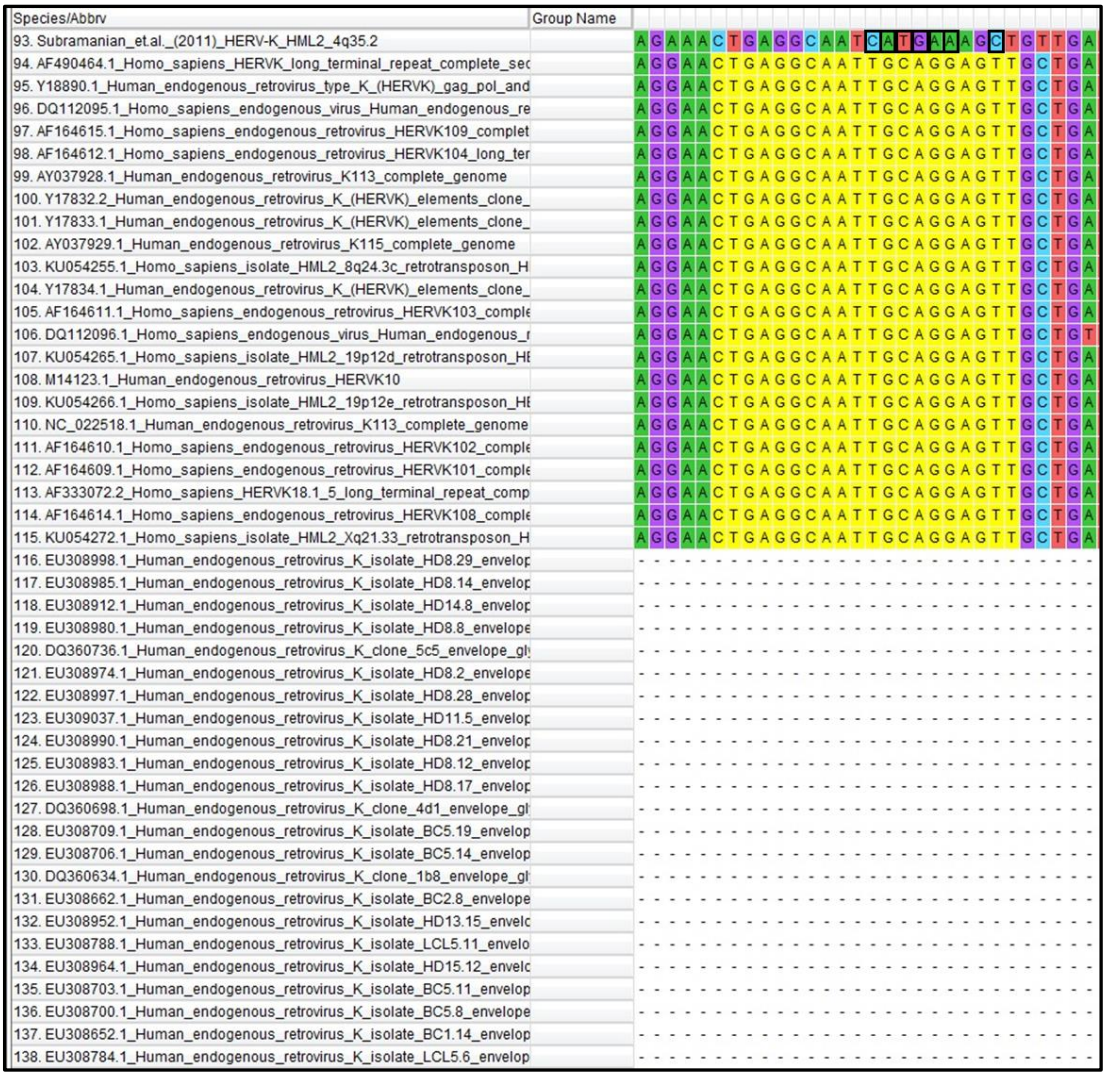


Figure S29. MEGA7 alignment for Li et.al. (2015) HERV-K *env* Forward Primer, showing full-length HERV-K gag-pol-env sequences 93-116, and HERV-K *env* fragments from 117-138 obtained from GenBank (NCBI, USA). Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

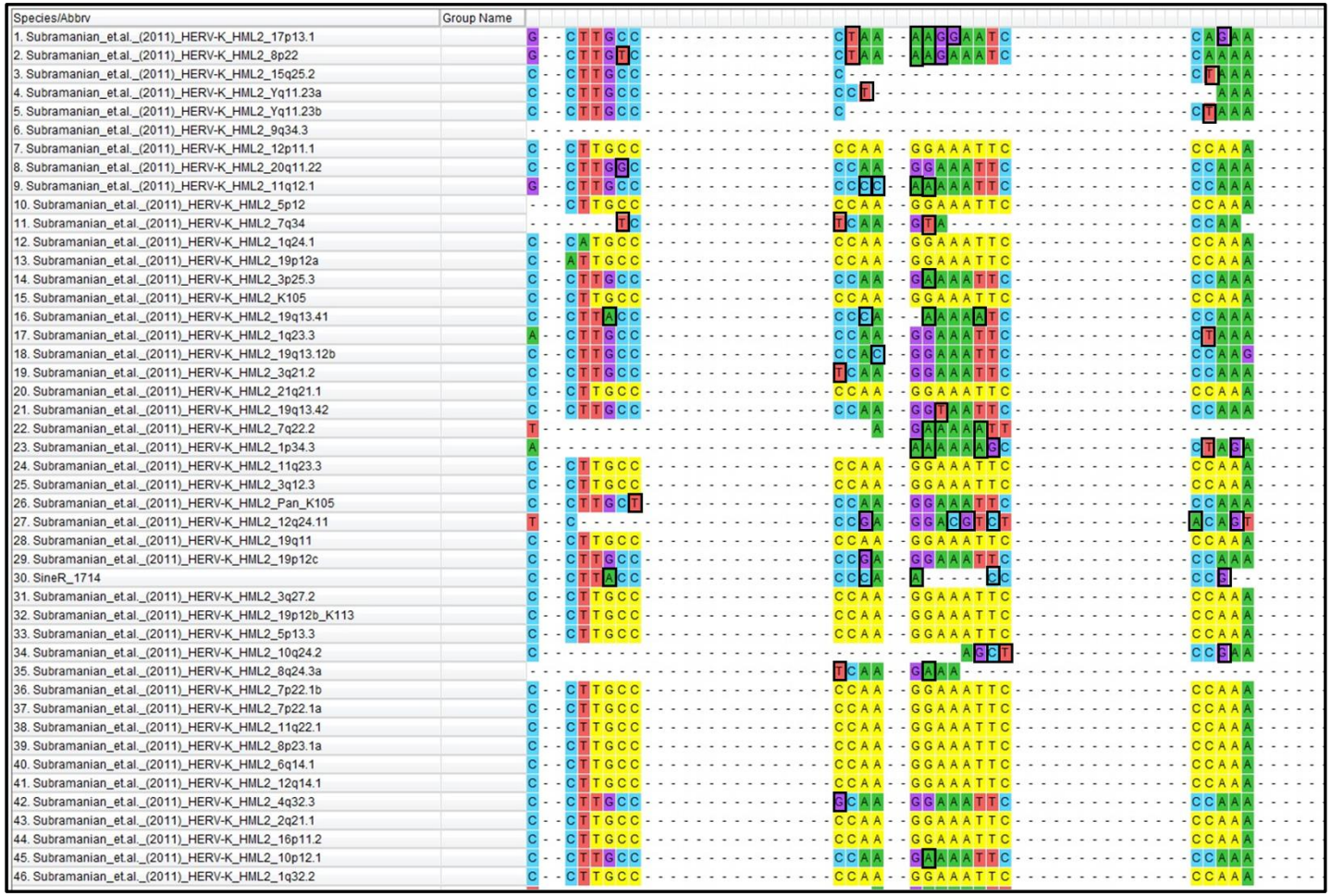


Figure S30. MEGA7 alignment for new HERV-K *env* Forward Primer, showing full-length HERV-K gag-pol-env sequences 1-46, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

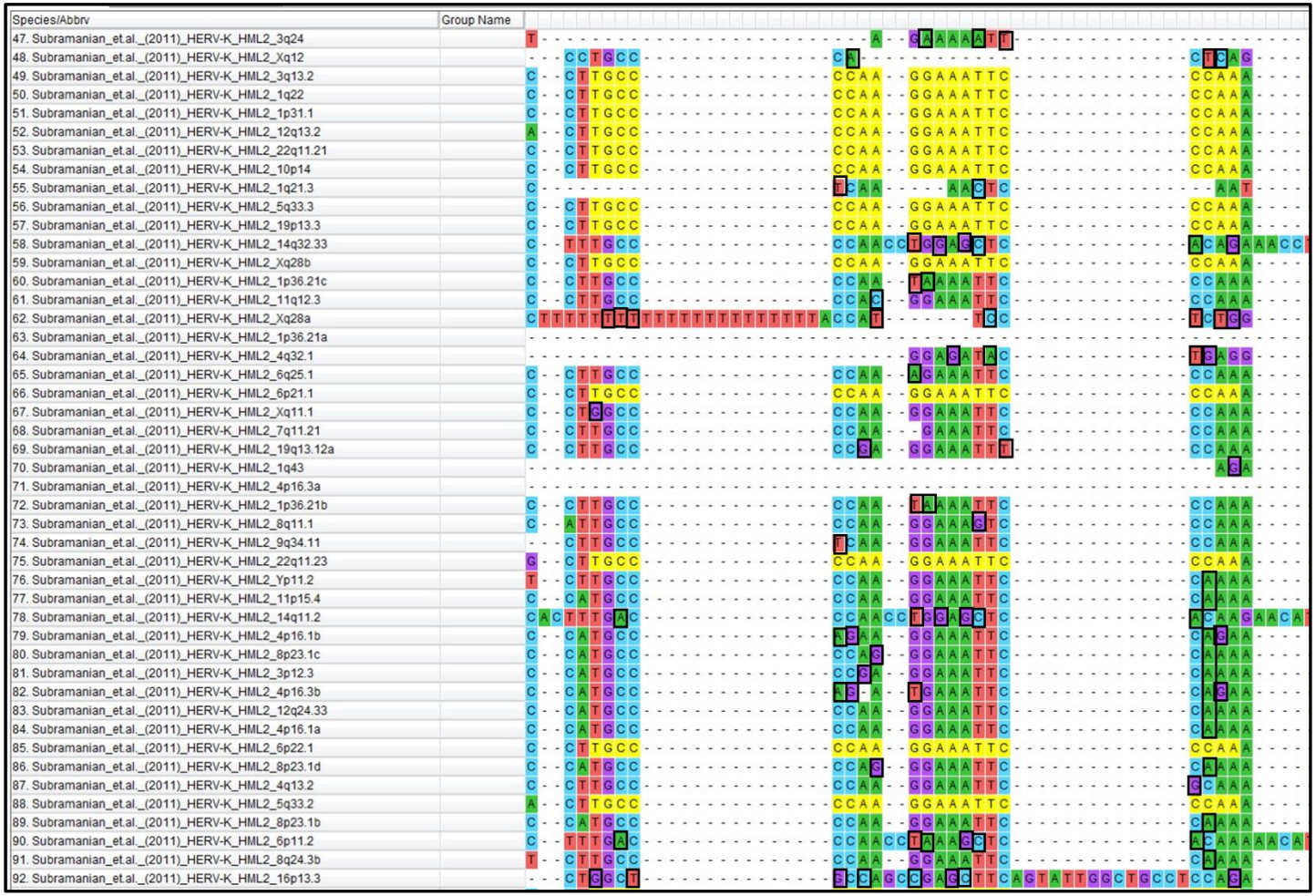


Figure S31. MEGA7 alignment for new HERV-K *env* Forward Primer, showing full-length HERV-K gag-pol-env sequences 47-92, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S32. MEGA7 alignment for new HERV-K *env* Forward Primer, showing full-length HERV-K gag-pol-env sequences 93-116, and HERV-K *env* fragments from 117-138 obtained from GenBank (NCBI, USA). Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

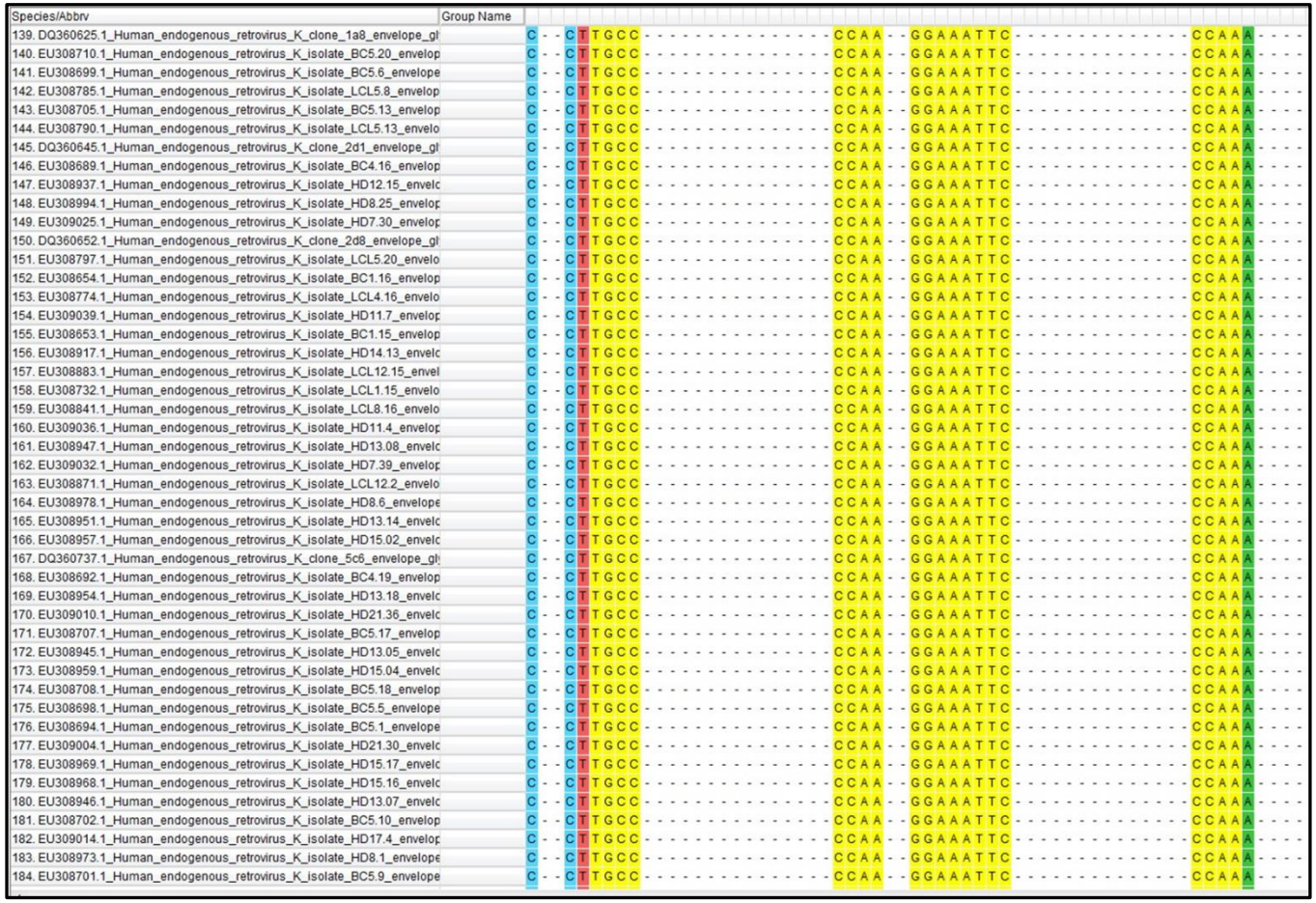


Figure S33. MEGA7 alignment for new HERV-K *env* Forward Primer, showing HERV-K *env* fragment sequences 139-184, obtained from GenBank (NCBI, USA). This is a representative image for the following 120 sequences. Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S34. MEGA7 alignment for Li et.al. (2015) HERV-K *env* Reverse Primer, showing full-length HERV-K gag-pol-env sequences 1-46, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S35. MEGA7 alignment for Li et.al. (2015) HERV-K *env* Reverse Primer, showing full-length HERV-K gag-pol-env sequences 47-92, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

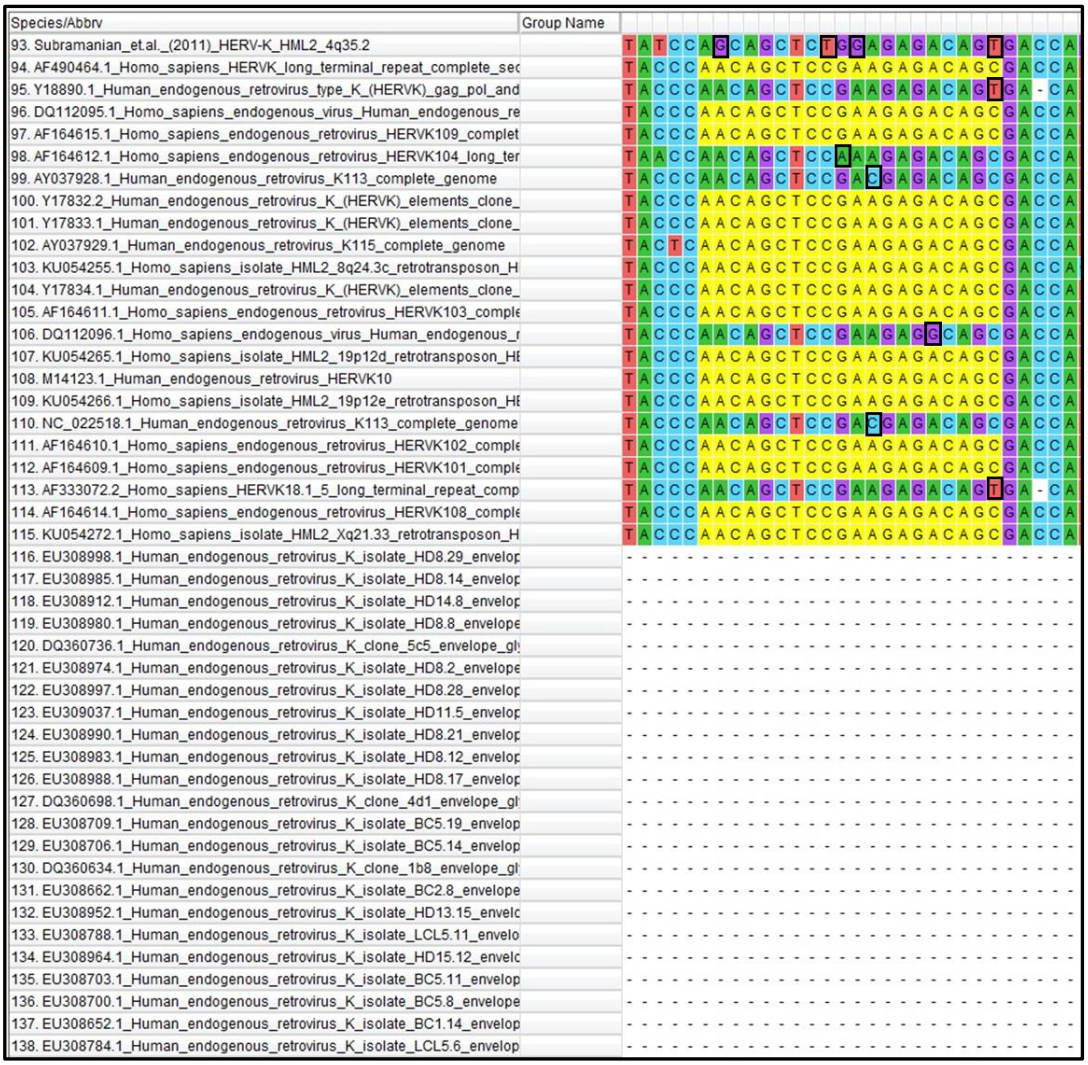


Figure S36. MEGA7 alignment for Li et.al. (2015) HERV-K *env* Reverse Primer, showing full-length HERV-K gag-pol-env sequences 93-116, and HERV-K *env* fragments from 117-138 obtained from GenBank (NCBI, USA). Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

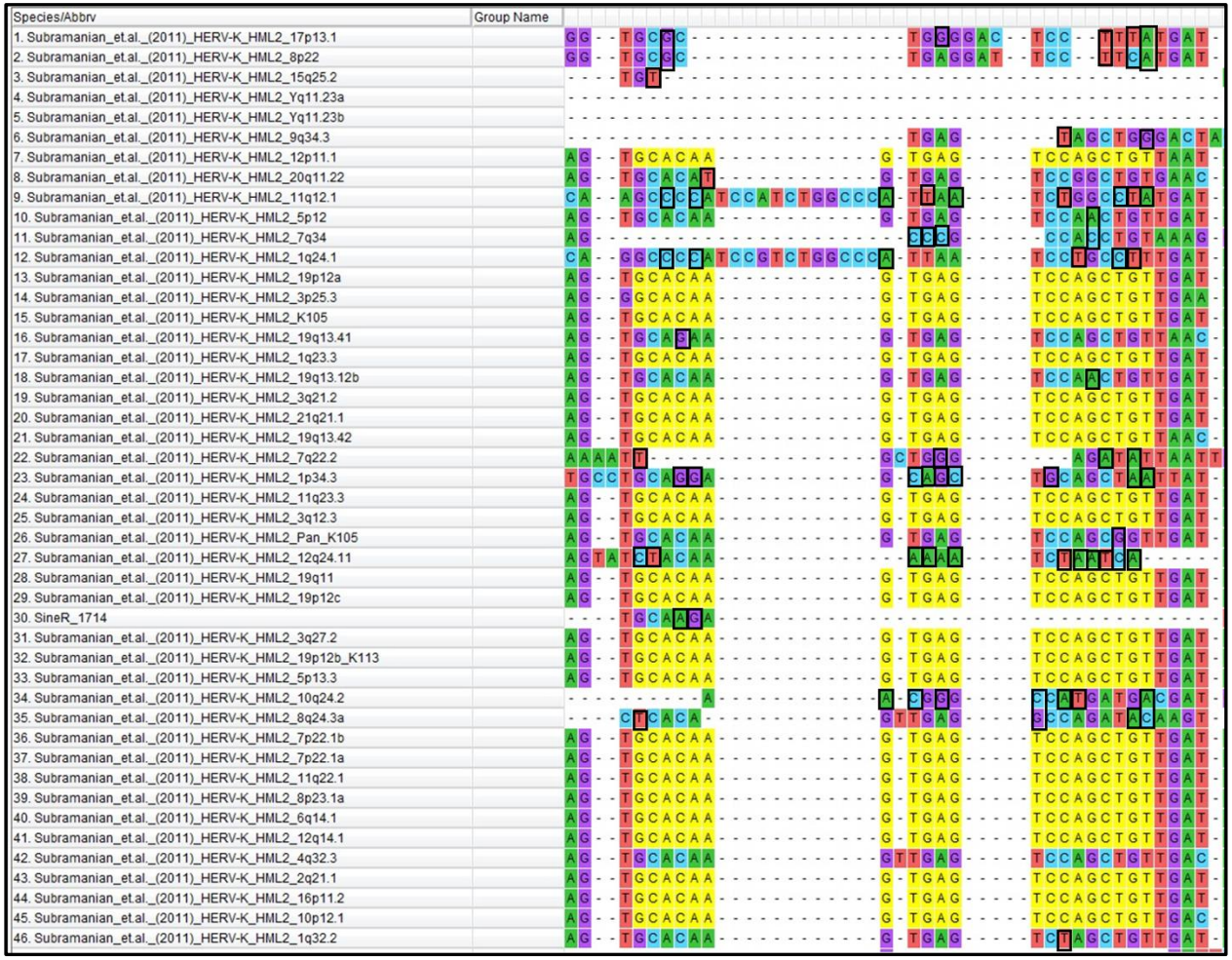


Figure S37. MEGA7 alignment for new HERV-K *env* Reverse Primer, showing full-length HERV-K gag-pol-env sequences 1-46, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S38. MEGA7 alignment for new HERV-K *env* Reverse Primer, showing full-length HERV-K gag-pol-env sequences 47-92, obtained from Subramanian et.al. (2011) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

Figure S39. MEGA7 alignment for Li et.al. (2015) HERV-K *env* Reverse Primer, showing full-length HERV-K gag-pol-env sequences 93-116, and HERV-K *env* fragments from 117-138 obtained from GenBank (NCBI, USA) Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

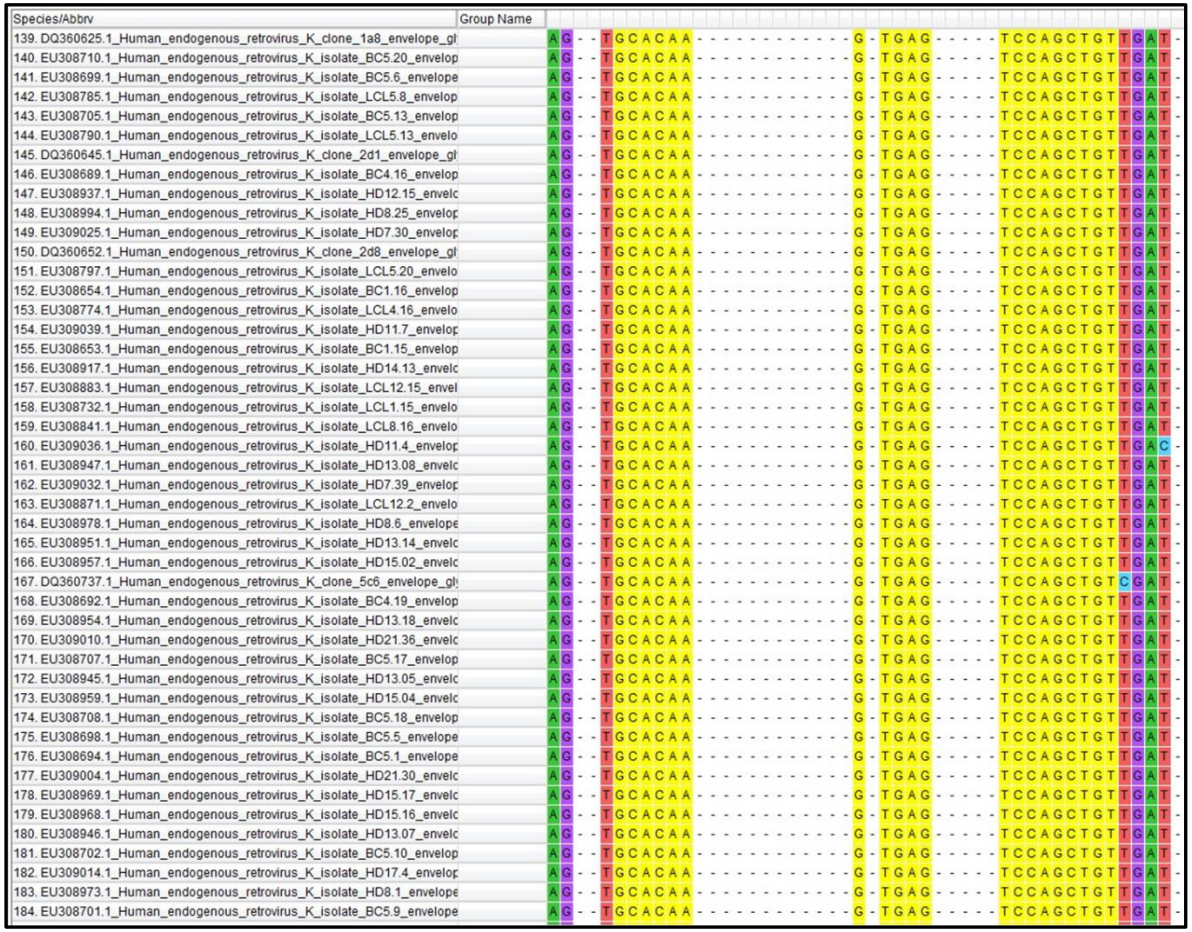


Figure S40. MEGA7 alignment for new HERV-K *env* Reverse Primer, showing HERV-K *env* fragment sequences 139-184, obtained from GenBank (NCBI, USA). This is a representative image for the following 120 sequences. Yellow highlighted regions represent exact matches to the primer sequence with black boxes indicating base pair differences to the primer.

S2. Supplementary Information for Chapter 4.0

Supplementary Table S1. Mean Ct information for n=5 ALS and n=5 non-ALS Control Samples used in geNorm Analysis.

		Candidate Reference Genes								
Sample ID	Disease Status	RPL13A	UBC	YWHAZ	EIF4A2	CYC1	SDHA	GAPDH	XPNPEP1	β-Actin
A292/09	Control	21.265	20.276	26.449	22.246	21.820	23.268	17.447	24.920	17.864
A265/08	Control	22.100	20.730	27.465	24.155	22.477	23.623	17.658	25.611	18.894
A012/12	Control	23.180	21.704	30.808	26.111	24.373	24.899	19.346	27.315	21.271
A346/10	Control	22.490	20.318	28.964	24.939	23.369	24.289	18.261	26.045	19.050
A273/12	Control	22.893	22.139	31.433	25.862	24.301	25.624	20.104	26.979	20.739
A151/10	ALS	21.341	19.179	26.522	21.894	21.699	22.849	17.061	24.562	17.650
A203/11	ALS	22.819	20.649	27.907	24.948	23.485	23.658	18.247	25.958	19.618
A401/08	ALS	21.640	19.478	26.279	22.040	21.836	22.784	17.207	24.457	17.778
A205/09	ALS	21.369	20.417	26.856	21.737	21.351	22.902	17.137	24.543	18.062
A115/08	ALS	23.481	21.563	28.248	25.038	23.458	23.845	18.570	26.195	19.477

Supplementary Table S2. Relevant clinical and RNA Integrity information for tissue samples used in the analysis of HERV-K differential expression in ALS and non-ALS controls.

Sample ID	Sex	RIN	Disease Status	Postmortem Delay (Hours)	Age at Time of Death (Years)
A375/12	Male	6.60	ALS	20.5	61
A073/12	Female	7.20	ALS	29	68
A342/13	Female	7.70	ALS	56	50
A254/15	Female	6.20	ALS	51.5	68
A355/15	Male	5.40	ALS	52.5	69
A151/10	Female	7.80	ALS	38	75
A115/08	Male	5.90	ALS	42	83
A414/12	Female	6.10	ALS	53	72
A203/11	Female	5.80	ALS	55	57
A447/13	Female	5.00	ALS	34	90
A381/11	Female	6.60	ALS	77	86
A251/09	Male	6.20	ALS	2:30	78
A205/09	Male	7.20	ALS	35	58
A162/09	Female	6.80	ALS	27.5	70
A041/07	Female	6.50	ALS	21.5	87
A308/15	Male	7.00	ALS	51	76
A401/08	Female	7.00	ALS	36.5	80
A348/08	Female	4.90	ALS	64	69
A218/09	Female	6.70	ALS	3.7	65
A292/09	Female	6.90	Control	43	43
A261/12	Male	5.80	Control	23	63
A132/14	Female	5.80	Control	50	66
A308/09	Male	4.90	Control	52	66
A346/10	Female	5.90	Control	34	84
A265/08	Male	6.60	Control	47	79
A012/12	Female	4.90	Control	33	51
A358/08	Female	5.40	Control	12	55
A033/11	Male	6.00	Control	47	82
A002/13	Male	6.60	Control	45	90
A007/15	Female	5.80	Control	66	74
A407/13	Female	6.60	Control	22	80
A153/06	Female	6.60	Control	17	92
A248/11	Female	6.80	Control	53	84
A177/14	Female	6.70	Control	62	67
A158/14	Female	4.10	Control	27	73
A273/12	Male	5.60	Control	25	67
A308/14	Female	4.40	Control	78	66
A319/14	Female	4.10	Control	44	90
A103/17	Female	4.30	Control	71	55

Supplementary Table S3. Ct Means for Reference Genes, HERV-K (*gag, pol, env* & *RT*) transcripts and HERV-W *env*

Sample ID	GAPDH	XPNPEP1	HERV-K *gag*	HERV-K *pol*	HERV-K *env*	HERV-K *RT*	HERV-W *env*
A375/12	16.545	23.681	24.983	24.354	23.275	23.400	25.602
A073/12	16.205	23.243	24.498	23.737	22.860	22.829	24.920
A342/13	15.855	22.770	23.852	23.640	22.374	22.569	24.805
A254/15	17.194	24.657	24.973	24.764	23.918	23.639	25.829
A355/15	17.881	25.454	25.890	25.818	24.732	24.621	26.926
A151/10	16.059	23.069	24.030	23.664	22.649	22.569	24.812
A115/08	17.155	24.395	25.032	24.712	23.780	23.645	25.392
A414/12	16.960	24.406	25.243	24.689	23.808	23.777	25.689
A203/11	17.753	25.517	25.888	25.779	24.716	24.182	26.895
A447/13	18.686	25.772	26.492	25.466	24.605	24.774	26.902
A381/11	16.950	23.879	23.100	23.557	22.177	22.381	25.590
A251/09	17.023	24.237	24.960	23.914	23.022	22.918	25.692
A205/09	16.325	23.370	24.395	23.862	23.256	22.747	25.437
A162/09	16.234	23.099	24.642	23.818	22.974	22.707	25.264
A041/07	16.741	23.782	25.059	24.539	23.519	23.428	25.484
A308/15	15.969	23.253	24.620	23.713	22.724	22.653	25.647
A401/08	16.180	22.967	24.091	23.729	22.684	22.376	25.348
A348/08	18.673	26.340	26.800	26.526	25.846	25.147	27.424
A218/09	16.917	23.837	24.519	23.907	23.016	22.925	25.622
A292/09	16.339	23.821	24.533	23.963	23.103	23.145	25.250
A261/12	17.394	24.813	25.436	25.077	24.344	24.011	26.095
A132/14	17.156	24.981	25.509	25.256	24.042	23.926	26.020
A308/09	18.150	26.435	26.993	27.737	26.270	26.817	27.519
A346/10	17.222	24.803	25.300	25.244	24.552	24.075	26.161
A265/08	17.719	25.484	26.030	25.306	24.223	24.601	26.991
A012/12	18.900	26.412	26.877	26.218	25.298	25.343	26.830
A358/08	18.032	25.583	25.779	25.161	24.334	24.194	26.165
A033/11	17.487	24.904	25.273	25.180	23.986	23.990	25.868
A002/13	16.327	23.744	24.679	24.218	23.153	23.419	25.448
A007/15	16.676	23.733	24.591	24.317	23.069	23.112	25.665
A407/13	17.055	23.864	24.959	24.151	23.372	23.107	25.601
A153/06	17.118	24.345	25.313	24.657	23.692	23.587	30.743
A248/11	16.608	23.685	24.774	23.653	22.560	22.841	25.526
A177/14	16.757	23.491	24.399	24.070	23.254	22.909	25.225
A158/14	19.695	27.468	27.562	27.506	26.201	26.288	27.880
A273/12	18.125	25.080	25.664	25.347	24.615	24.241	26.606
A308/14	18.683	26.875	28.495	27.189	25.907	26.500	27.963
A319/14	20.070	27.202	28.793	28.082	26.991	27.231	29.335
A103/17	18.388	25.863	26.427	26.746	25.737	25.041	27.492

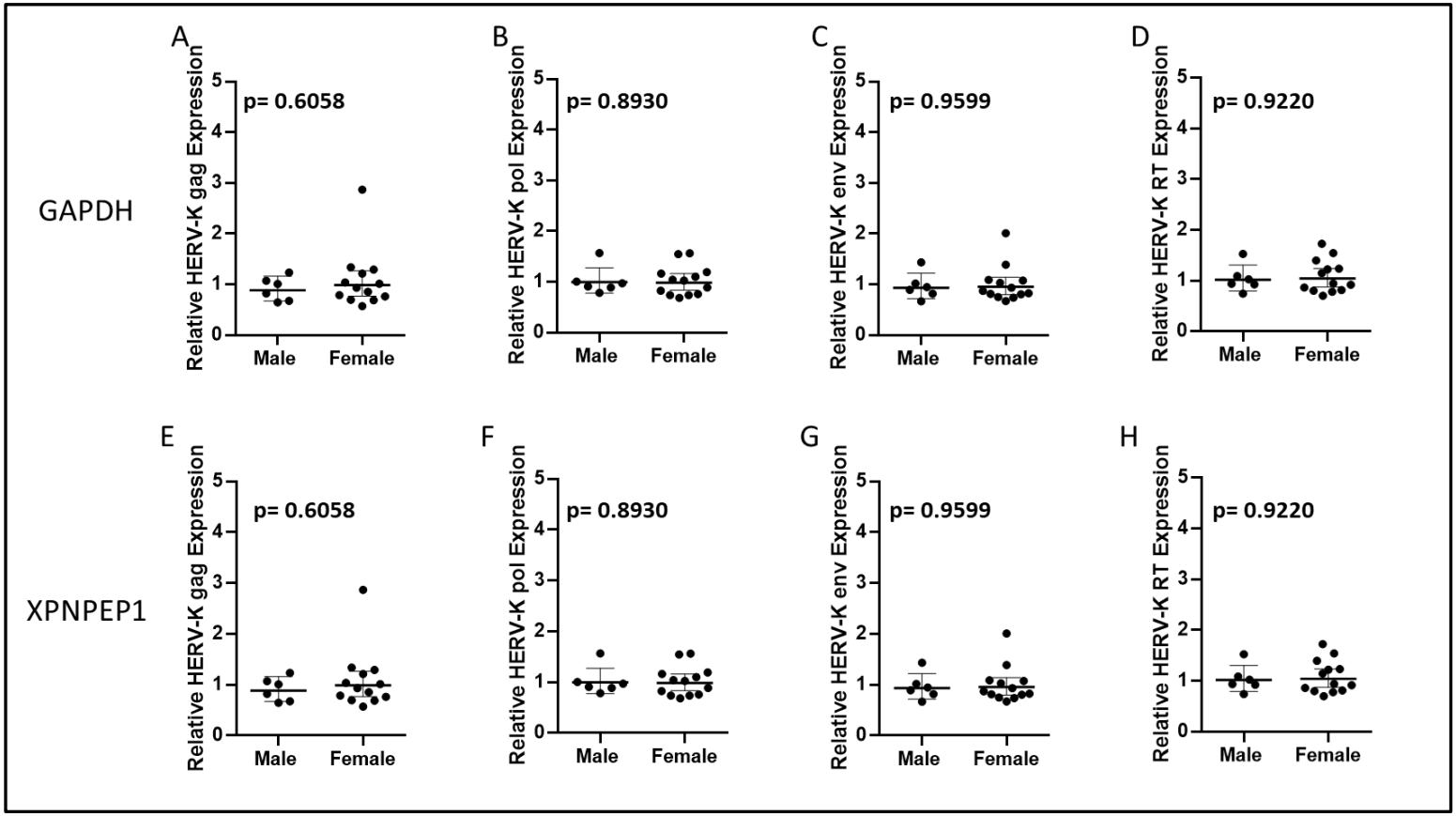


Figure S41. No significant correlation between gender and HERV-K RNA expression in Postmortem Premotor Cortex Tissue from ALS Patients
There is no significant difference in HERV-K transcript expression between male and female groups when the data is normalised to either GAPDH (A, B, C and D). or when normalised to XPNPEP1 (E, F, G, and H). This graph utilises data from ΔΔCt Normalisation method.

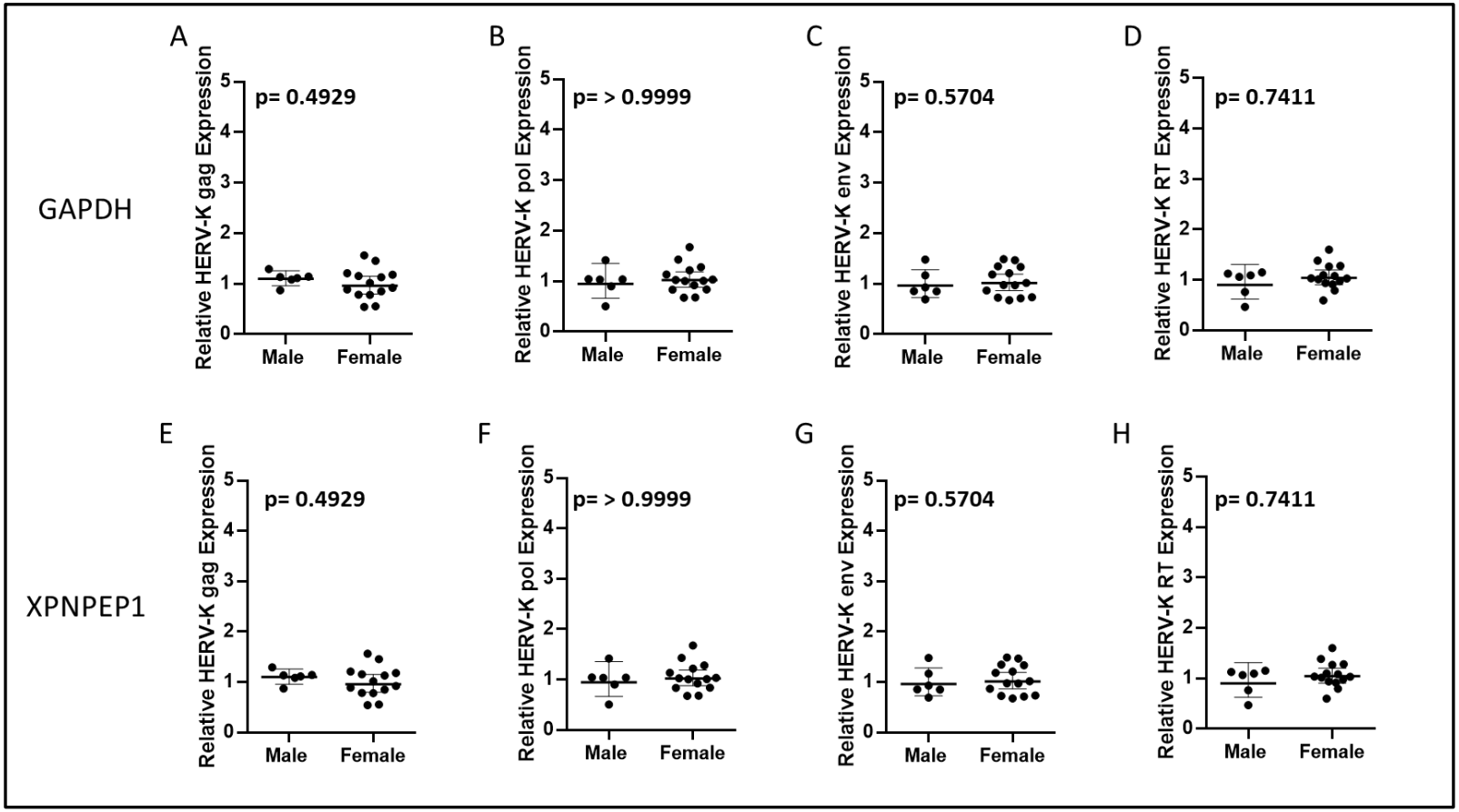


Figure S42. No significant correlation between gender and HERV-K RNA expression in Postmortem Premotor Cortex Tissue from non-ALS control Patients

There is no significant difference in HERV-K transcript expression between male and female groups when the data is normalised to either GAPDH (A, B, C and D). or when normalised to XPNPEP1 (E, F, G, and H). This graph utilises data from ΔΔCt Normalisation method.

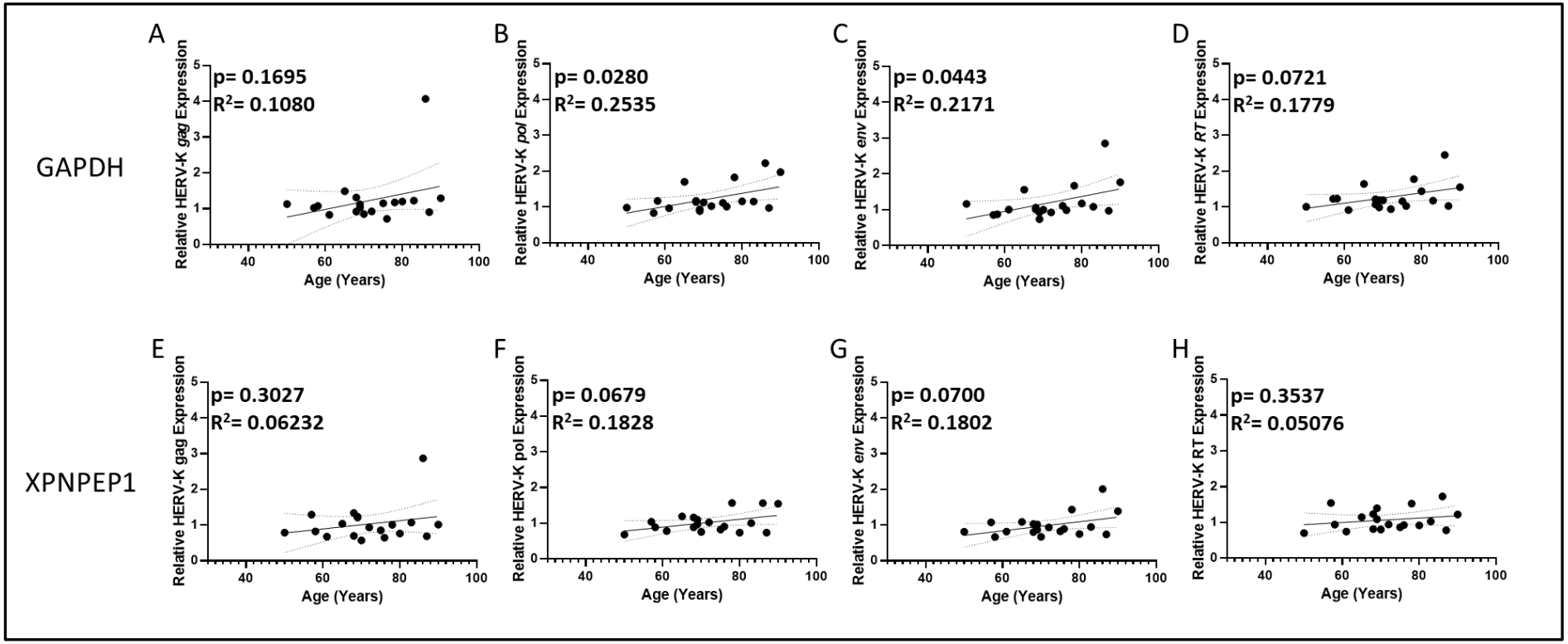


Figure S43. Effect of increasing age of patient at time of death on HERV-K Transcript expression in ALS samples.

Data shown in the figure above were normalised against 2 separate reference genes, with A, B, C and D normalised against GAPDH and E, F, G and H normalised against XPNPEP1. Normalisation against different reference genes had no effect on results R^2 values were calculated in GraphPad Prism v8.0. This graph utilises data from ΔΔCt Normalisation method.

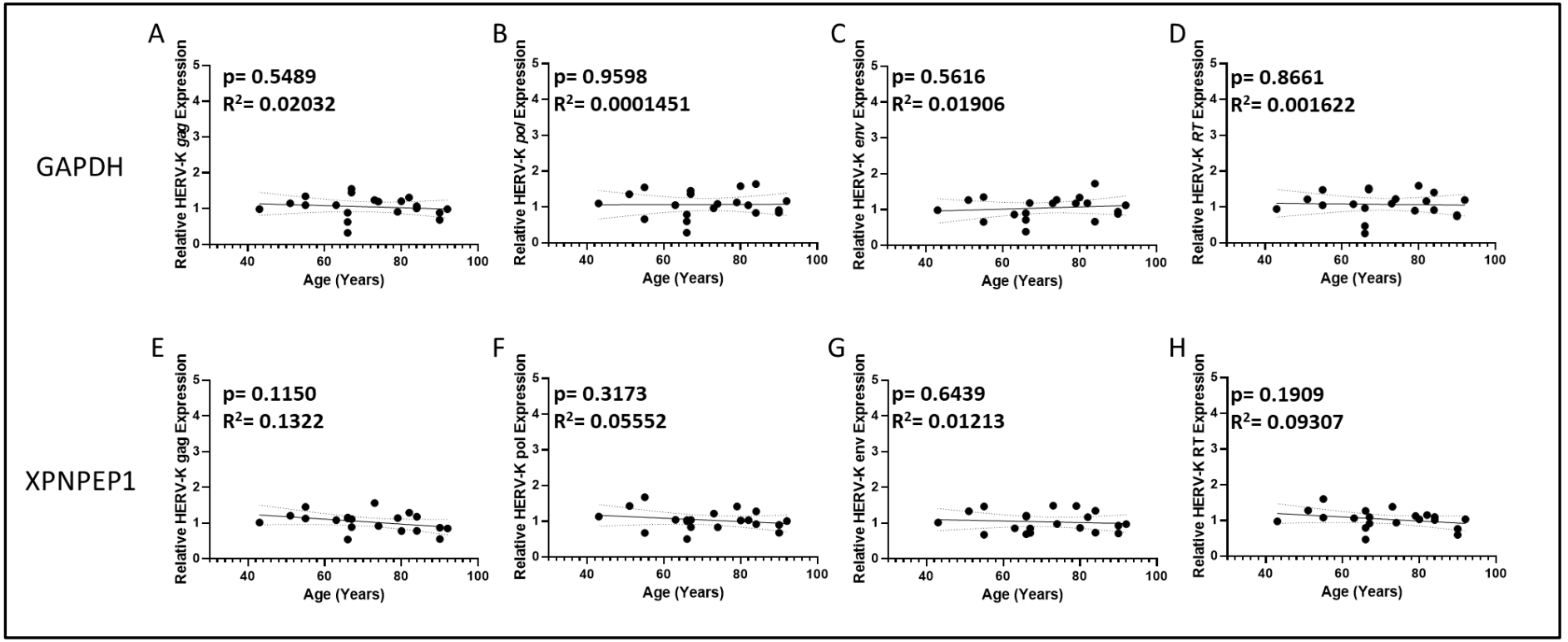


Figure S44. Effect of increasing age of patient at time of death on HERV-K Transcript expression in Non-ALS Control samples.

Data shown in the figure above were normalised against 2 separate reference genes, with A, B, C and D normalised against GAPDH and E, F, G and H normalised against XPNPEP1. Normalisation against different reference genes had no effect on results R^2 values were calculated in GraphPad Prism v8.0. This graph utilises data from ΔΔCt Normalisation method.

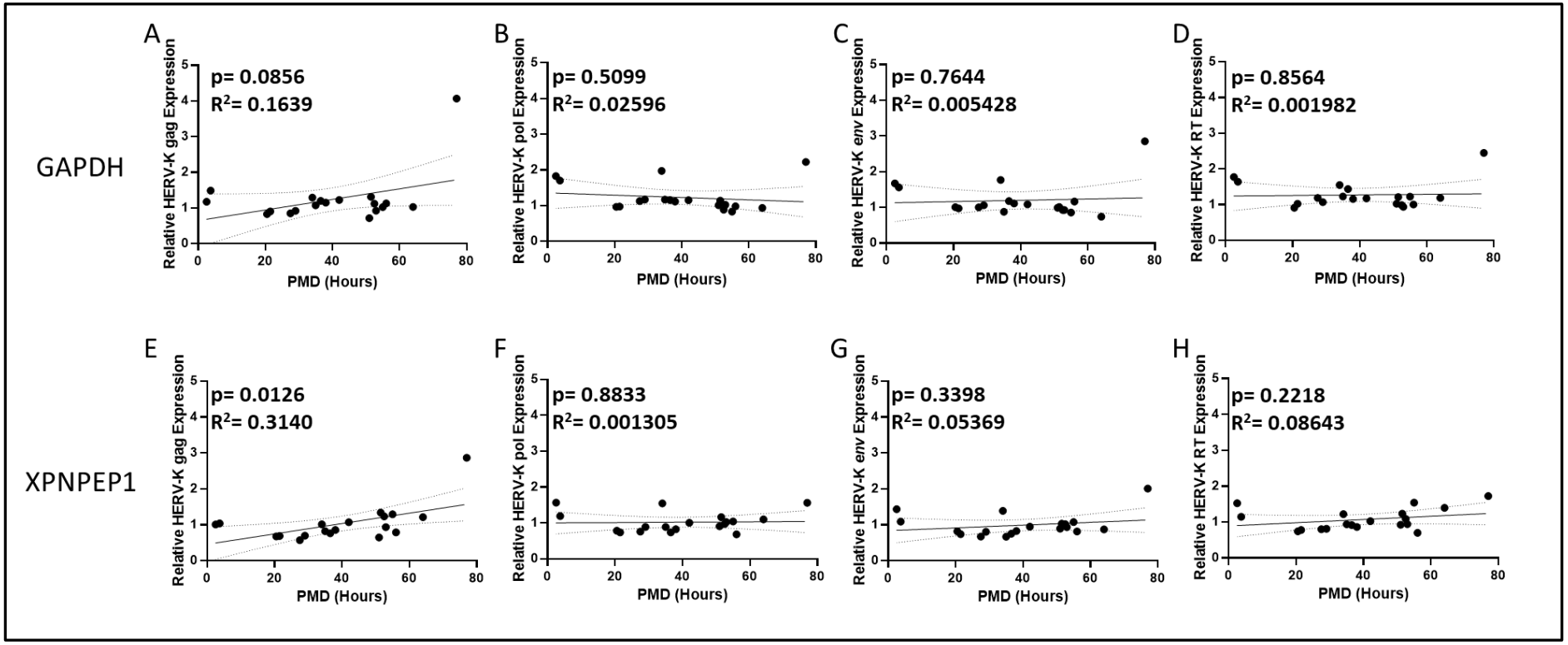


Figure S45. Effect of PMD on HERV-K transcript levels when normalised to GAPDH and XPNPEP1 in ALS Patient Tissue.

The figure above shows the effect of PMD on HERV-K expression for *gag, pol, env* and *RT* genomic regions. A, B, C and D are normalised to GAPDH which shows no or only very slight effect of PMD on HERV-K pol expression increasing PMD (B). There was no correlation to HERV-K expression when the data was normalised to XPNPEP1 (E, F, G and H). R^2 values and p values were calculated in GraphPad Prism v8.0. This graph utilises data from ΔΔCt Normalisation method.

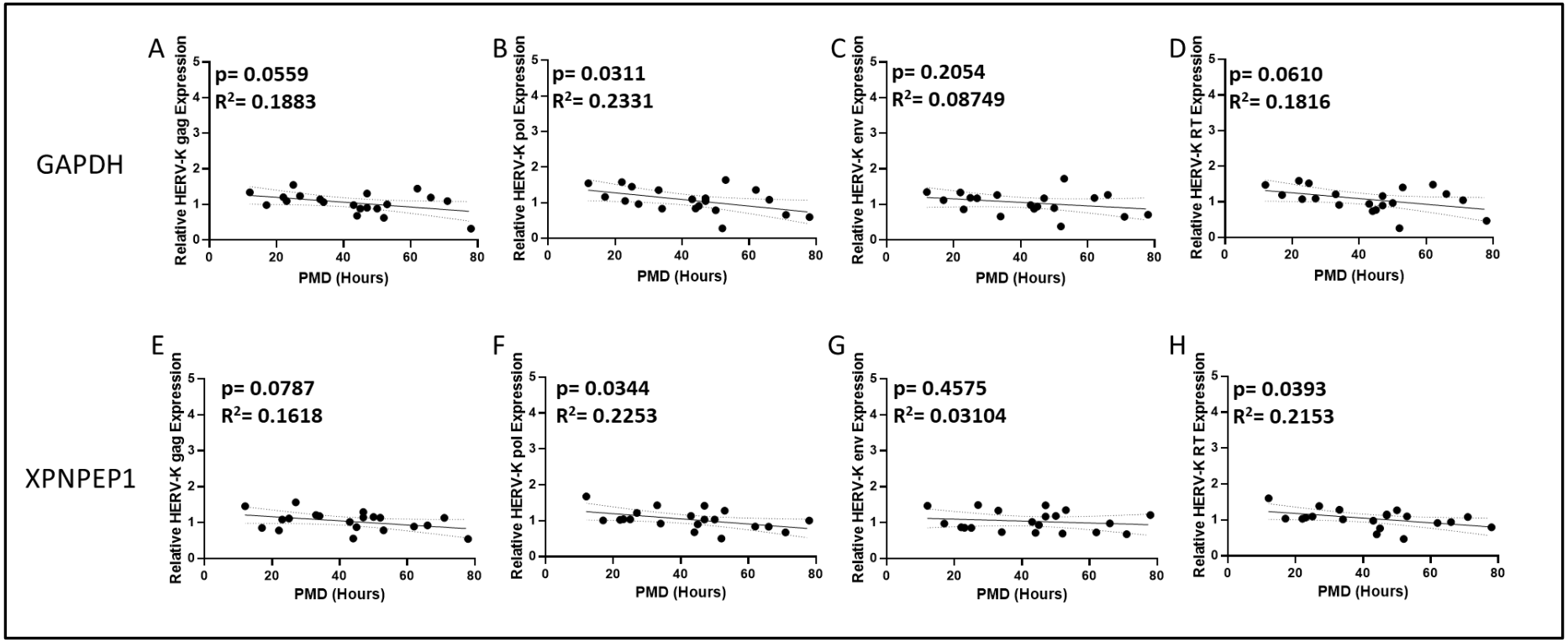


Figure S46. Effect of PMD on HERV-K transcript levels when normalised to GAPDH and XPNPEP1 in Non-ALS Control Patient Tissue.
The figure above shows the effect of PMD on HERV-K expression for *gag, pol, env* and *RT* genomic regions. A, B, C and D are normalised to GAPDH which shows no or only very slight effect of PMD on HERV-K pol expression increasing PMD (B). There was no correlation to HERV-K expression when the data was normalised to XPNPEP1 (E, F, G and H). R^2 values and p values were calculated in GraphPad Prism v8.0. This graph utilises data from ΔΔCt Normalisation method.

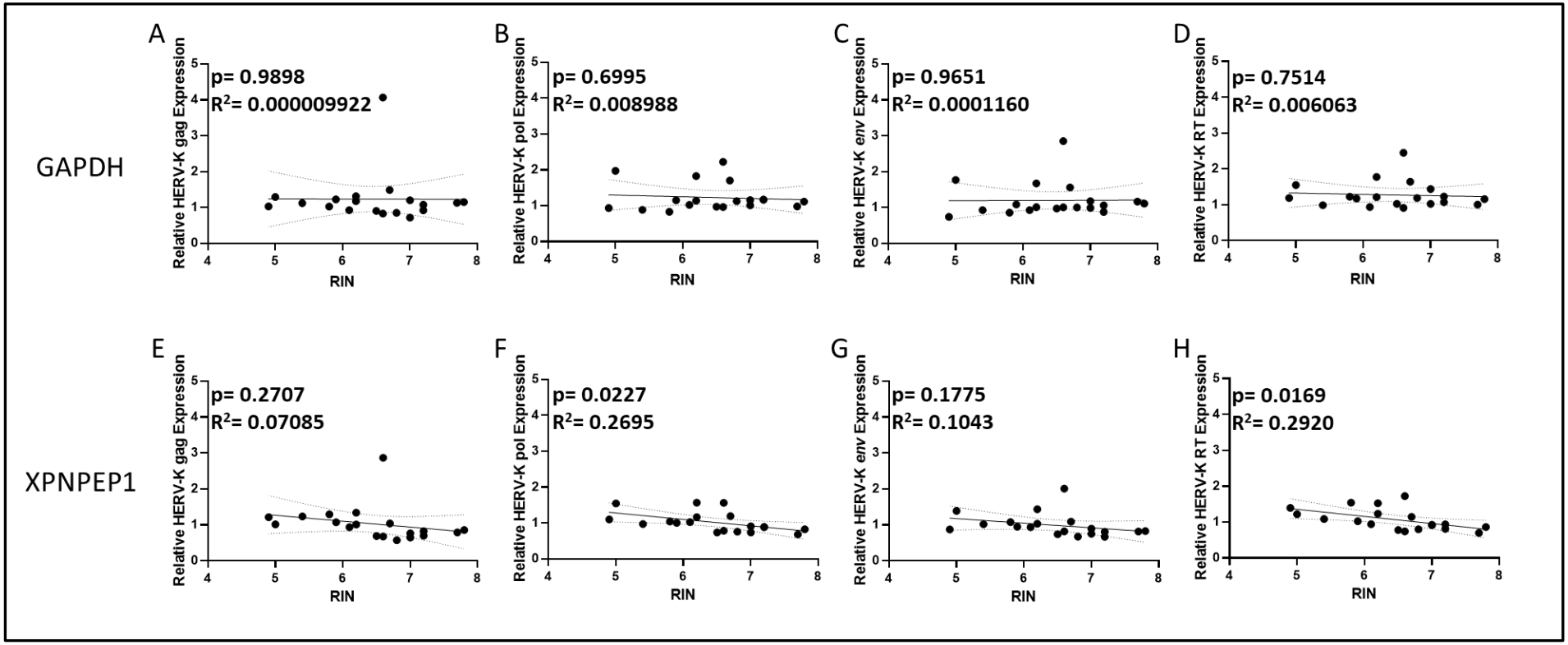


Figure S47. Effect of RNA integrity value on HERV-K gene transcript expression In ALS Patient Tissue Samples.

The data displayed in the graph above shows HERV-K expression when normalized against 2 separate reference genes, GAPDH and XPNPEP1. Linear regression graphs A, B, C and D show expression data normalised against GAPDH while E, F, G, and H display expression data normalised against XPNPEP1. R^2 values and P values were calculated in GraphPad Prism v8.0. All p values were not significant. This graph utilises data from ΔΔCt Normalisation method.

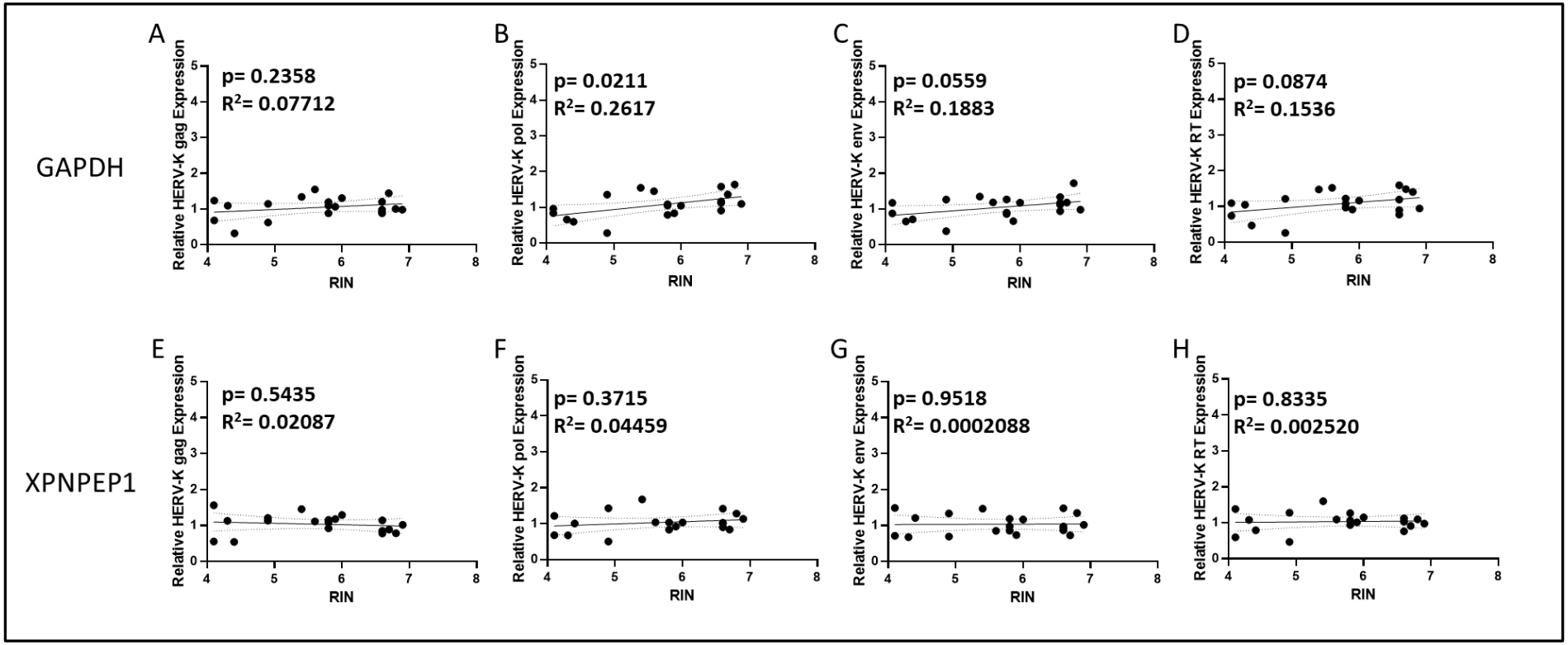


Figure S48. Effect of RNA integrity value on HERV-K gene transcript expression In Non-ALS Control Patient Tissue Samples.

The data displayed in the graph above shows HERV-K expression when normalized against 2 separate reference genes, GAPDH and XPNPEP1. Linear regression graphs A, B, C and D show expression data normalised against GAPDH while E, F, G, and H display expression data normalised against XPNPEP1. R^2 values and P values were calculated in GraphPad Prism v8.0. All p values were not significant. This graph utilises data from ΔΔCt Normalisation method.

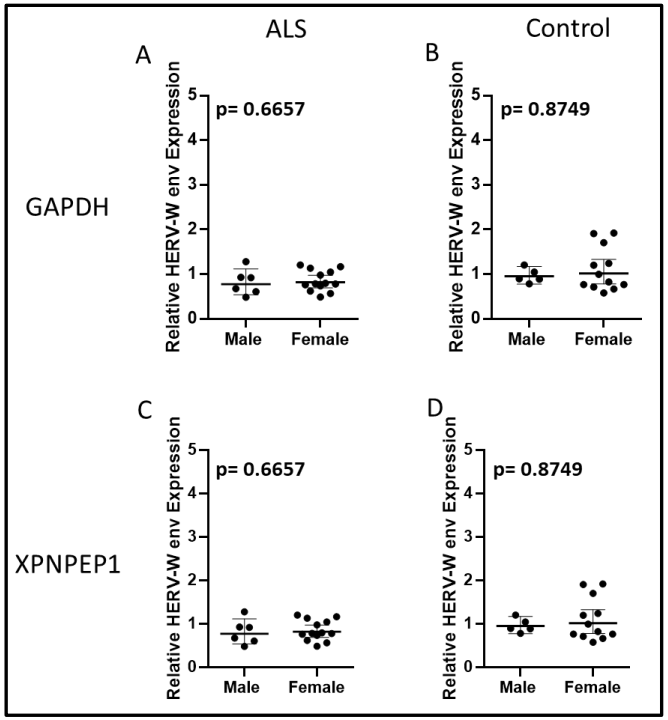


Figure S49. HERV-W *env* relative expression shows no correlation between male and female sample groups.

The data displayed in the graphs is normalised against 2 separate reference genes, being normalised to GAPDH or normalised to XPNPEP1. Horizontal lines correspond to the geometric mean with error bars representing a 95% confidence in the means position utilising data from ΔΔCt Normalisation method.

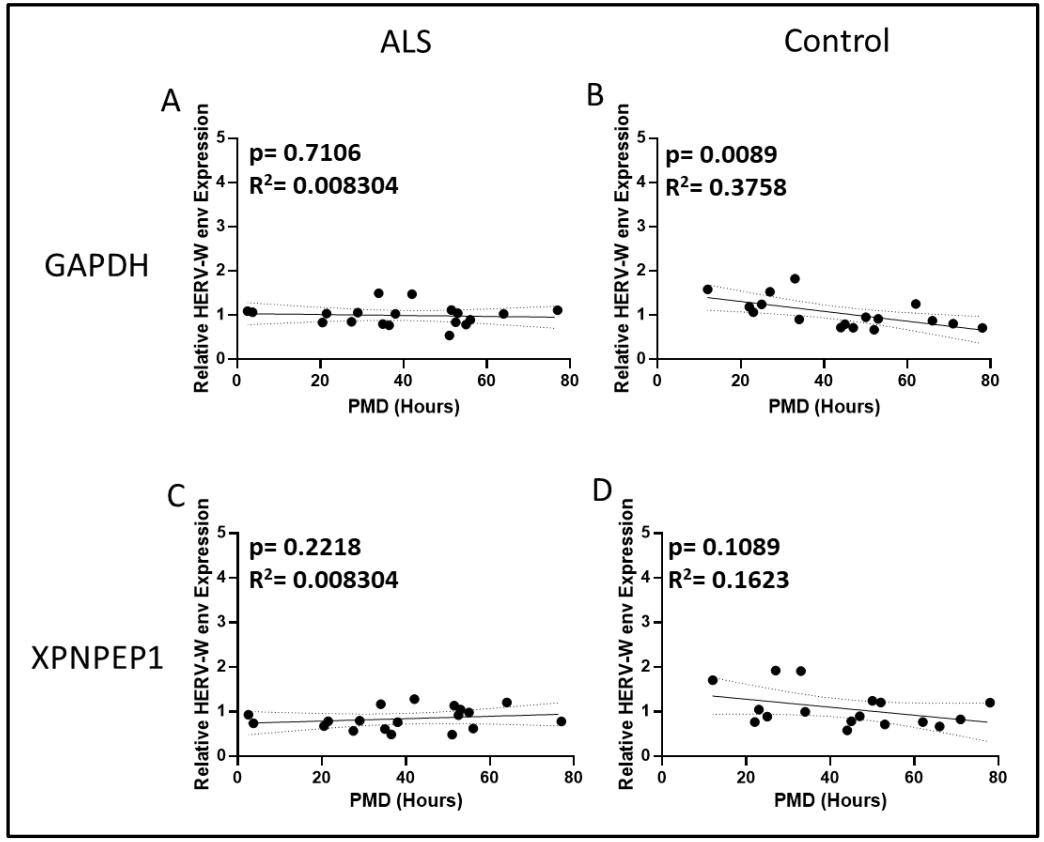


Figure S50. HERV-W *env* shows no significant expression with increasing post-mortem delay.

The data displayed in the graphs shows HERV-W *env* differential expression values normalised to 2 separate reference genes, with data in A normalised to GAPDH and data shown in B normalised to XPNPEP1. R^2 values and P values were calculated in GraphPad Prism v8.0 utilising data from ΔΔCt Normalisation method. All p values were not significant.

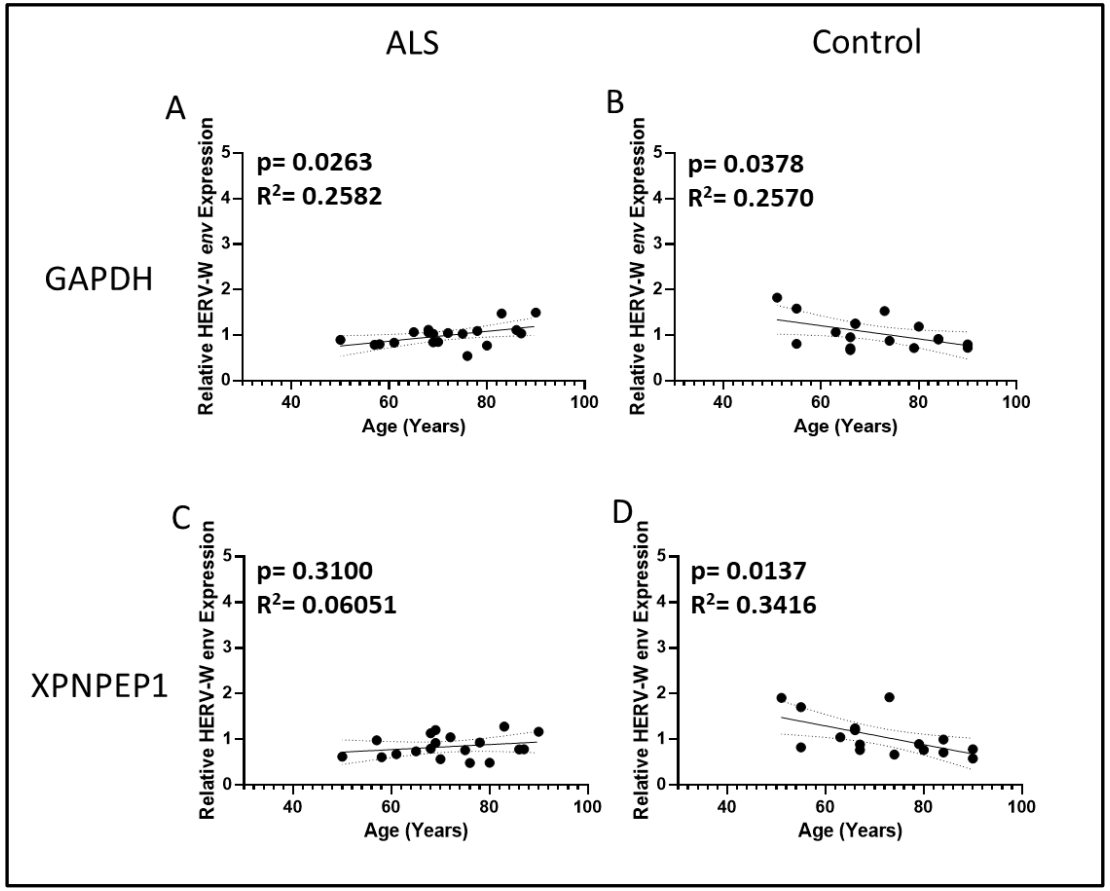


Figure S51. HERV-W *env* has no significant correlation with increasing age of patients at time of death.

There is no observable difference in HERV-W *env* differential expression when normalised against either GAPDH or XPNEP1. R^2 values and p values were calculated in GraphPad Prism v8.0 utilising data from ΔΔCt Normalisation method. All p values were not significant.

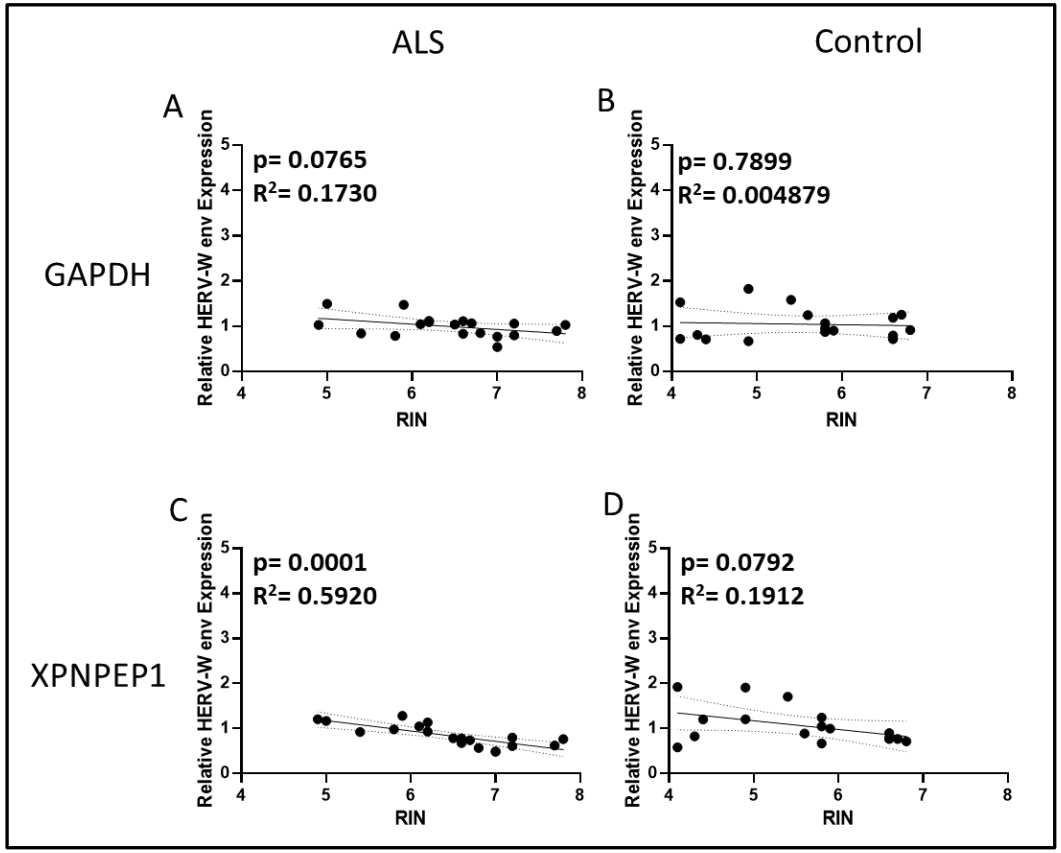


Figure S52. HERV-W *env* has no significant correlation with RNA integrity values.

As shown in the supplementary figure above HERV-W-env has no correlation with increasing RNA integrity value when normalised against either GAPDH or XPNPEP1. R^2 values and p values were calculated in GraphPad Prism v8.0 utilising data from ΔΔCt Normalisation method. All p values were not significant.

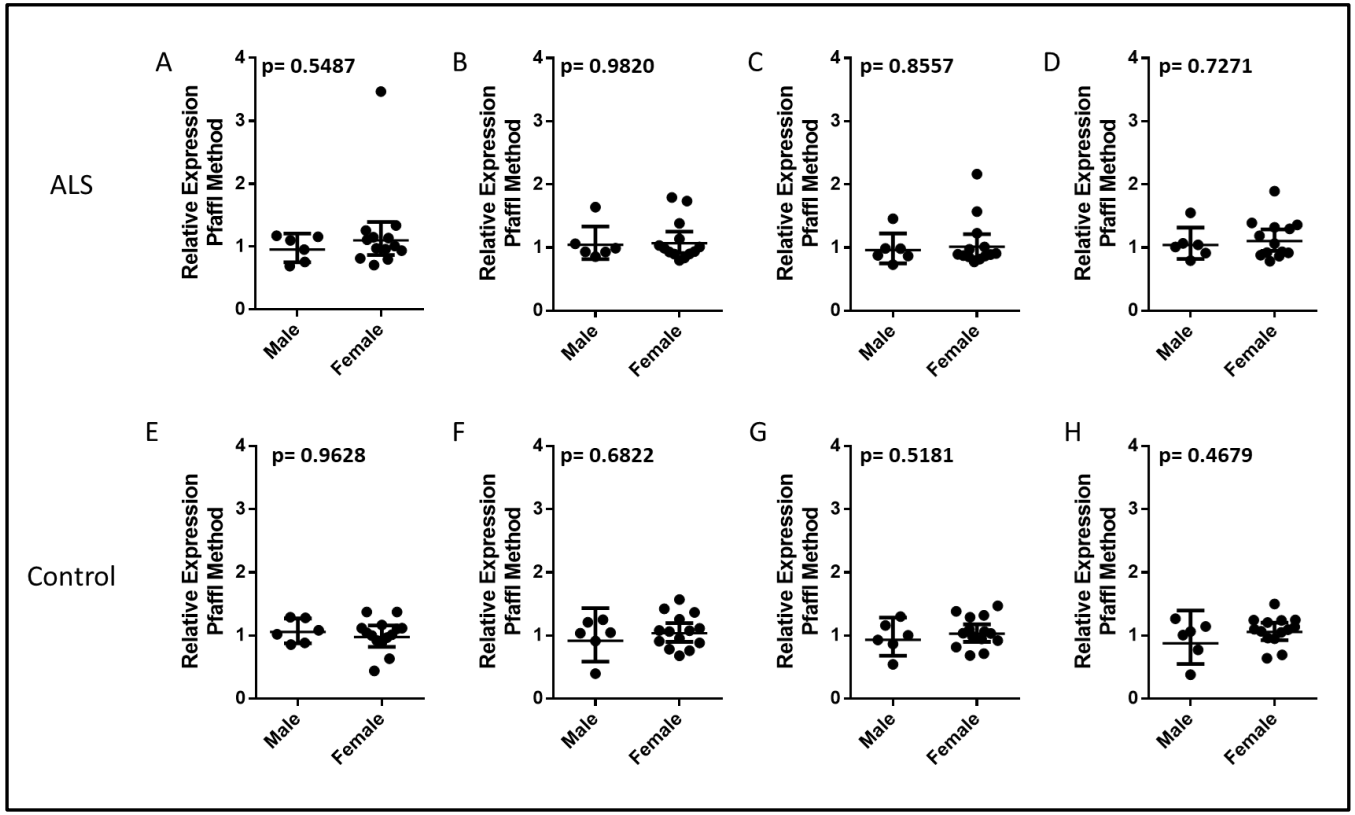


Figure S53. Effect of Sex on Expression Data from ALS and non-ALS control Samples when normalised using Pfaffl Differential Expression Method.

The Figure above shows the differential expression data for the effect of gender in ALS and non-ALS Postmortem Premotor cortex tissue samples. Graphs A & E show data for HERV-K *gag* transcripts, B & F show data for HERV-K *pol* transcripts, C & G show data for HERV-K *env* transcripts and D & H show data for HERV-K *env* transcripts. This data was generated using GraphPad v8.0.

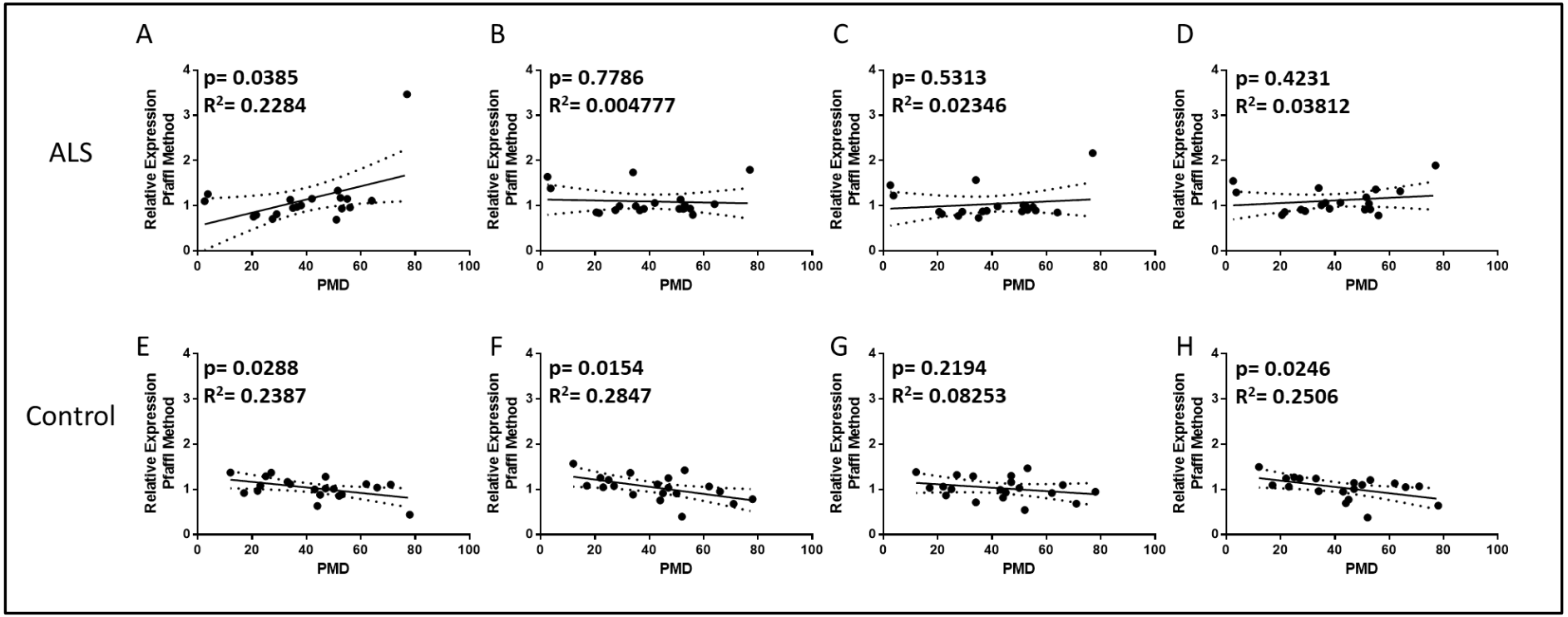


Figure S54. Effect of Postmortem Delay on Expression Data from ALS and non-ALS control Samples when normalised using Pfaffl Differential Expression Method.

The Figure above shows the differential expression data for the effect of postmortem delay in ALS and non-ALS Postmortem Premotor cortex tissue samples. Graphs A & E show data for HERV-K *gag* transcripts, B & F show data for HERV-K *pol* transcripts, C & G show data for HERV-K *env* transcripts and D & H show data for HERV-K *env* transcripts. This data was generated using GraphPad v8.0.

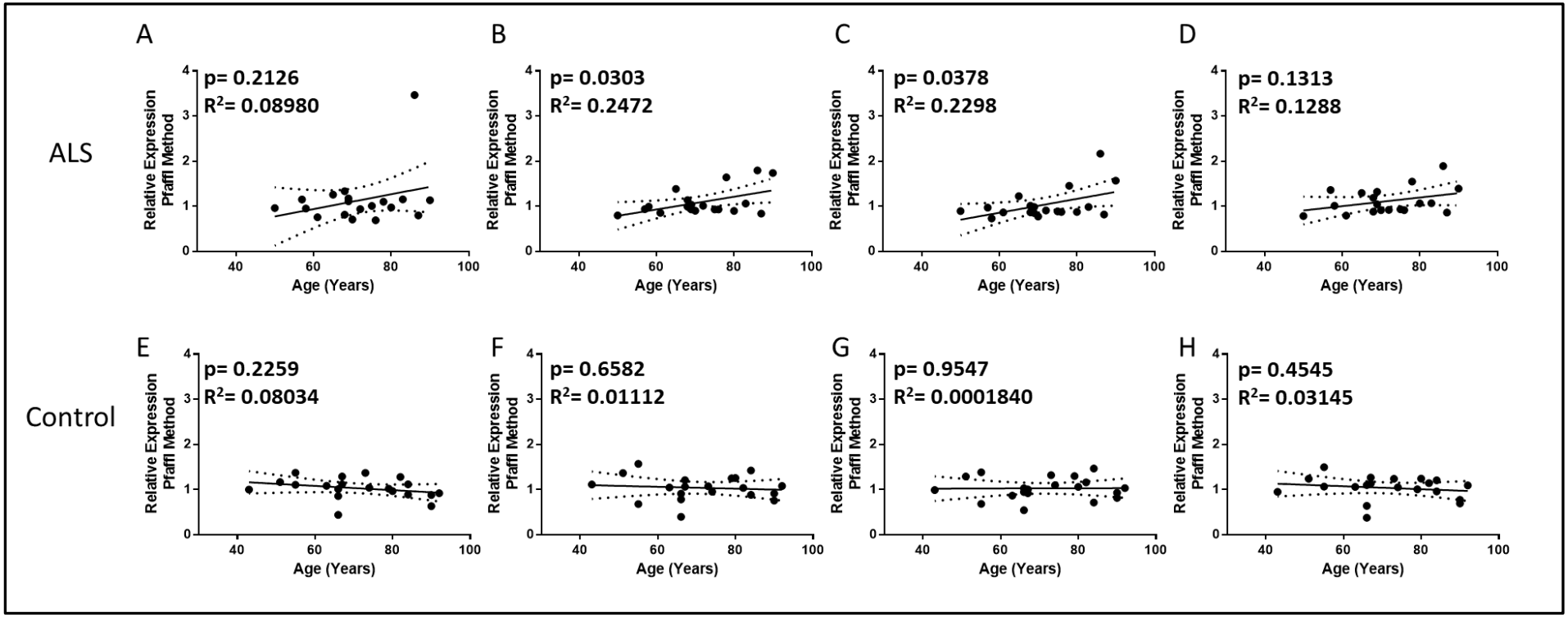


Figure S55. Effect of Age of Patient at time of Death on Expression Data from ALS and non-ALS control Samples when normalised using Pfaffl Differential Expression Method.

The Figure above shows the differential expression data for the effect of patient age at time of death in ALS and non-ALS Postmortem Premotor cortex tissue samples. Graphs A & E show data for HERV-K *gag* transcripts, B & F show data for HERV-K *pol* transcripts, C & G show data for HERV-K *env* transcripts and D & H show data for HERV-K *env* transcripts. This data was generated using GraphPad v8.0.

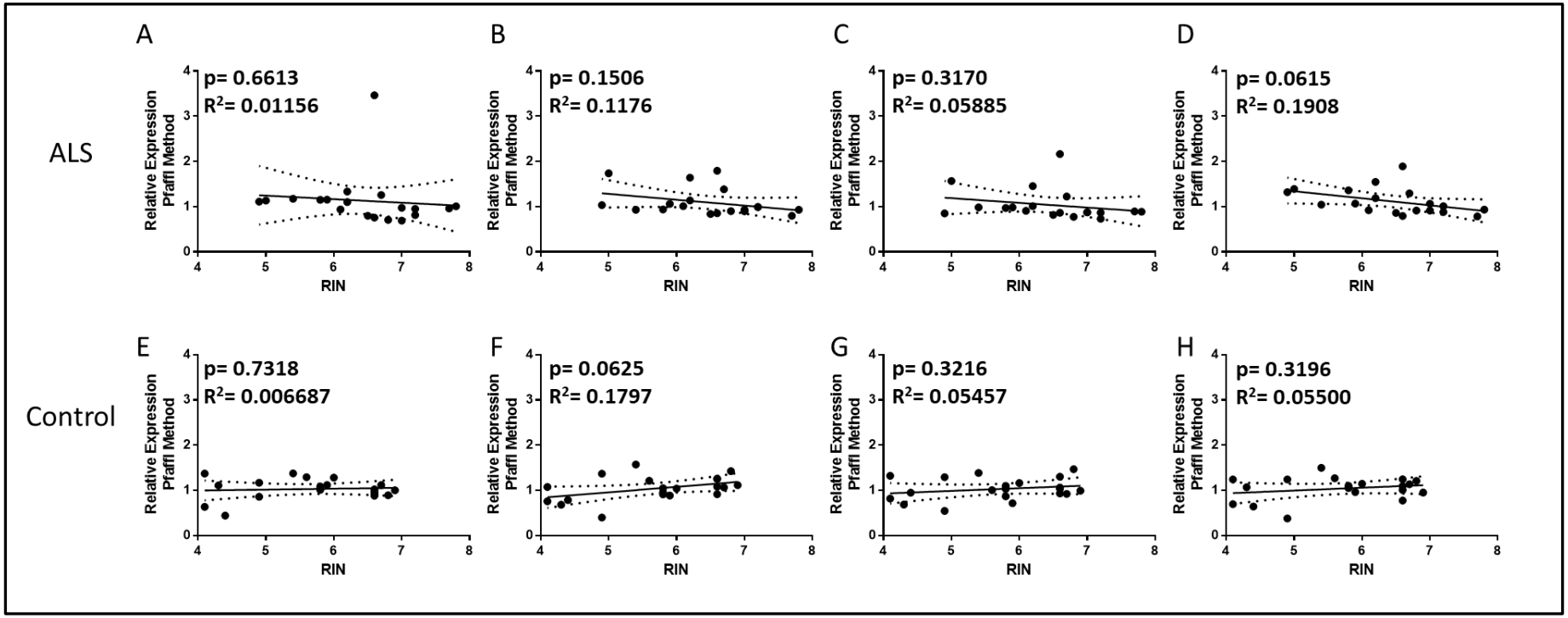


Figure S56. Effect of RNA Integrity Value on Expression Data from ALS and non-ALS control Samples when normalised using Pfaffl Differential Expression Method.

The Figure above shows the differential expression data for the effect of patient age at time of death in ALS and non-ALS Postmortem Premotor cortex tissue samples. Graphs A & E show data for HERV-K *gag* transcripts, B & F show data for HERV-K *pol* transcripts, C & G show data for HERV-K *env* transcripts and D & H show data for HERV-K *env* transcripts. This data was generated using GraphPad v8.0.

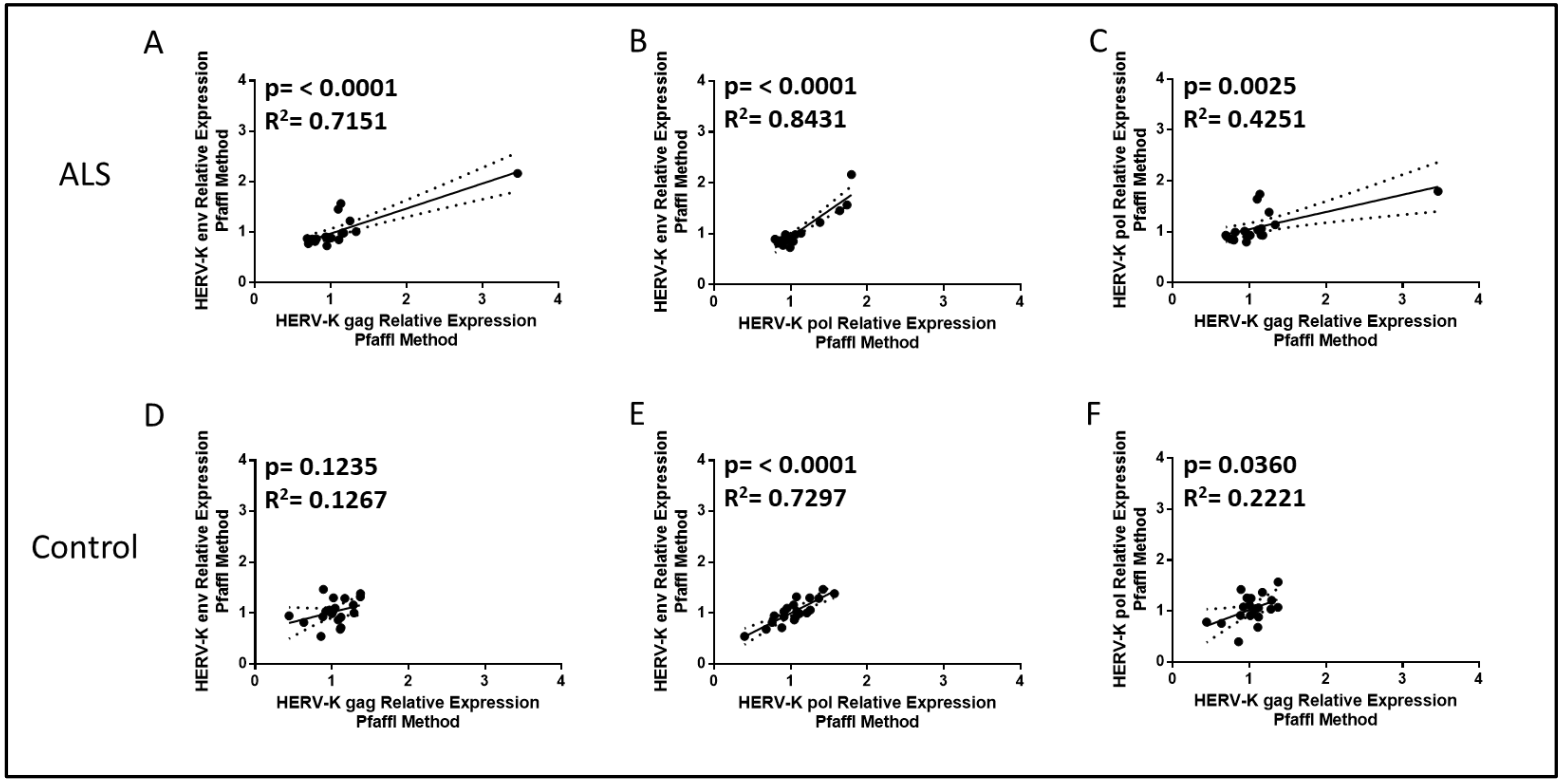


Figure S57. Correlation of HERV-K *gag, pol, env* Transcript Expression Data from ALS and non-ALS control Samples when normalised using Pfaffl Differential Expression Method.

The Figure above shows the correlation of HERV-K *gag, pol env* & *RT* differential expression data in ALS and non-ALS Postmortem Premotor cortex tissue samples. Graphs A & D shows comparison of HERV-K *env* & *gag* transcripts, B & E shows correlation of HERV-K *env* & *pol* transcripts and C & F shows data for the correlation of HERV-K *pol* & *gag* transcripts. This data was generated using GraphPad v8.0.

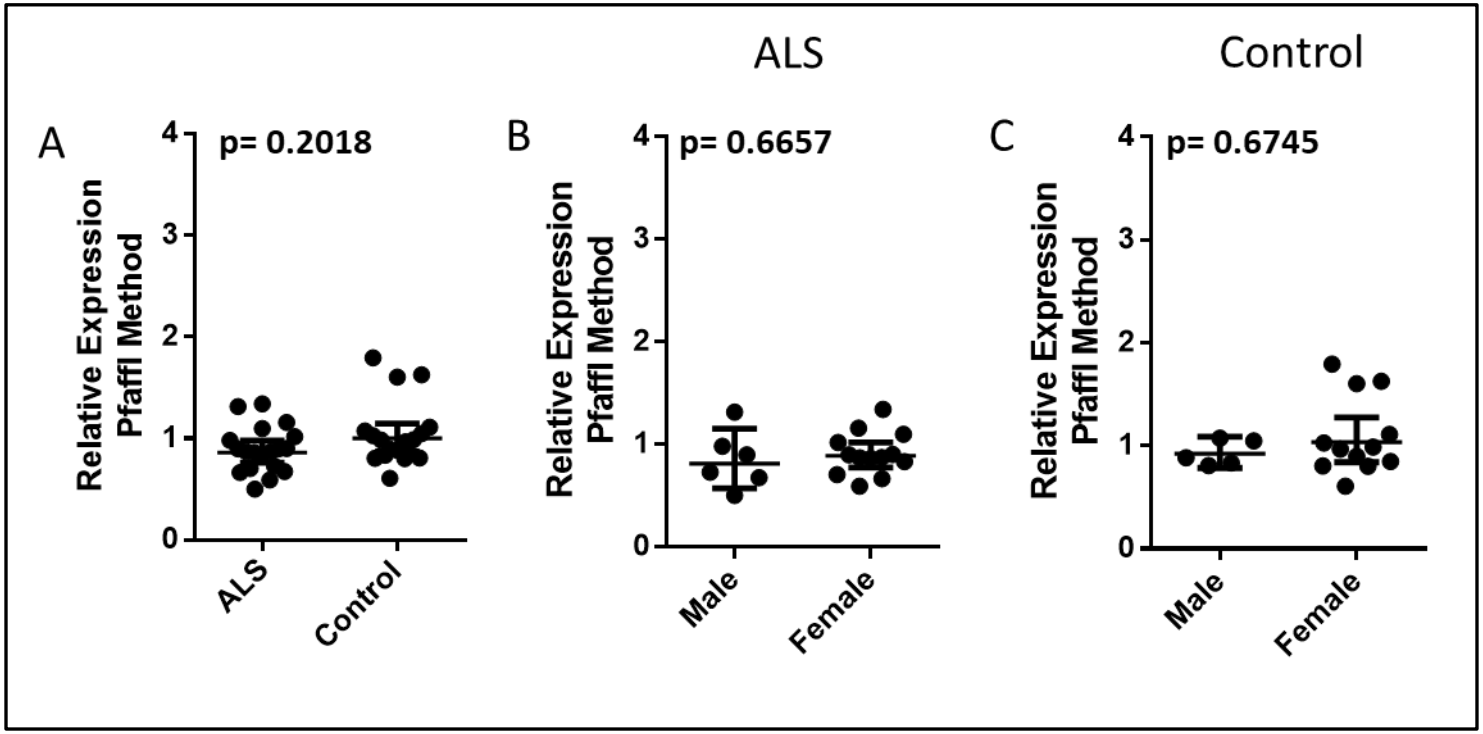


Figure S58. The effect of Disease status and Gender on HERV-W *env* Expression in ALS and non-ALS Control Postmortem Premotor Cortex Tissue Samples using Pfaffl Gene Expression Method.

The Figure above shows the differential expression data for HERV-W *env* transcripts between A) ALS and non-ALS Postmortem Premotor cortex tissue samples and the effect of gender in ALS and Controls (B&C). This data was generated using GraphPad v8.0.

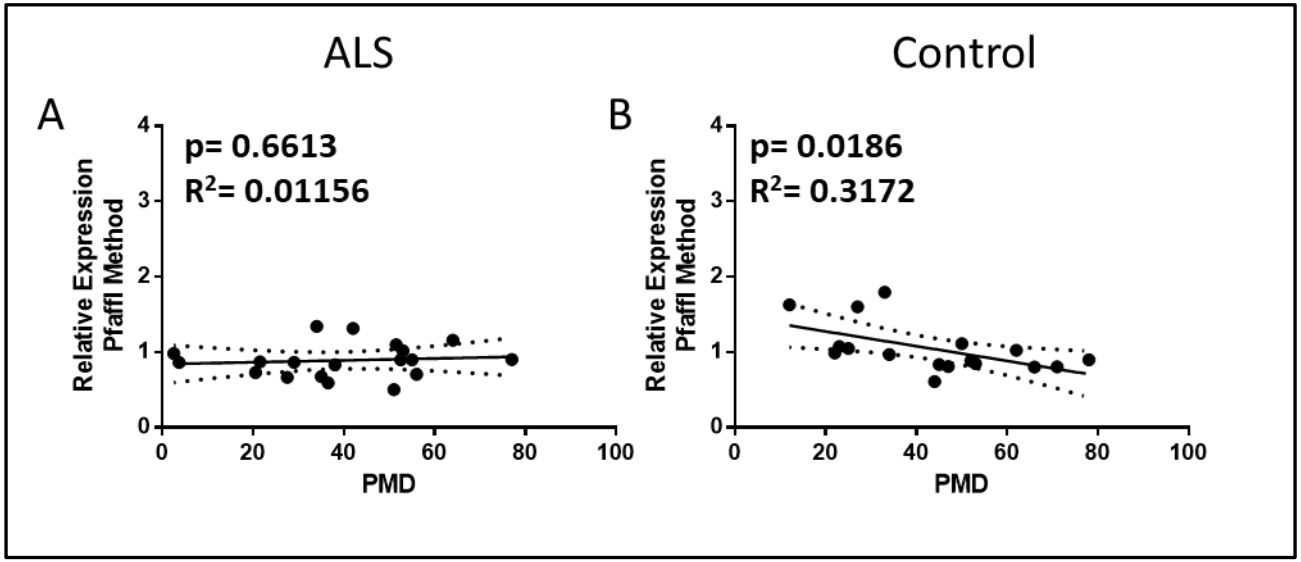


Figure S59. The effect of Postmortem Delay on HERV-W *env* Expression in ALS and non-ALS Control Postmortem Premotor Cortex Tissue Samples using Pfaffl Gene Expression Method.

The Figure above shows the effect of Postmortem Delay on HERV-W *env* transcript differential expression data for A) ALS and B) non-ALS Postmortem Premotor cortex tissue samples. This data was generated using GraphPad v8.0.

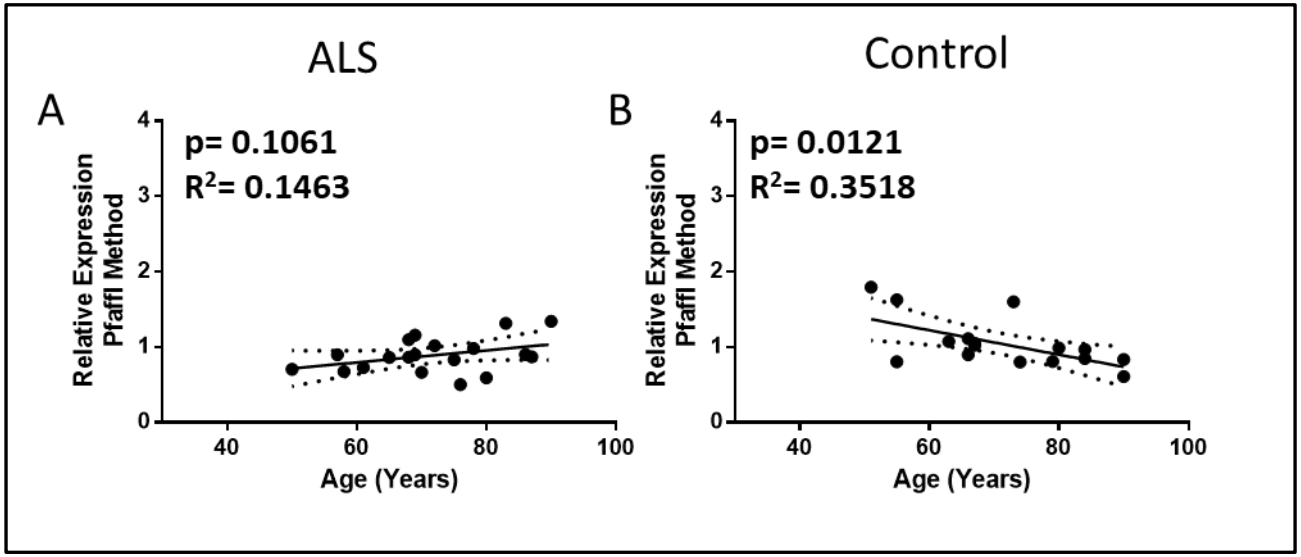


Figure S60. The effect of Age of Patient at Time of Death on HERV-W *env* Expression in ALS and non-ALS Control Postmortem Premotor Cortex Tissue Samples using Pfaffl Gene Expression Method.

The Figure above shows the effect of patient age on HERV-W *env* transcript differential expression data for A) ALS and B) non-ALS Postmortem Premotor cortex tissue samples. This data was generated using GraphPad v8.0.

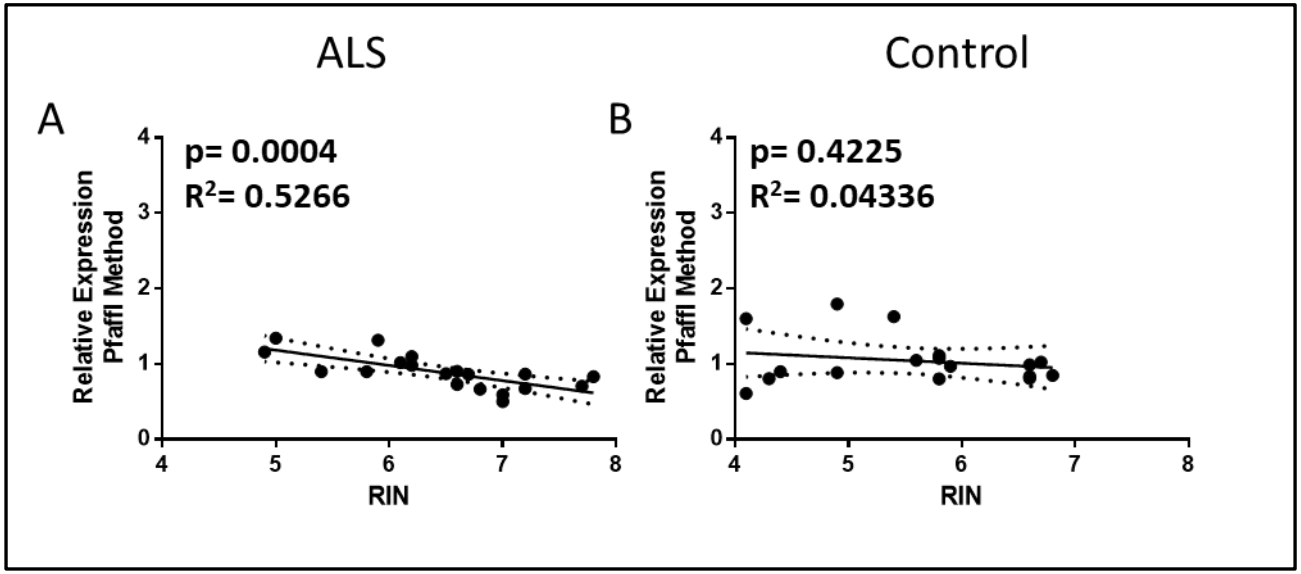


Figure S61. The effect of RNA Integrity on HERV-W *env* Expression in ALS and non-ALS Control Postmortem Premotor Cortex Tissue Samples using Pfaffl Gene Expression Method.

The Figure above shows the effect of RNA Integrity (RIN) on HERV-W *env* transcript differential expression data for A) ALS and B) non-ALS Postmortem Premotor cortex tissue samples. This data was generated using GraphPad v8.0.

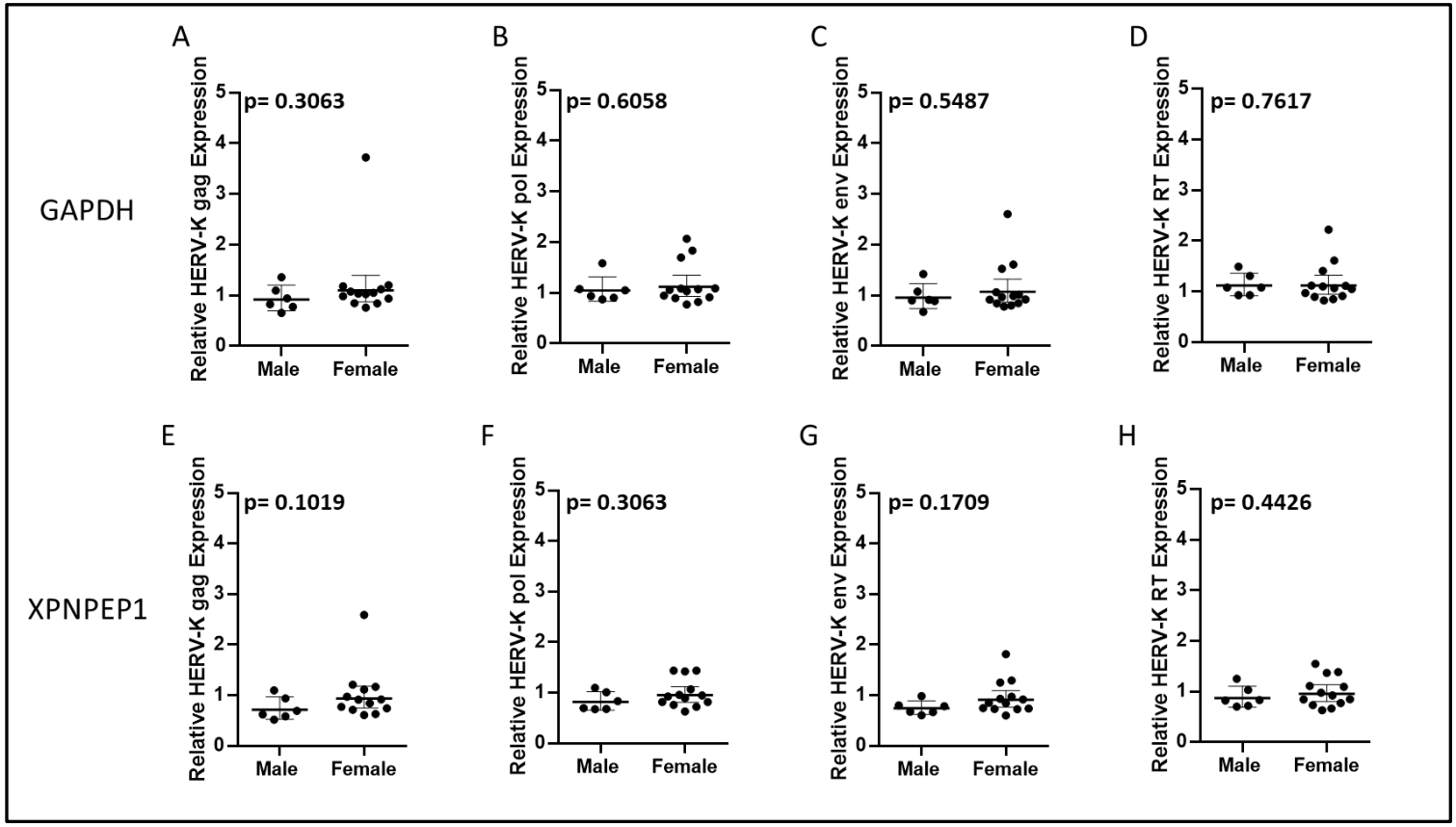


Figure S62. No significant correlation between gender and HERV-K RNA expression in Postmortem Premotor Cortex Tissue from n=19 ALS Patients from No-Cancer Control ΔΔCt Differential Expression Analysis.

There is no significant difference in HERV-K transcript expression between male and female groups when the data is normalised to either GAPDH (A, B, C and D). or when normalised to XPNPEP1 (E, F, G, and H). This graph utilises data from ΔΔCt Normalisation method.

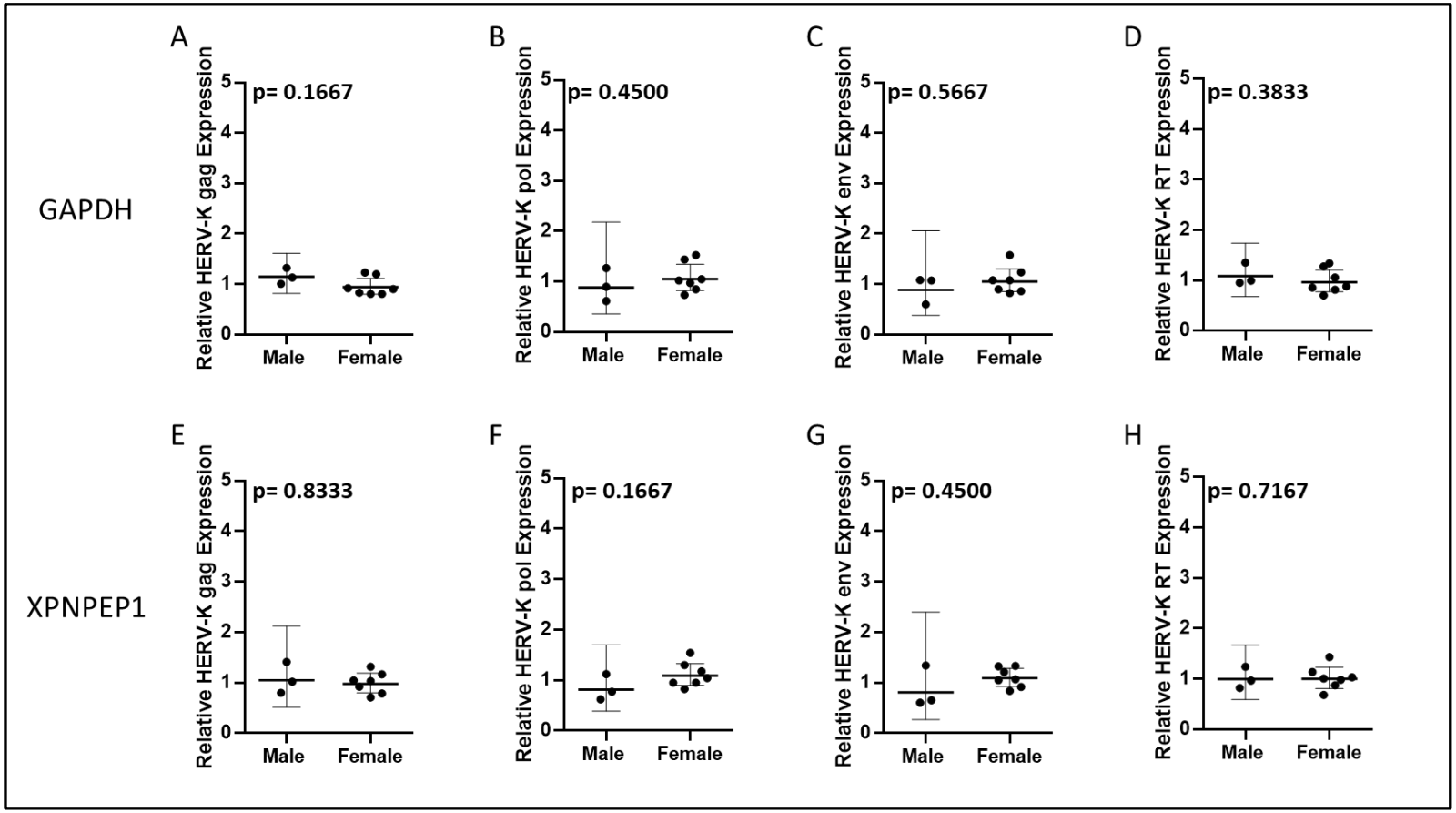


Figure S63. No significant correlation between gender and HERV-K RNA expression in Postmortem Premotor Cortex Tissue from n=8 no-Cancer Control Cases from No-Cancer Control ΔΔCt Differential Expression Analysis.

There is no significant difference in HERV-K transcript expression between male and female groups when the data is normalised to either GAPDH (A, B, C and D). or when normalised to XPNPEP1 (E, F, G, and H). This graph utilises data from ΔΔCt Normalisation method.

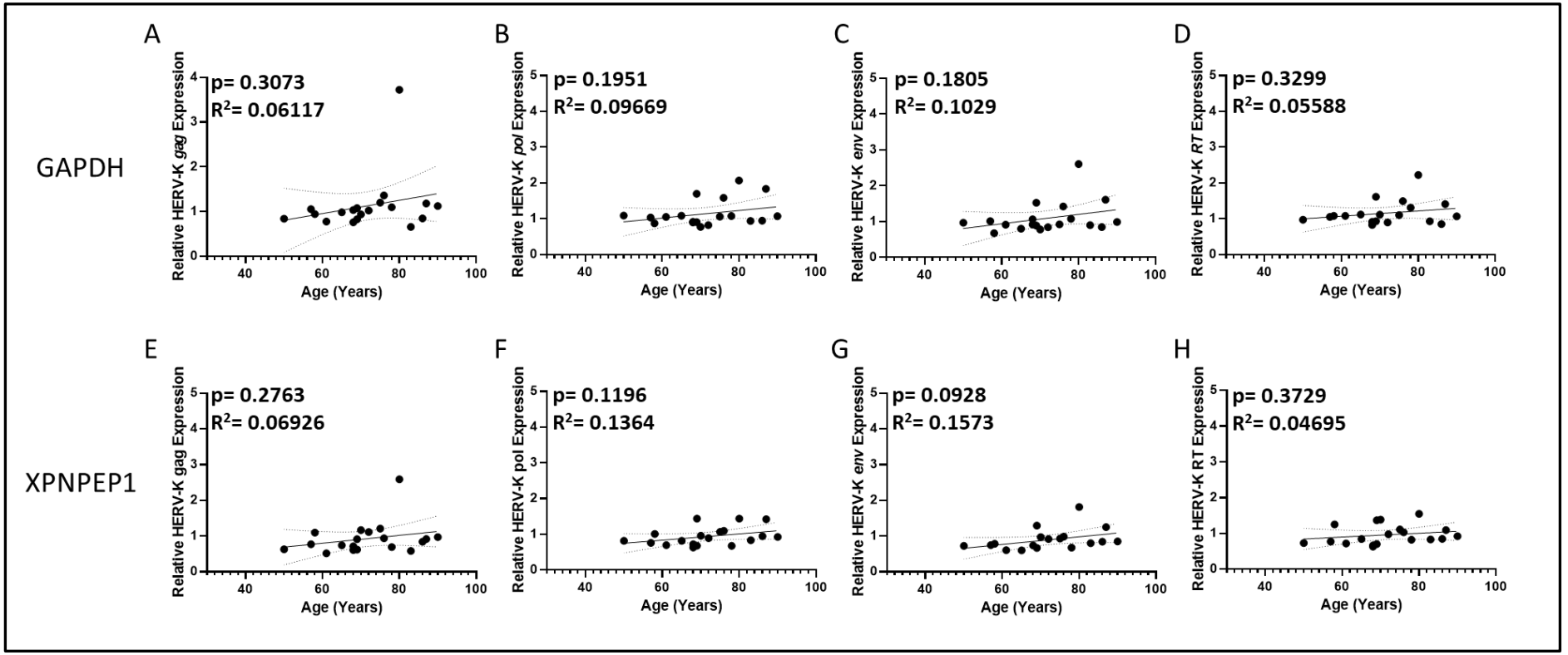


Figure S64. Effect of increasing age of patient at time of death on HERV-K Transcript expression from n=19 ALS Patients from No-Cancer Control ΔΔCt Differential Expression Analysis.

Data shown in the figure above were normalised against 2 separate reference genes, with A, B, C and D normalised against GAPDH and E, F, G and H normalised against XPNPEP1. Normalisation against different reference genes had no effect on results R^2 values were calculated in GraphPad Prism v8.0. This graph utilises data from ΔΔCt Normalisation method.

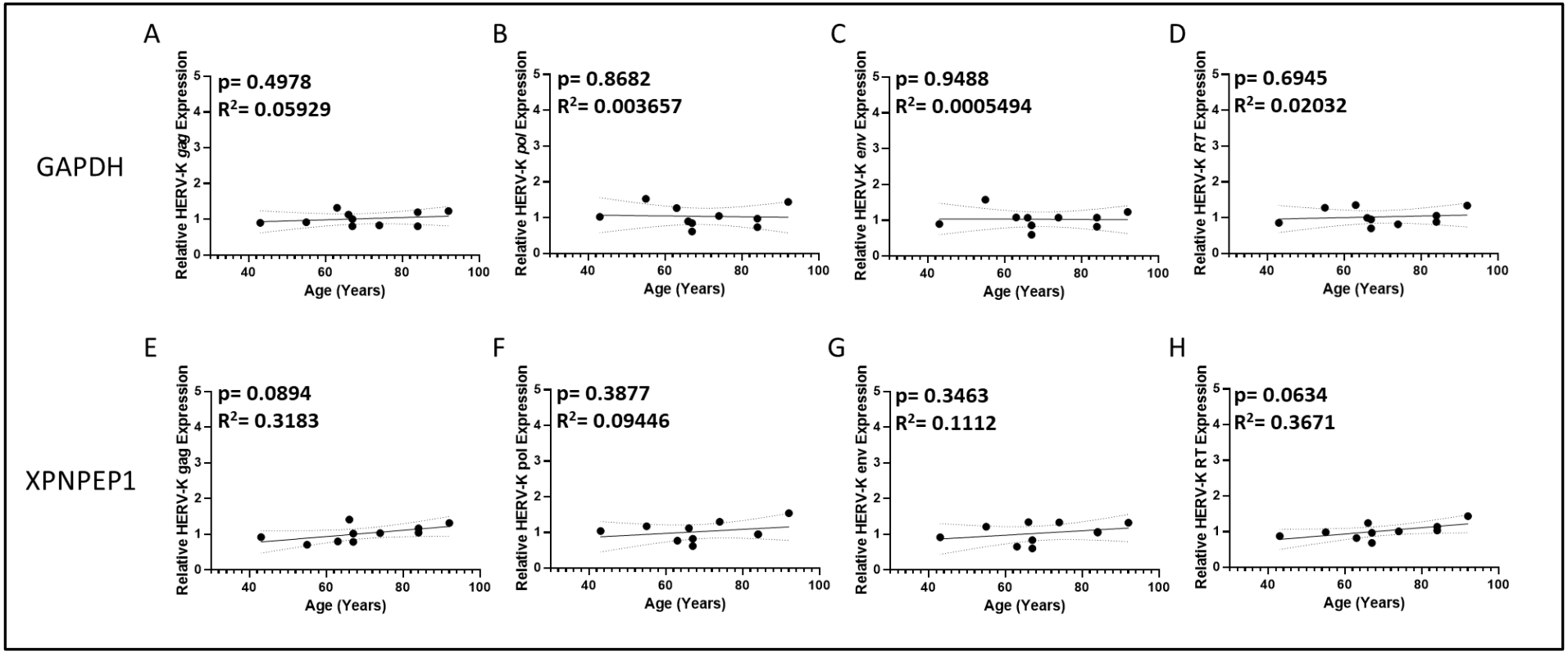


Figure S65. Effect of increasing age of patient at time of death on HERV-K Transcript expression from n=8 no-Cancer Control Cases from No-Cancer Control ΔΔCt Differential Expression Analysis.

Data shown in the figure above were normalised against 2 separate reference genes, with A, B, C and D normalised against GAPDH and E, F, G and H normalised against XPNPEP1. Normalisation against different reference genes had no effect on results R^2 values were calculated in GraphPad Prism v8.0. This graph utilises data from ΔΔCt Normalisation method.

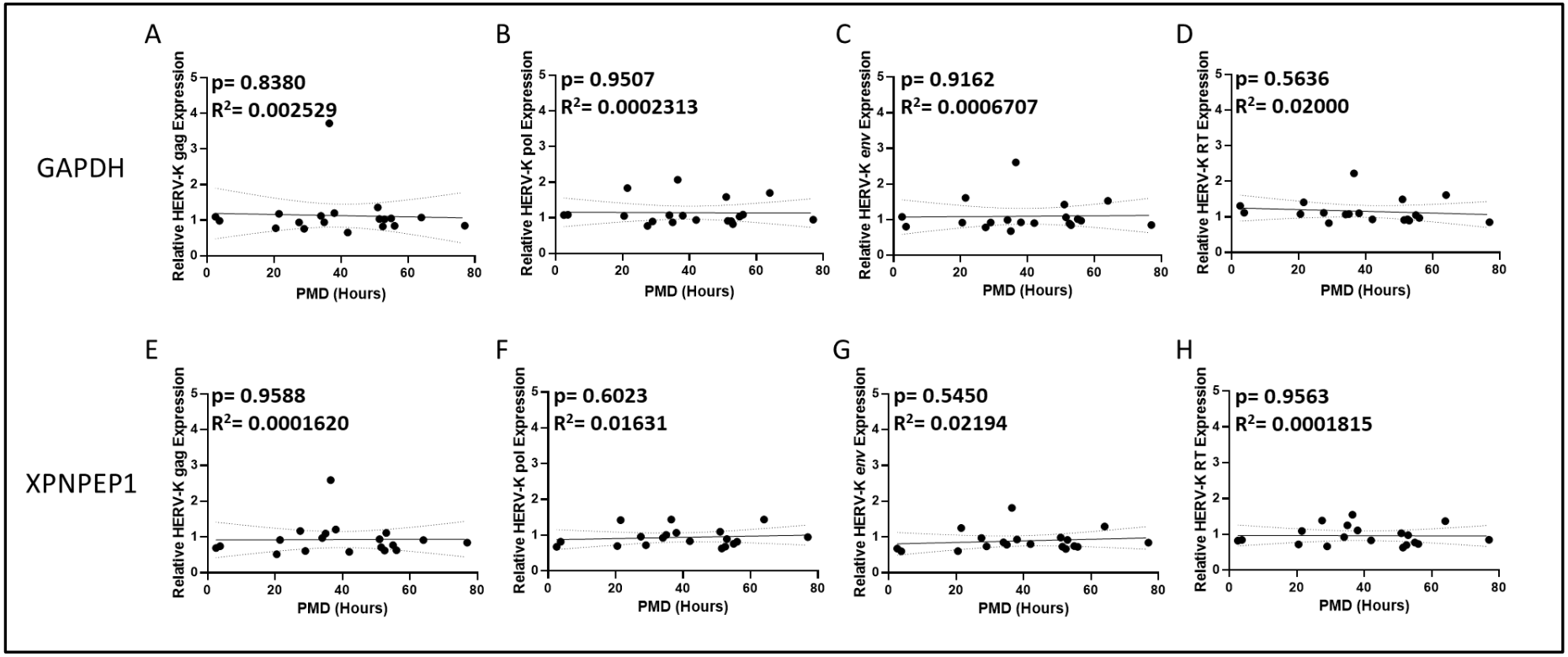


Figure S66. Effect of PMD on HERV-K transcript levels when normalised to GAPDH and XPNPEP1 from n=19 ALS Patients from No-Cancer Control ΔΔCt Differential Expression Analysis.

The figure above shows the effect of PMD on HERV-K expression for *gag, pol, env* and *RT* genomic regions. A, B, C and D are normalised to GAPDH which shows no or only very slight effect of PMD on HERV-K pol expression increasing PMD (B). There was no correlation to HERV-K expression when the data was normalised to XPNPEP1 (E, F, G and H). R^2 values and p values were calculated in GraphPad Prism v8.0. This graph utilises data from ΔΔCt Normalisation method.

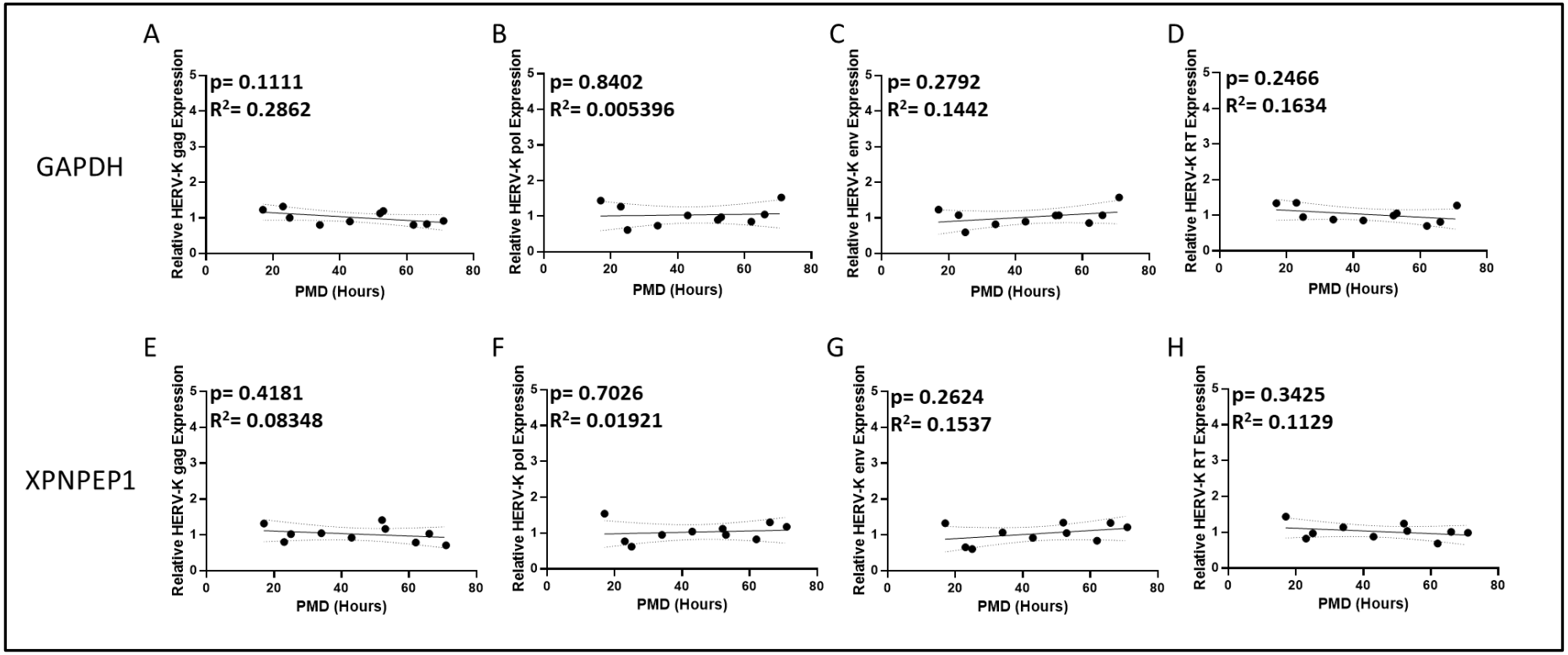


Figure S67. Effect of PMD on HERV-K transcript levels when normalised to GAPDH and XPNPEP1 from n=8 no-Cancer Control Cases from No-Cancer Control ΔΔCt Differential Expression Analysis.

The figure above shows the effect of PMD on HERV-K expression for *gag, pol, env* and *RT* genomic regions. A, B, C and D are normalised to GAPDH which shows no or only very slight effect of PMD on HERV-K pol expression increasing PMD (B). There was no correlation to HERV-K expression when the data was normalised to XPNPEP1 (E, F, G and H). R^2 values and p values were calculated in GraphPad Prism v8.0. This graph utilises data from ΔΔCt Normalisation method.

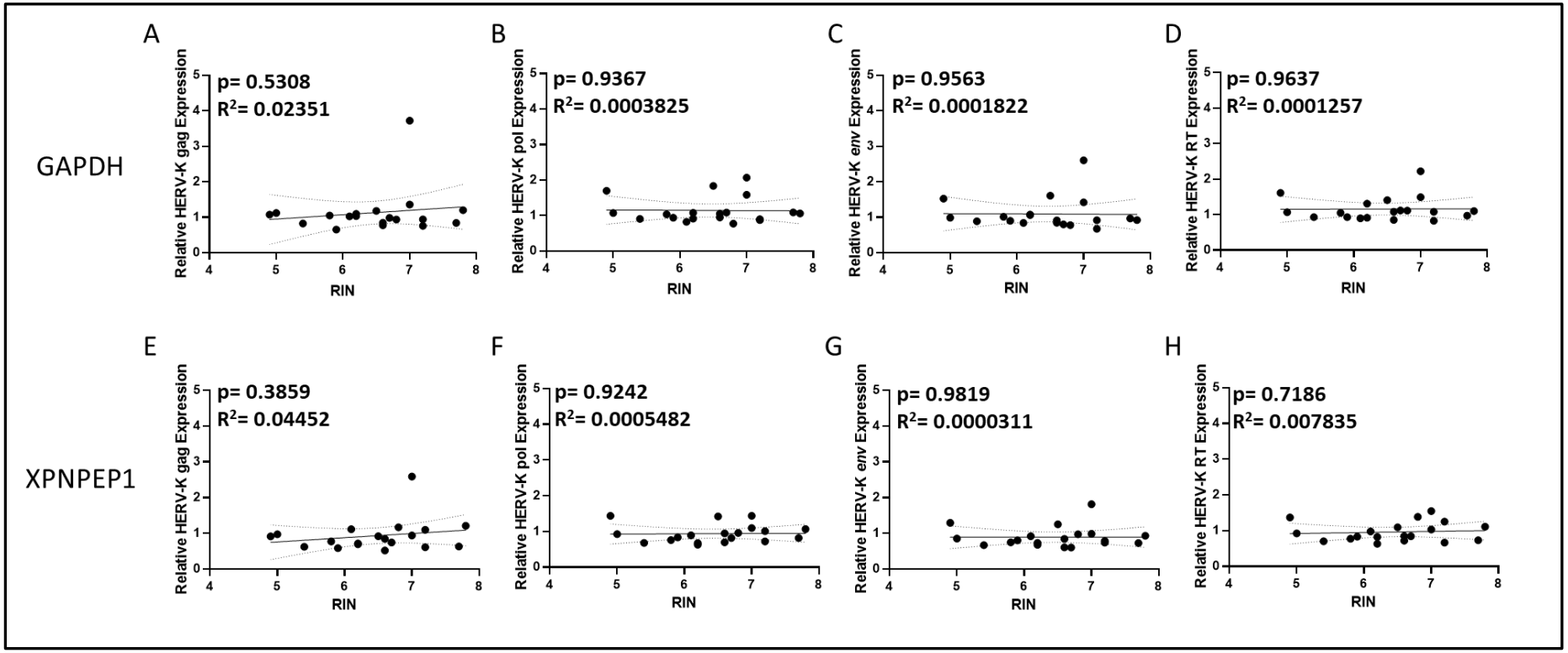


Figure S68. Effect of RNA integrity value on HERV-K gene transcript expression from n=19 ALS Patients from No-Cancer Control ΔΔCt Differential Expression Analysis.

The data displayed in the graph above shows HERV-K expression when normalized against 2 separate reference genes, GAPDH and XPNPEP1. Linear regression graphs A, B, C and D show expression data normalised against GAPDH while E, F, G, and H display expression data normalised against XPNPEP1. R^2 values and P values were calculated in GraphPad Prism v8.0. All p values were not significant. This graph utilises data from ΔΔCt Normalisation method.

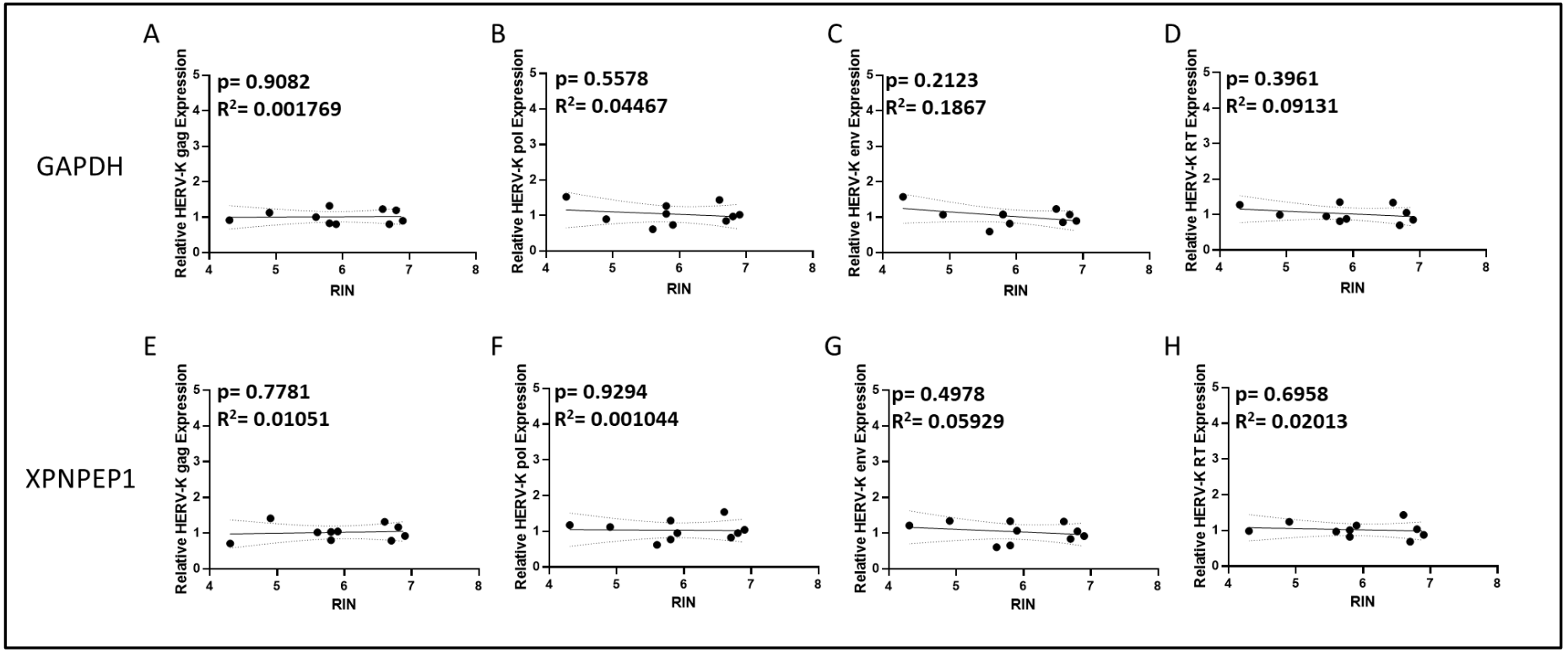


Figure S69. Effect of RNA integrity value on HERV-K gene transcript expression from n=8 no-Cancer Control Cases from No-Cancer Control ΔΔCt Differential Expression Analysis.

The data displayed in the graph above shows HERV-K expression when normalized against 2 separate reference genes, GAPDH and XPNPEP1. Linear regression graphs A, B, C and D show expression data normalised against GAPDH while E, F, G, and H display expression data normalised against XPNPEP1. R^2 values and P values were calculated in GraphPad Prism v8.0. All p values were not significant. This graph utilises data from ΔΔCt Normalisation method.

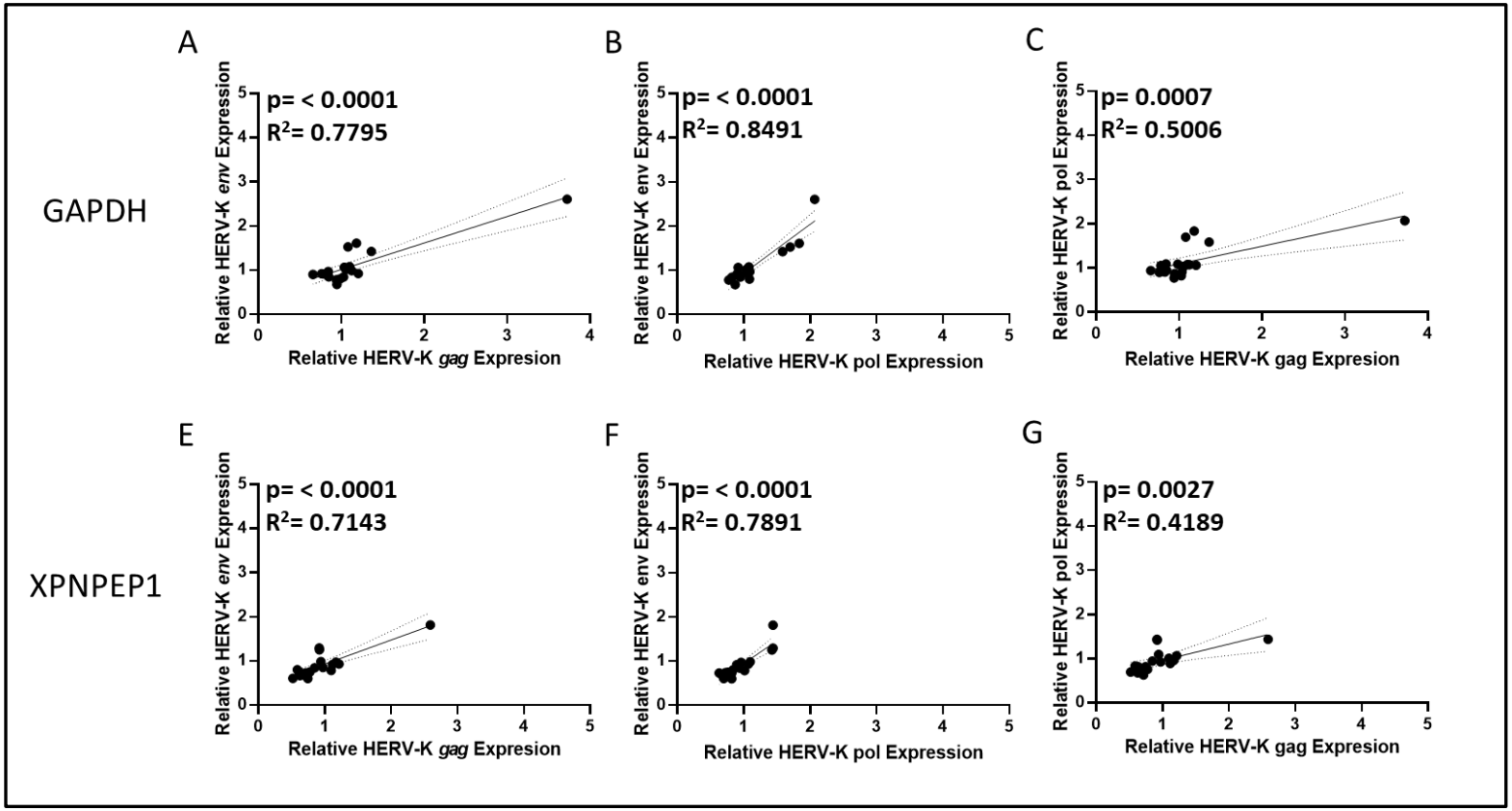


Figure S70. Correlation of HERV-K *gag, pol, env* & *RT* Transcript Expression Data from n=19 ALS Samples against no-Cancer Control Cases using ΔΔCt Method

The Figure above shows the correlation of HERV-K *gag, pol env* & *RT* differential expression data from n=19 ALS Cases when normalised to GAPDH or XPNPEP1. Graphs A & D shows comparison of HERV-K *env* & *gag* transcripts, B & E shows correlation of HERV-K *env* & *pol* transcripts and C & F shows data for the correlation of HERV-K *pol* & *gag* transcripts. This data was generated using GraphPad v8.0.

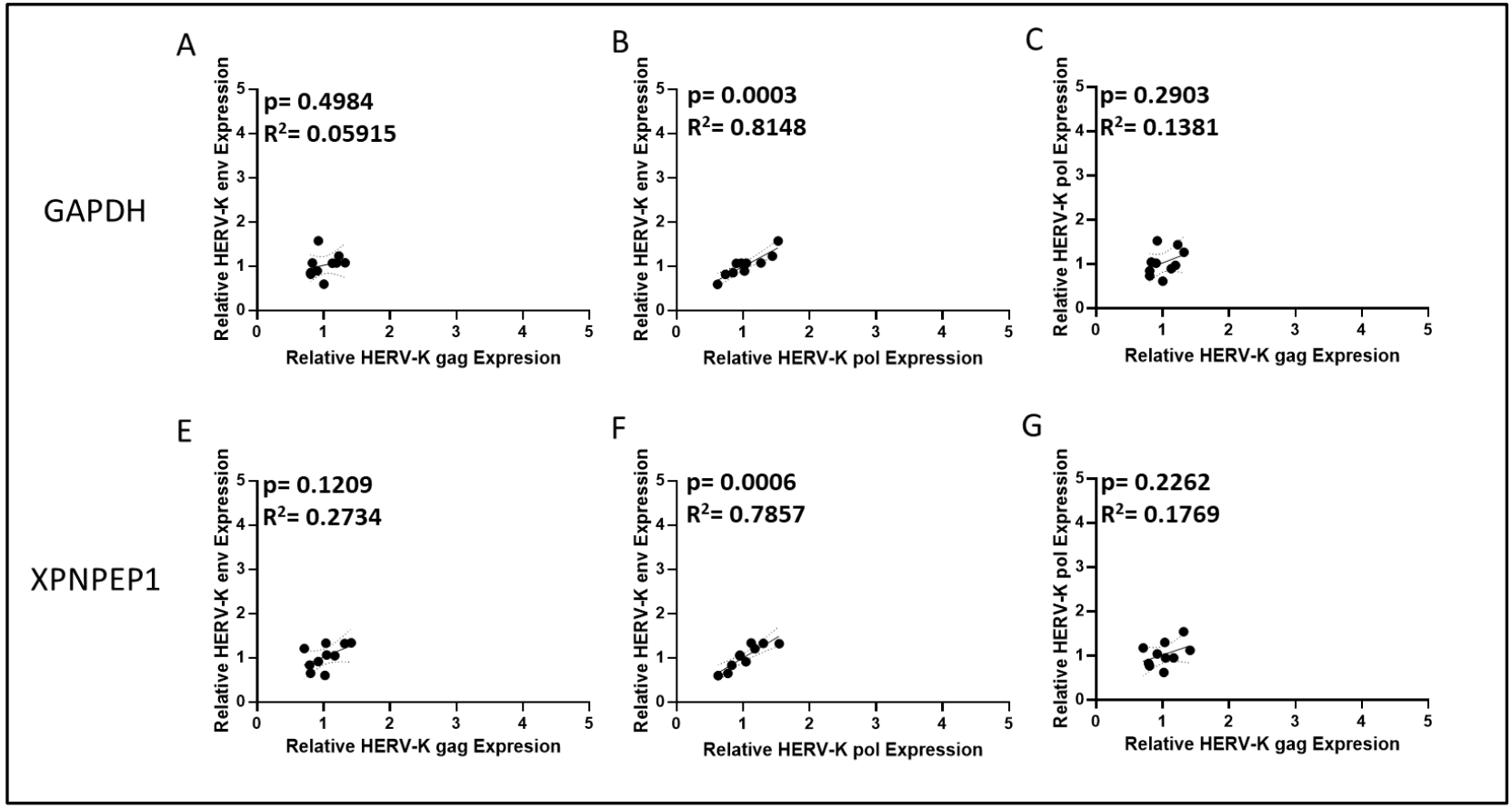


Figure S71. Correlation of HERV-K *gag, pol, env* & *RT* Transcript Expression from n=8 no-Cancer Control Cases using ΔΔCt Method

The Figure above shows the correlation of HERV-K *gag, pol env* & *RT* differential expression data from n=8 no-Cancer Control Cases when normalised to GAPDH or XPNPEP1. Graphs A & D shows comparison of HERV-K *env* & *gag* transcripts, B & E shows correlation of HERV-K *env* & *pol* transcripts and C & F shows data for the correlation of HERV-K *pol* & *gag* transcripts. This data was generated using GraphPad v8.0.

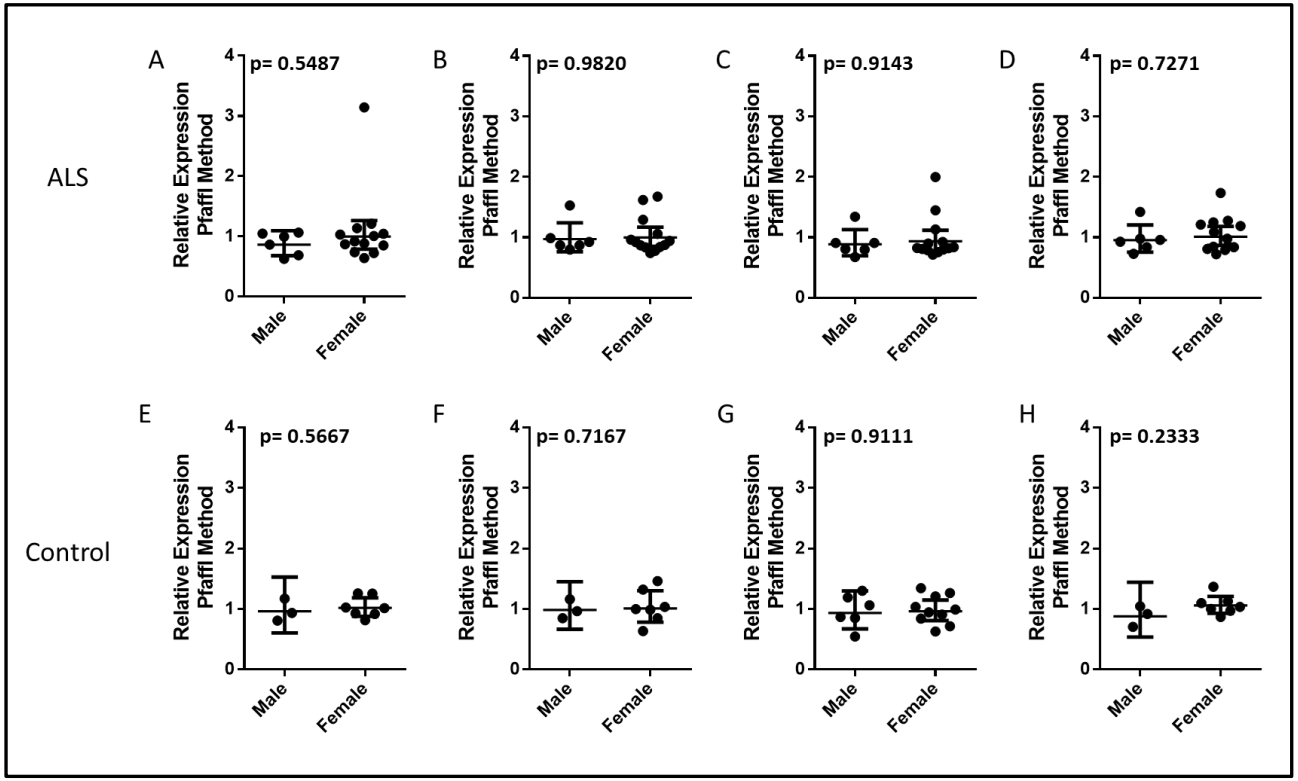


Figure S72. Effect of Sex on Expression Data from n=19 ALS and n=8 No-Cancer control Samples when normalised using Pfaffl Differential Expression Method.

The Figure above shows the differential expression data for the effect of gender in ALS and non-ALS Postmortem Premotor cortex tissue samples. Graphs A & E show data for HERV-K *gag* transcripts, B & F show data for HERV-K *pol* transcripts, C & G show data for HERV-K *env* transcripts and D & H show data for HERV-K *env* transcripts. This data was generated using GraphPad v8.0.

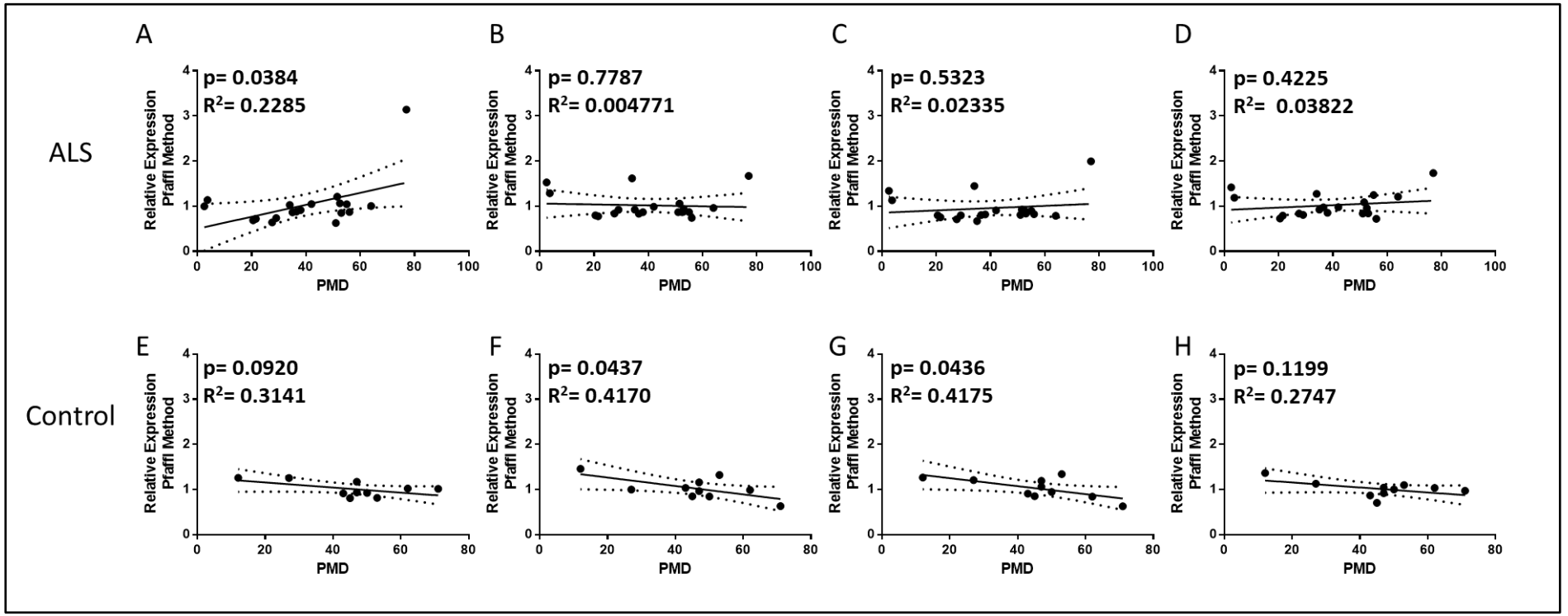


Figure S73. Effect of Postmortem Delay on Expression Data from n=19 ALS and n=8 No-Cancer control Samples when normalised using Pfaffl Differential Expression Method.

The Figure above shows the differential expression data for the effect of postmortem delay in ALS and non-ALS Postmortem Premotor cortex tissue samples. Graphs A & E show data for HERV-K *gag* transcripts, B & F show data for HERV-K *pol* transcripts, C & G show data for HERV-K *env* transcripts and D & H show data for HERV-K *env* transcripts. This data was generated using GraphPad v8.0.

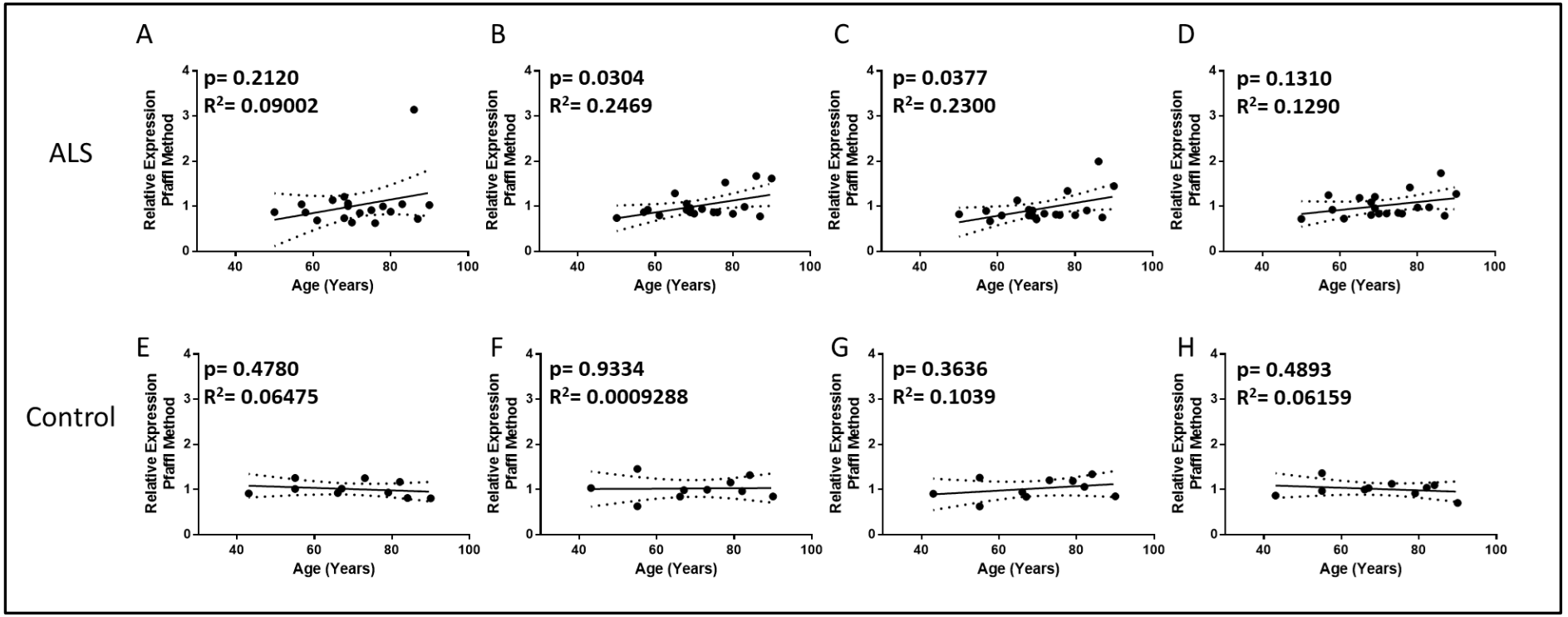


Figure S74. Effect of Age of Patient at time of Death on Expression Data from n=19 ALS and n=8 No-Cancer control Samples when normalised using Pfaffl Differential Expression Method.

The Figure above shows the differential expression data for the effect of patient age at time of death in ALS and non-ALS Postmortem Premotor cortex tissue samples. Graphs A & E show data for HERV-K *gag* transcripts, B & F show data for HERV-K *pol* transcripts, C & G show data for HERV-K *env* transcripts and D & H show data for HERV-K *env* transcripts. This data was generated using GraphPad v8.0.

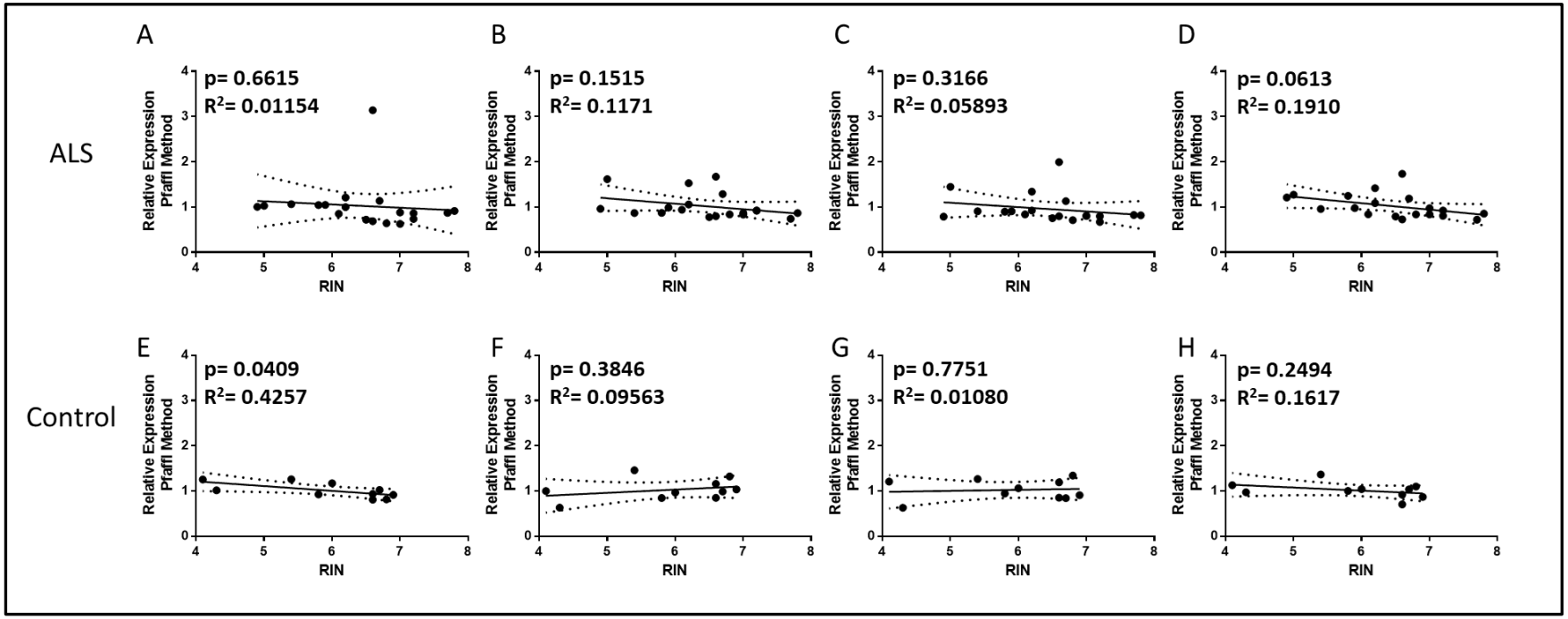


Figure S75. Effect of RNA Integrity Value on Expression Data from n=19 ALS and n=8 No-Cancer control Samples when normalised using Pfaffl Differential Expression Method.

The Figure above shows the differential expression data for the effect of patient age at time of death in ALS and non-ALS Postmortem Premotor cortex tissue samples. Graphs A & E show data for HERV-K *gag* transcripts, B & F show data for HERV-K *pol* transcripts, C & G show data for HERV-K *env* transcripts and D & H show data for HERV-K *env* transcripts. This data was generated using GraphPad v8.0.

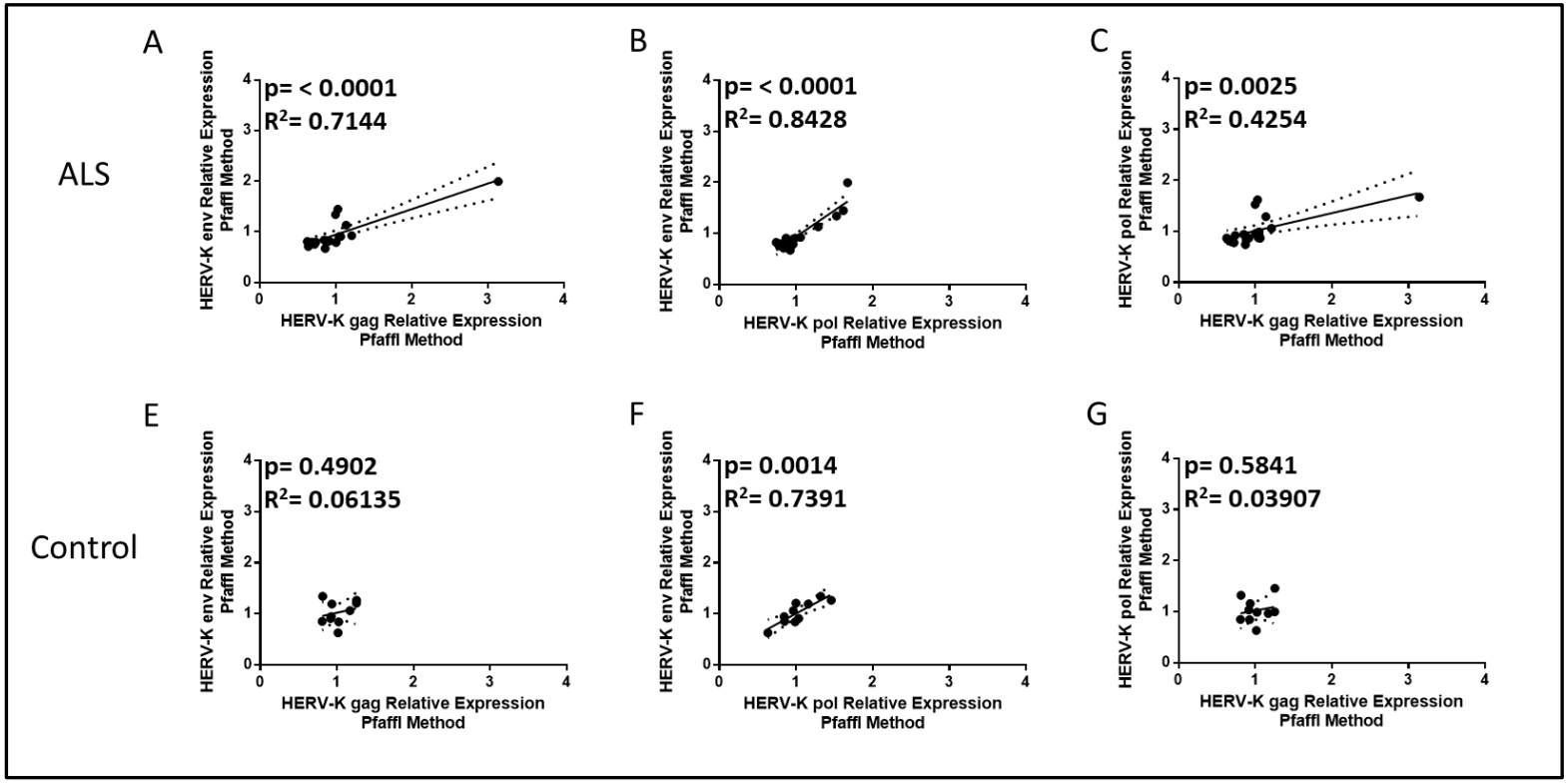


Figure S76. Correlation of HERV-K *gag, pol, env* & *RT* Transcript Expression Data from n=19 ALS and n=8 No-Cancer control Samples when normalised using Pfaffl Differential Expression Method.

The Figure above shows the correlation of HERV-K *gag, pol env* & *RT* differential expression data in ALS and non-ALS Postmortem Premotor cortex tissue samples. Graphs A & D shows comparison of HERV-K *env* & *gag* transcripts, B & E shows correlation of HERV-K *env* & *pol* transcripts and C & F shows data for the correlation of HERV-K *pol* & *gag* transcripts. This data was generated using GraphPad v8.0.

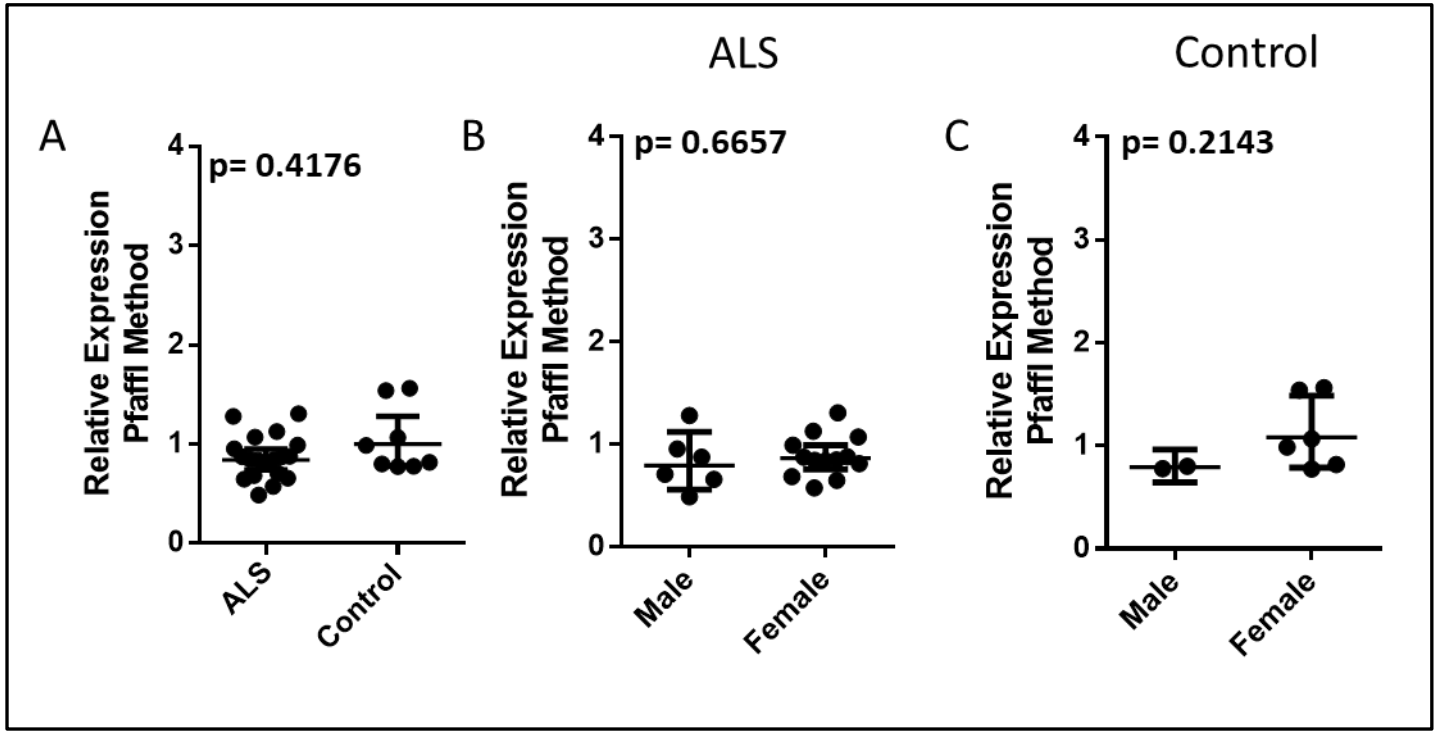


Figure S77. The effect of Disease status and Gender on HERV-W *env* Expression in n=19 ALS and n=8 No-Cancer control Postmortem Premotor Cortex Tissue Samples using Pfaffl Gene Expression Method.

The Figure above shows the differential expression data for HERV-W *env* transcripts between A) ALS and non-ALS Postmortem Premotor cortex tissue samples and the effect of gender in ALS and Controls (B&C). This data was generated using GraphPad v8.0.

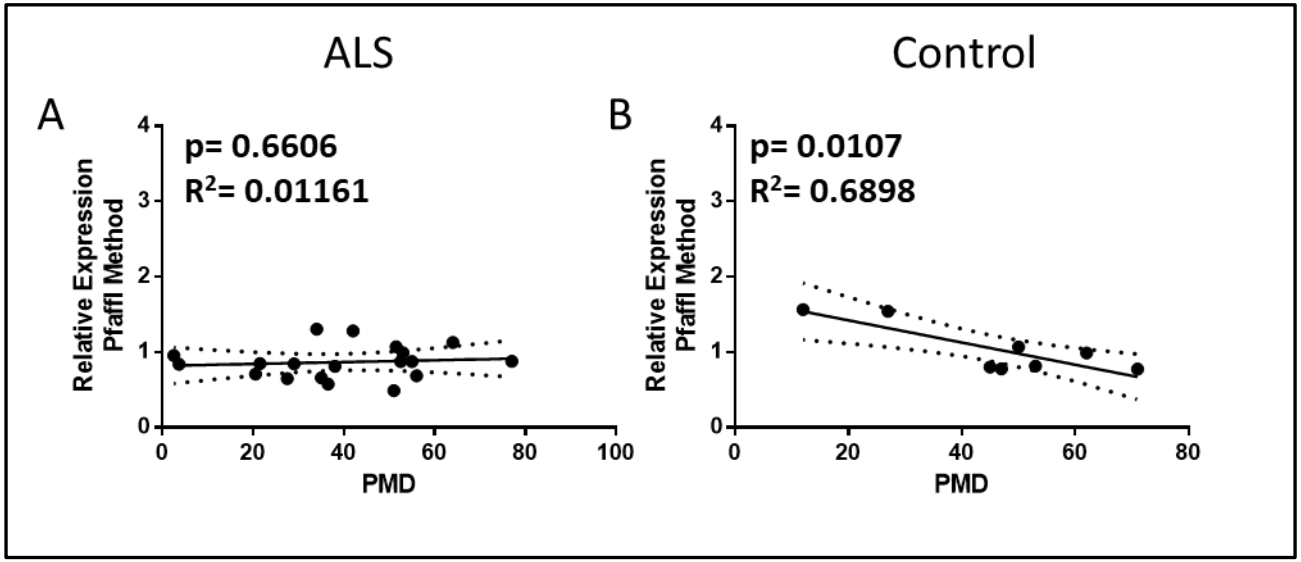


Figure S78. The effect of Postmortem Delay on HERV-W *env* Expression in n=19 ALS and n=8 No-Cancer control Postmortem Premotor Cortex Tissue Samples using Pfaffl Gene Expression Method.

The Figure above shows the effect of Postmortem Delay on HERV-W *env* transcript differential expression data for A) ALS and B) non-ALS Postmortem Premotor cortex tissue samples. This data was generated using GraphPad v8.0.

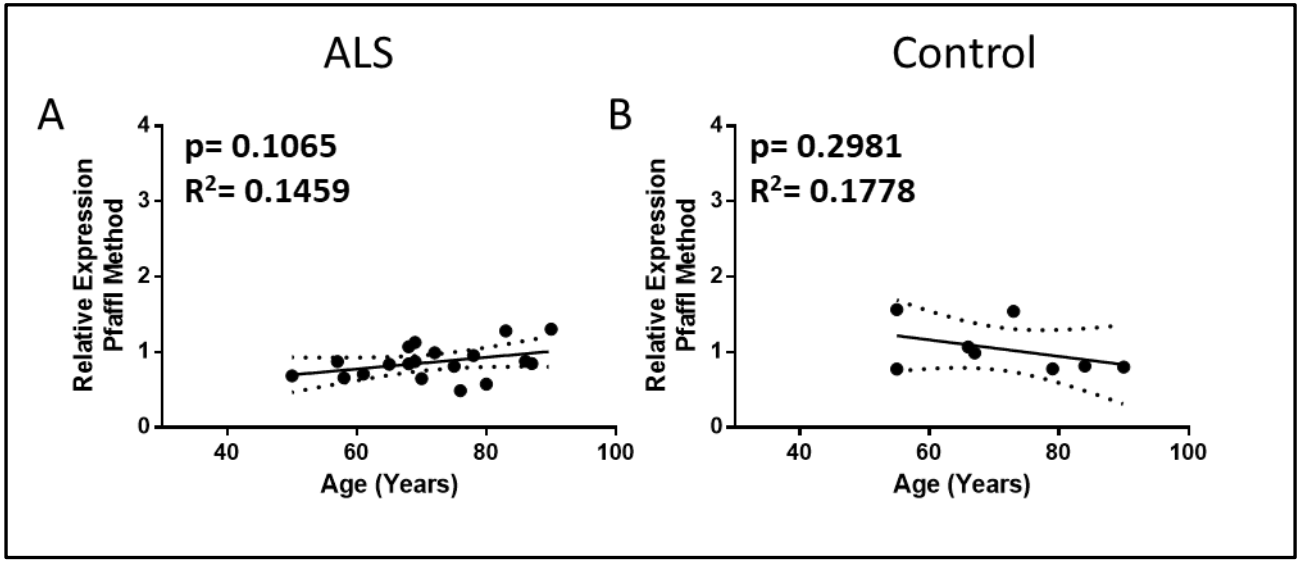


Figure S79. The effect of Age of Patient at time of Death on HERV-W *env* Expression in n=19 ALS and n=8 No-Cancer control Postmortem Premotor Cortex Tissue Samples using Pfaffl Gene Expression Method.

The Figure above shows the effect of patient age at time of death on HERV-W *env* transcript differential expression data for A) ALS and B) non-ALS Postmortem Premotor cortex tissue samples. This data was generated using GraphPad v8.0.

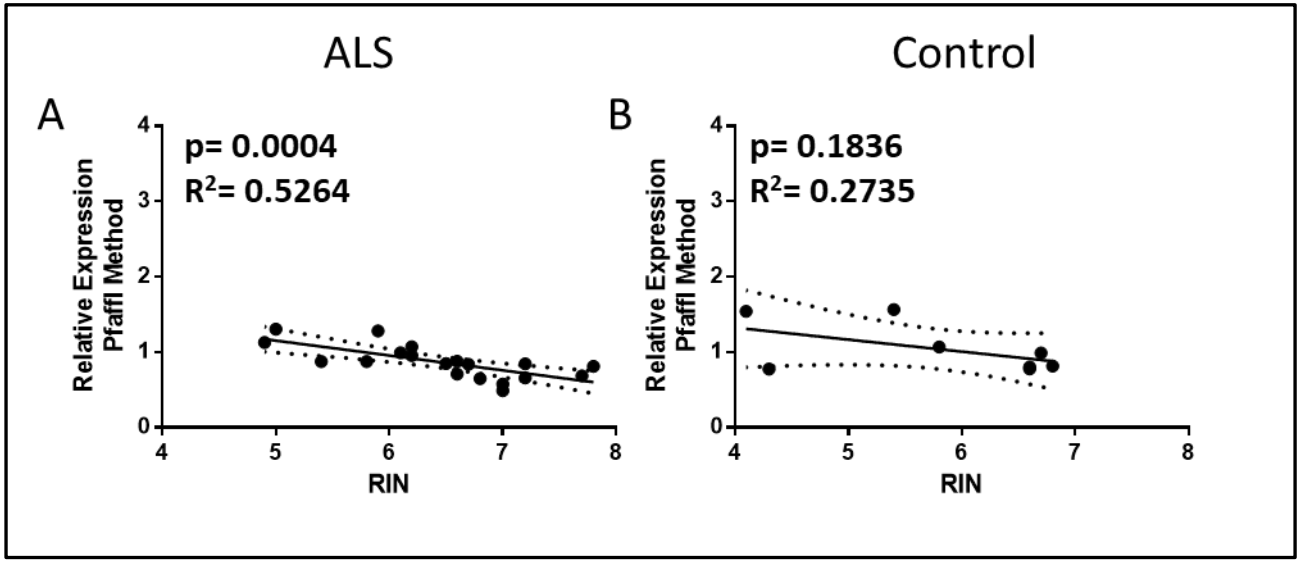


Figure S80. The effect of RNA Integrity on HERV-W *env* Expression in n=19 ALS and n=8 No-Cancer control Postmortem Premotor Cortex Tissue Samples using Pfaffl Gene Expression Method.

The Figure above shows the effect of RNA Integrity (RIN) on HERV-W *env* transcript differential expression data for A) ALS and B) non-ALS Postmortem Premotor cortex tissue samples. This data was generated using GraphPad v8.0.

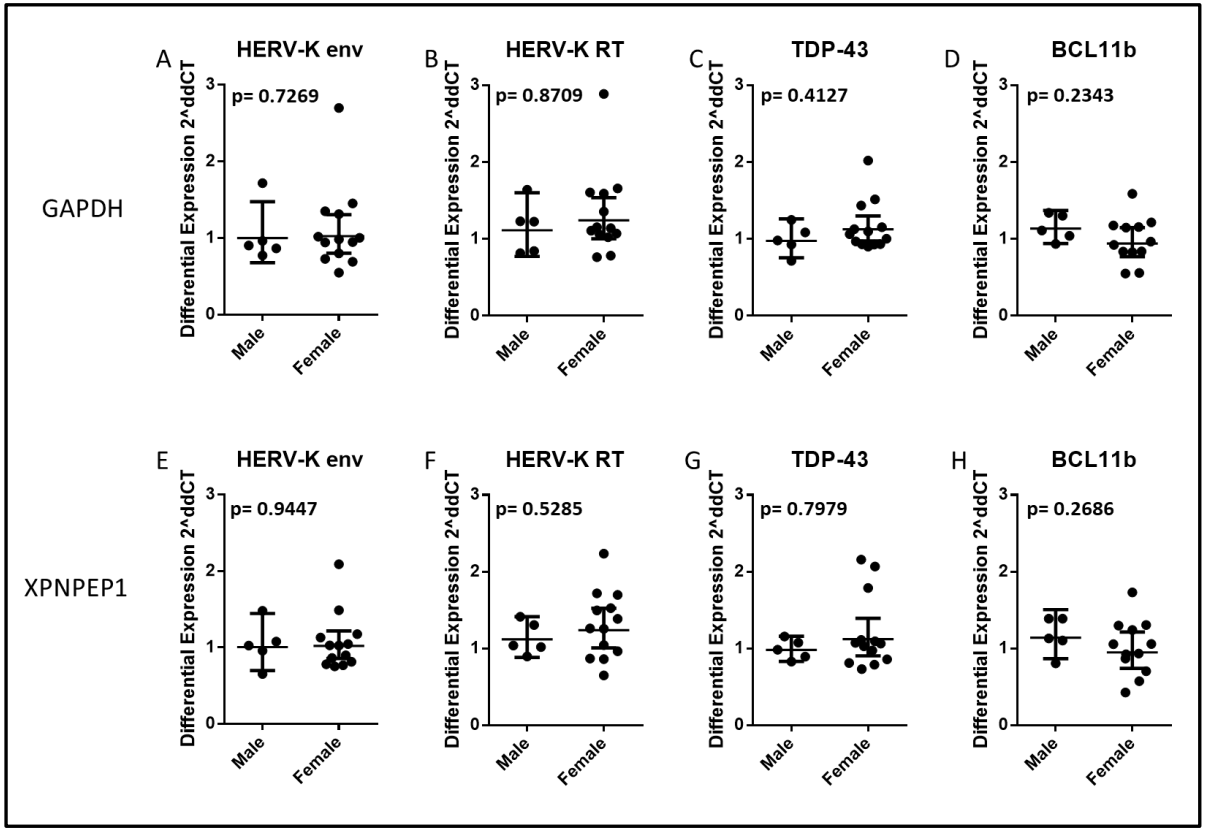


Figure S81. No significant correlation between gender, HERV-K *env* & *RT*, TDP-43 and BCL11b expression in Postmortem Premotor Cortex Tissue from ALS Patients

There is no significant difference in HERV-K env & RT, TDP-43 and BCL11b transcript expression between male and female groups when the data is normalised to either GAPDH (A, B, C and D). or when normalised to XPNPEP1 (E, F, G, and H). This graph utilises data from ΔΔCt Normalisation method.

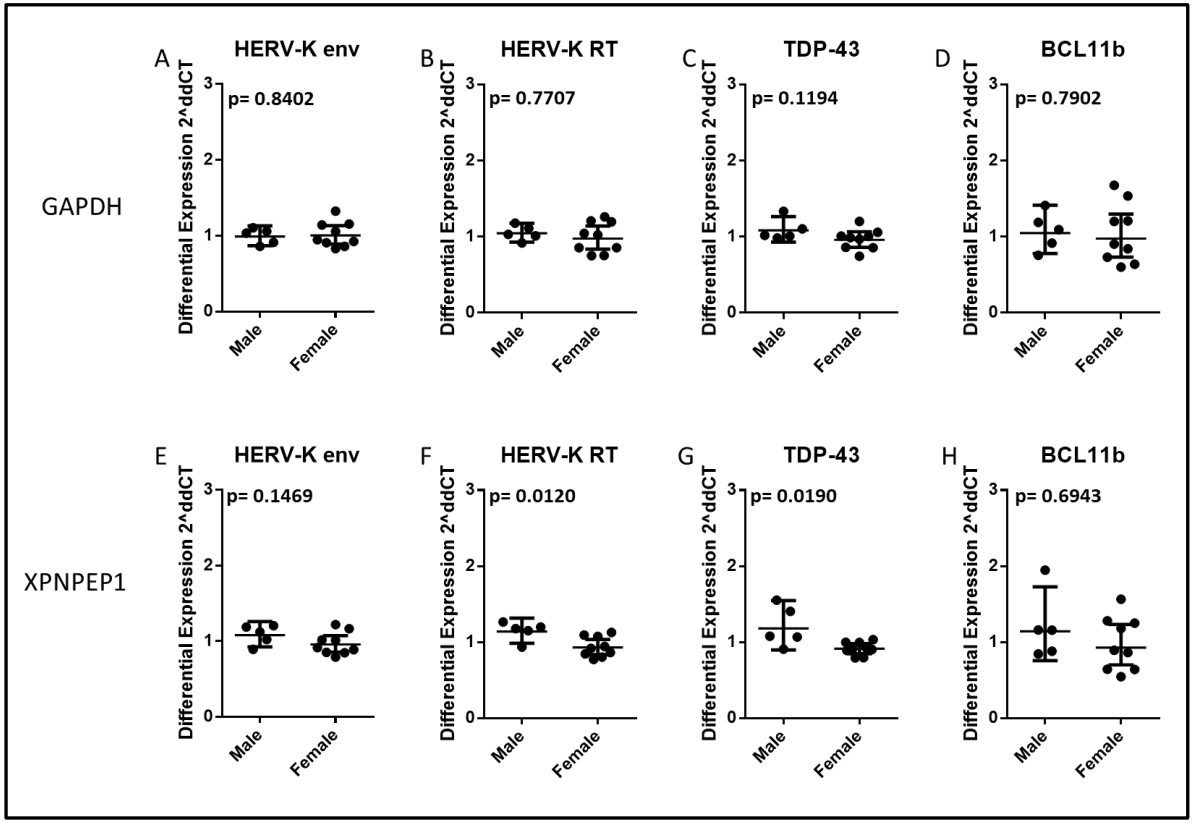


Figure S82. No significant correlation between gender HERV-K *env* & *RT*, TDP-43 and BCL11b expression in Postmortem Premotor Cortex Tissue from non-ALS control Patients

There is no significant difference in HERV-K env & RT, TDP-43 and BCL11b transcript expression between male and female groups when the data is normalised to either GAPDH (A, B, C and D). or when normalised to XPNPEP1 (E, F, G, and H). This graph utilises data from ΔΔCt Normalisation method.

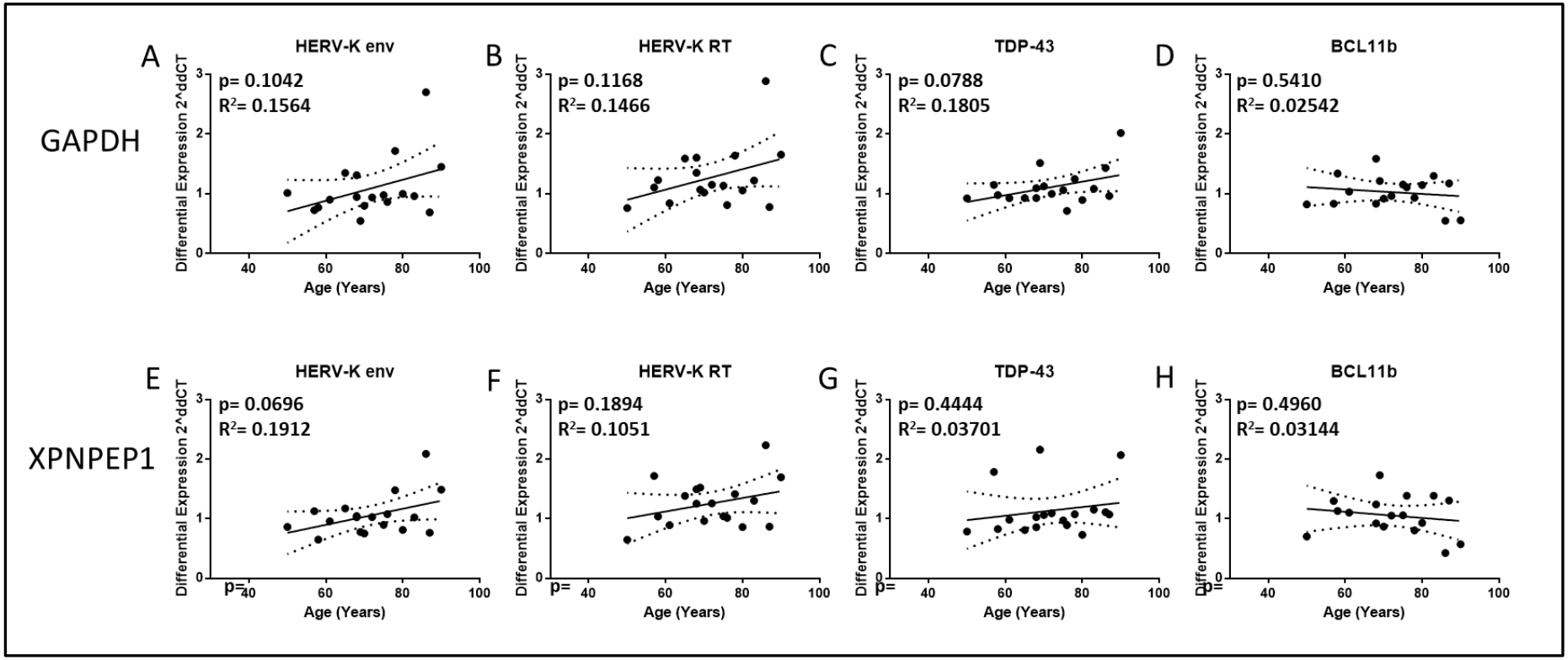


Figure S83. Effect of increasing age of patient at time of death on HERV-K *env* & *RT*, TDP-43 and BCL11b expression in ALS samples.

Data shown in the figure above were normalised against 2 separate reference genes, with A, B, C and D normalised against GAPDH and E, F, G and H normalised against XPNPEP1. Normalisation against different reference genes had no effect on results R^2 values were calculated in GraphPad Prism v8.0. This graph utilises data from ΔΔCt Normalisation method.

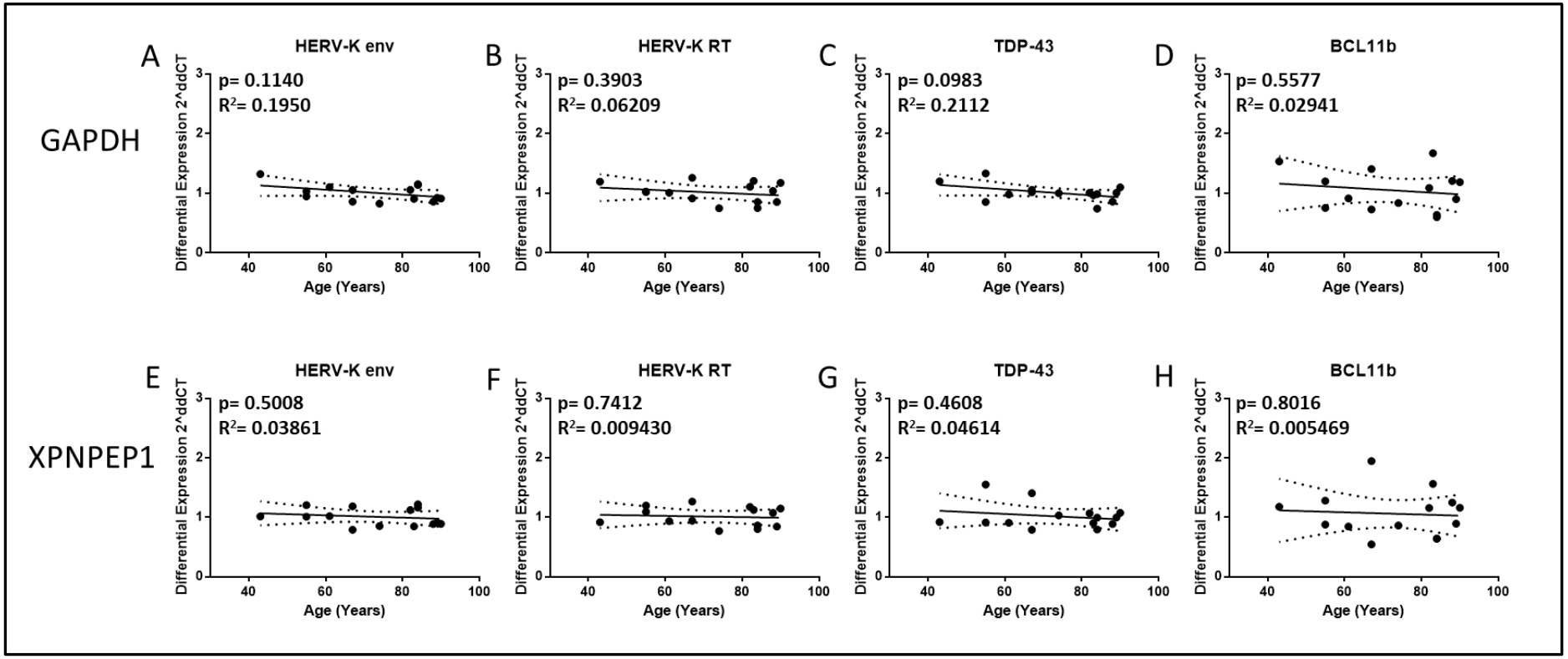


Figure S84. Effect of increasing age of patient at time of death on HERV-K *env* & *RT*, TDP-43 and BCL11b expression in Non-ALS Control samples.

Data shown in the figure above were normalised against 2 separate reference genes, with A, B, C and D normalised against GAPDH and E, F, G and H normalised against XPNPEP1. Normalisation against different reference genes had no effect on results R^2 values were calculated in GraphPad Prism v8.0. This graph utilises data from ΔΔCt Normalisation method.

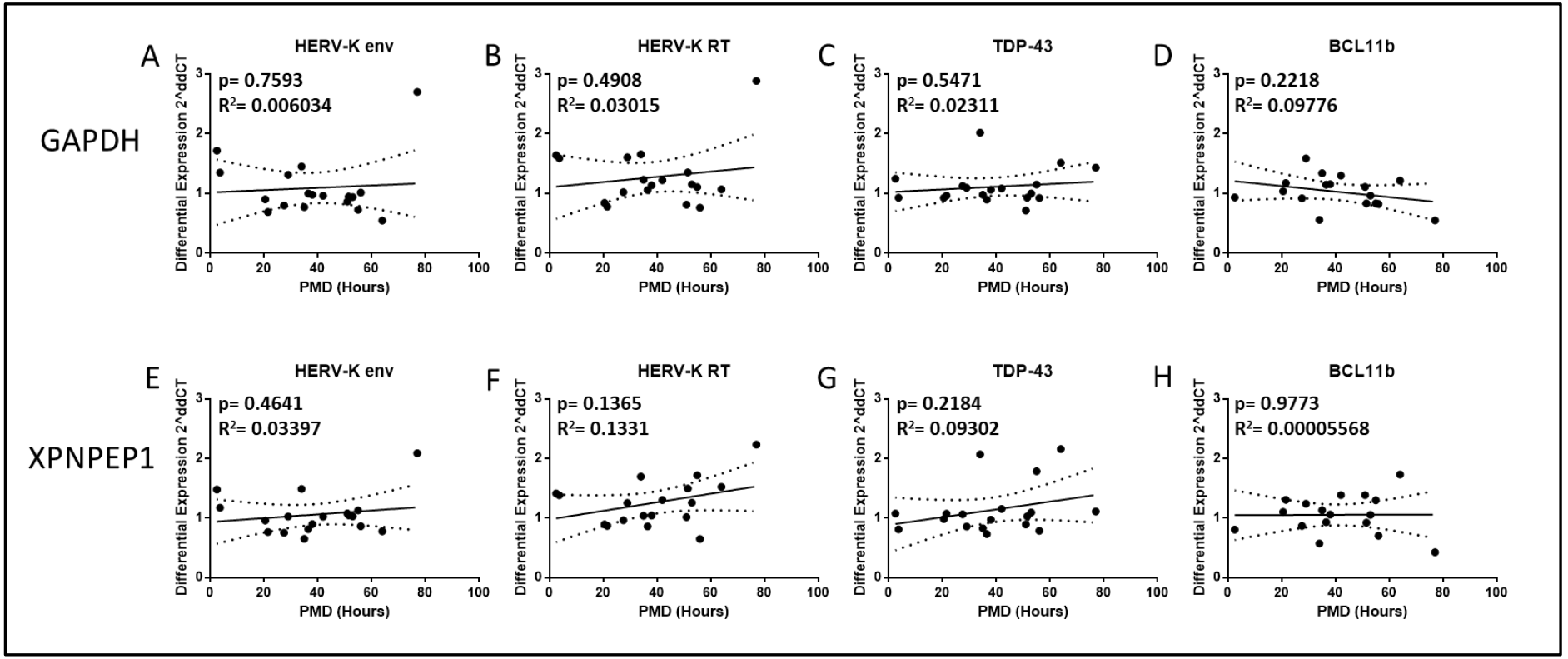


Figure S85. Effect of PMD on HERV-K *env* & *RT*, TDP-43 and BCL11b expression when normalised to GAPDH and XPNPEP1 in ALS Patient Tissue. The figure above shows the effect of PMD on HERV-K *env & RT,* TDP-43 and BCL11b genomic regions. A, B, C and D are normalised to GAPDH which shows no or only very slight effect of PMD on HERV-K pol expression increasing PMD (B). There was no correlation to HERV-K expression when the data was normalised to XPNPEP1 (E, F, G and H). R^2 values and p values were calculated in GraphPad Prism v8.0. This graph utilises data from ΔΔCt Normalisation method.

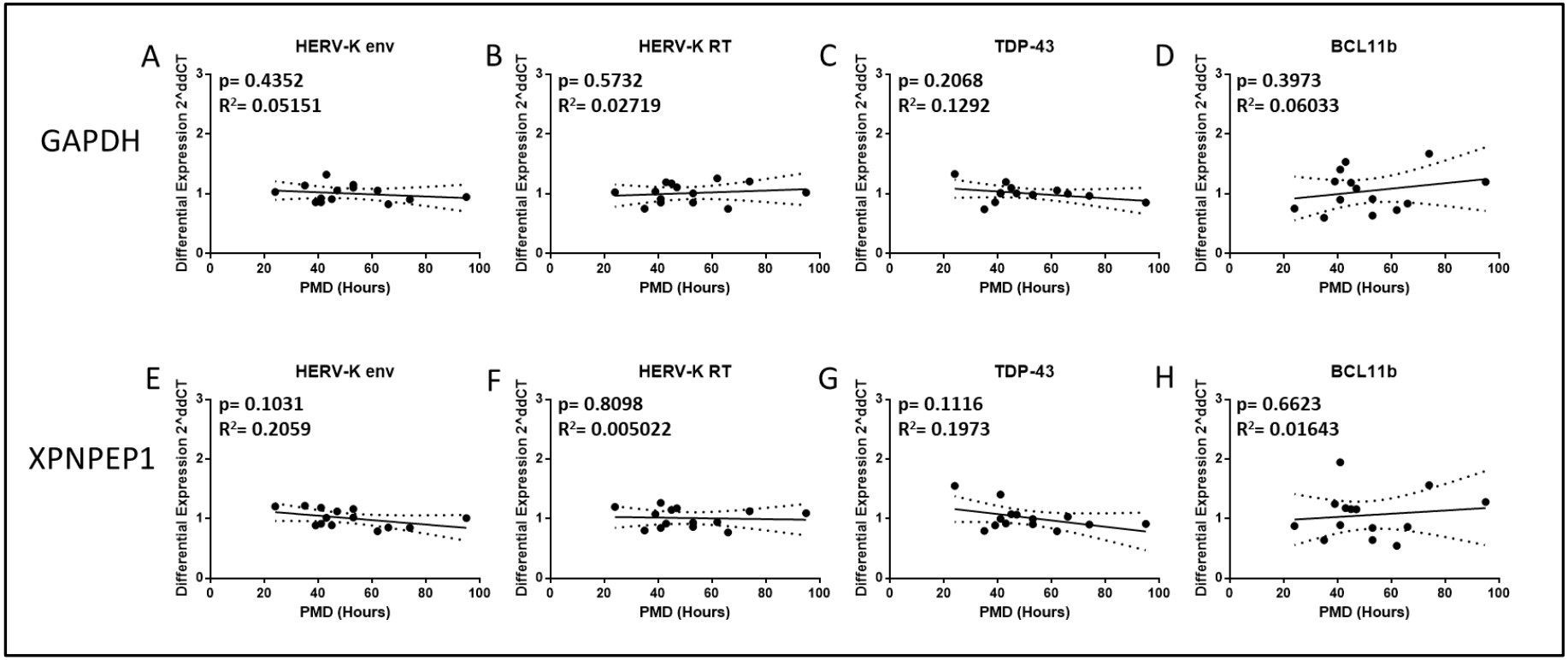


Figure S86. Effect of PMD on HERV-K *env* & *RT*, TDP-43 and BCL11b expression when normalised to GAPDH and XPNPEP1 in Non-ALS Control Patient Tissue.

The figure above shows the effect of PMD on HERV-K *env* & *RT*, TDP-43 and BCL11b genomic regions. A, B, C and D are normalised to GAPDH which shows no or only very slight effect of PMD on HERV-K pol expression increasing PMD (B). There was no correlation to HERV-K expression when the data was normalised to XPNPEP1 (E, F, G and H). R^2 values and p values were calculated in GraphPad Prism v8.0. This graph utilises data from ΔΔCt Normalisation method.

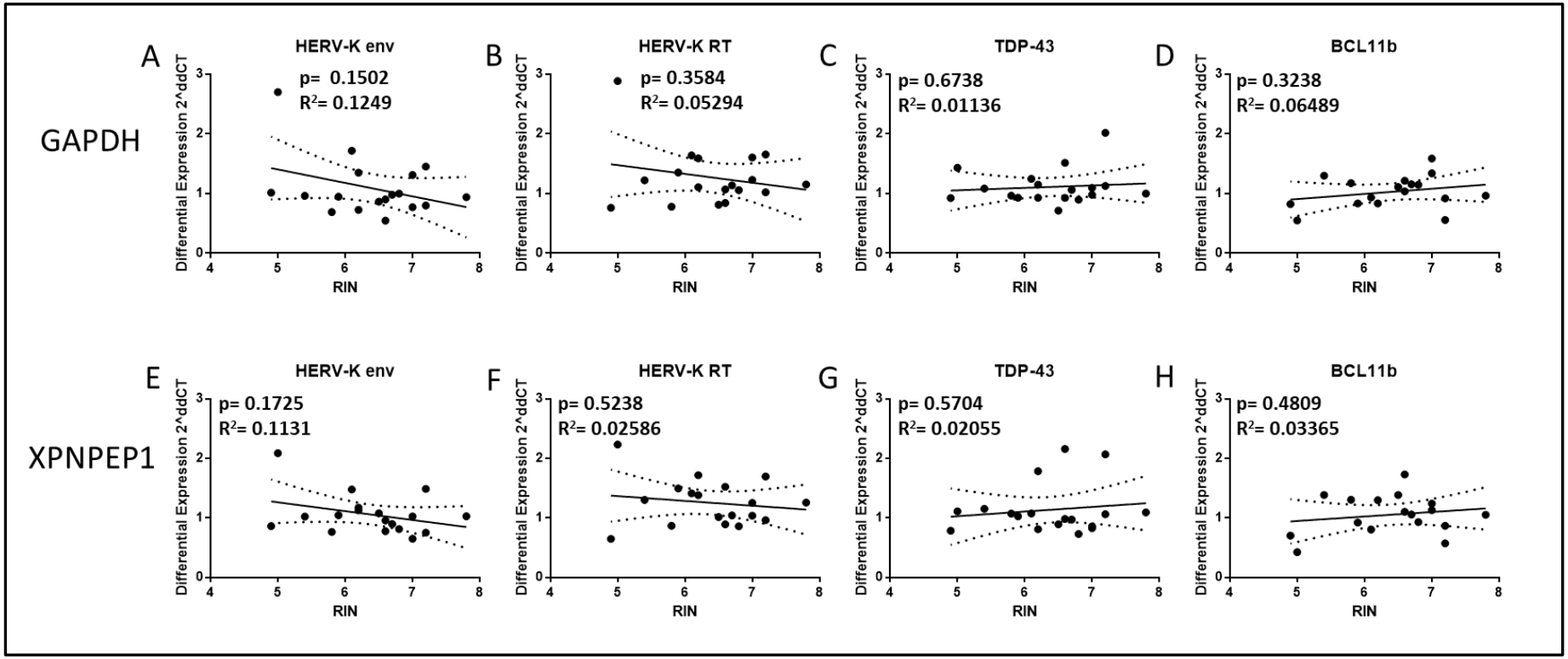


Figure S87. Effect of RNA integrity value on HERV-K *env* & *RT*, TDP-43 and BCL11b expression In ALS Patient Tissue Samples.

The data displayed in the graph above shows HERV-K *env & RT*, TDP-43 and BCL11b expression when normalized against 2 separate reference genes, GAPDH and XPNPEP1. Linear regression graphs A, B, C and D show expression data normalised against GAPDH while E, F, G, and H display expression data normalised against XPNPEP1. R^2 values and P values were calculated in GraphPad Prism v8.0. All p values were not significant. This graph utilises data from ΔΔCt Normalisation method.

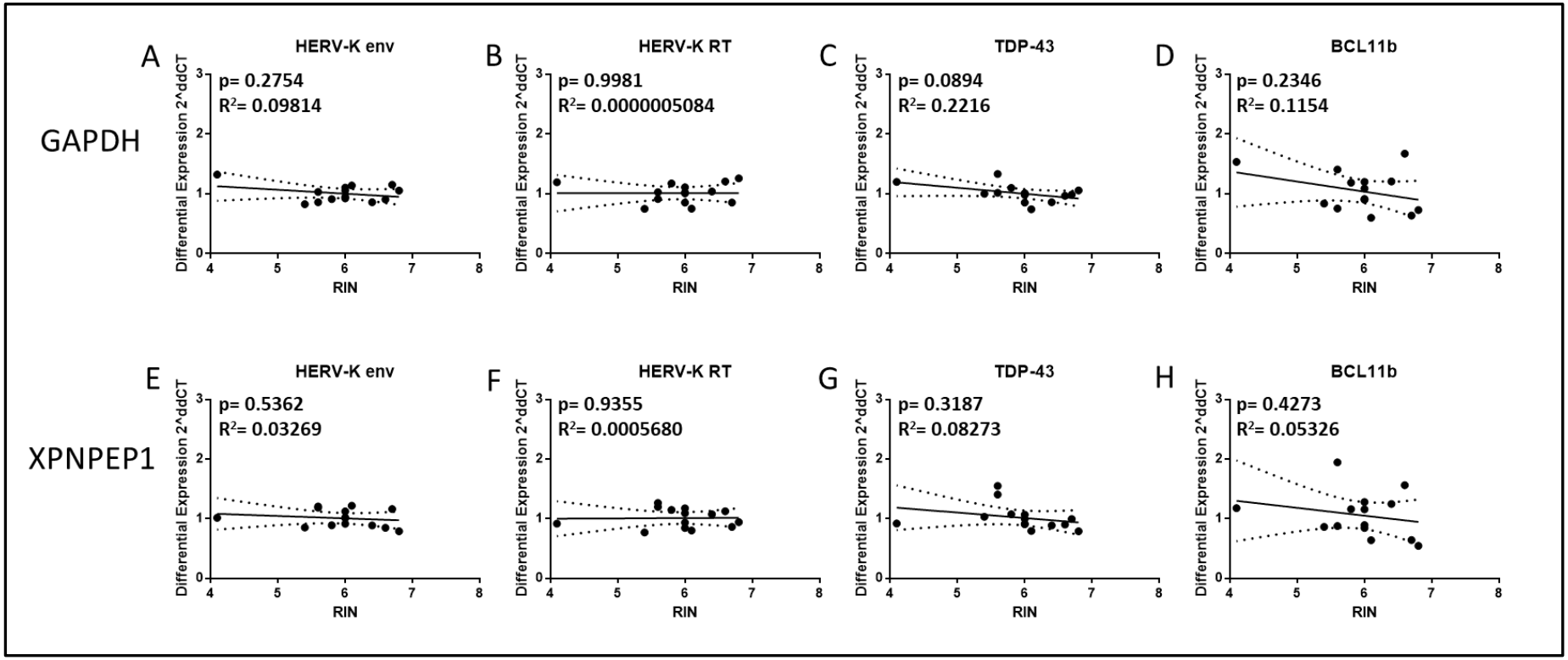


Figure S88. Effect of RNA integrity value on HERV-K *env* & *RT*, TDP-43 and BCL11b expression In Non-ALS Control Patient Tissue Samples.
The data displayed in the graph above shows HERV-K *env* & *RT*, TDP-43 and BCL11b expression when normalized against 2 separate reference genes, GAPDH and XPNPEP1. Linear regression graphs A, B, C and D show expression data normalised against GAPDH while E, F, G, and H display expression data normalised against XPNPEP1. R^2 values and P values were calculated in GraphPad Prism v8.0. All p values were not significant. This graph utilises data from ΔΔCt Normalisation method.

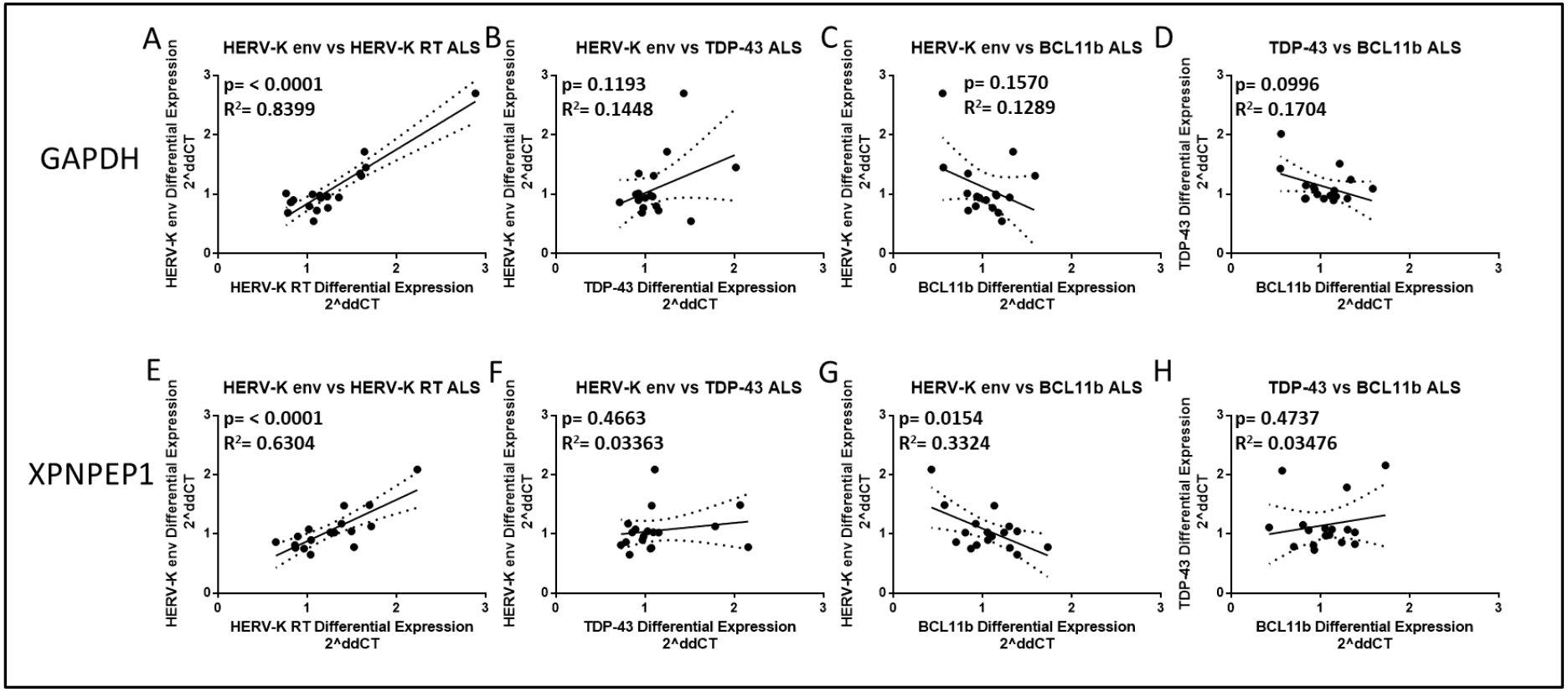


Figure S89. Correlation of HERV-K *env* & *RT*, TDP-43 and BCL11b Expression Data from n=18 ALS Samples using ΔΔCt Method

The Figure above shows the correlation of HERV-K *env* & *RT*, TDP-43 and BCL11b differential expression data from n=19 ALS Cases when normalised to GAPDH or XPNPEP1. Graphs A & E shows comparison of HERV-K *env* & *RT* transcripts, B & F shows correlation of HERV-K *env* & TDP-43 transcripts, C & G shows data for the correlation of HERV-K *env* & BCL11b transcripts and D&H shows the correlation between TDP-43 and BCL11b. This data was generated using GraphPad v8.0.

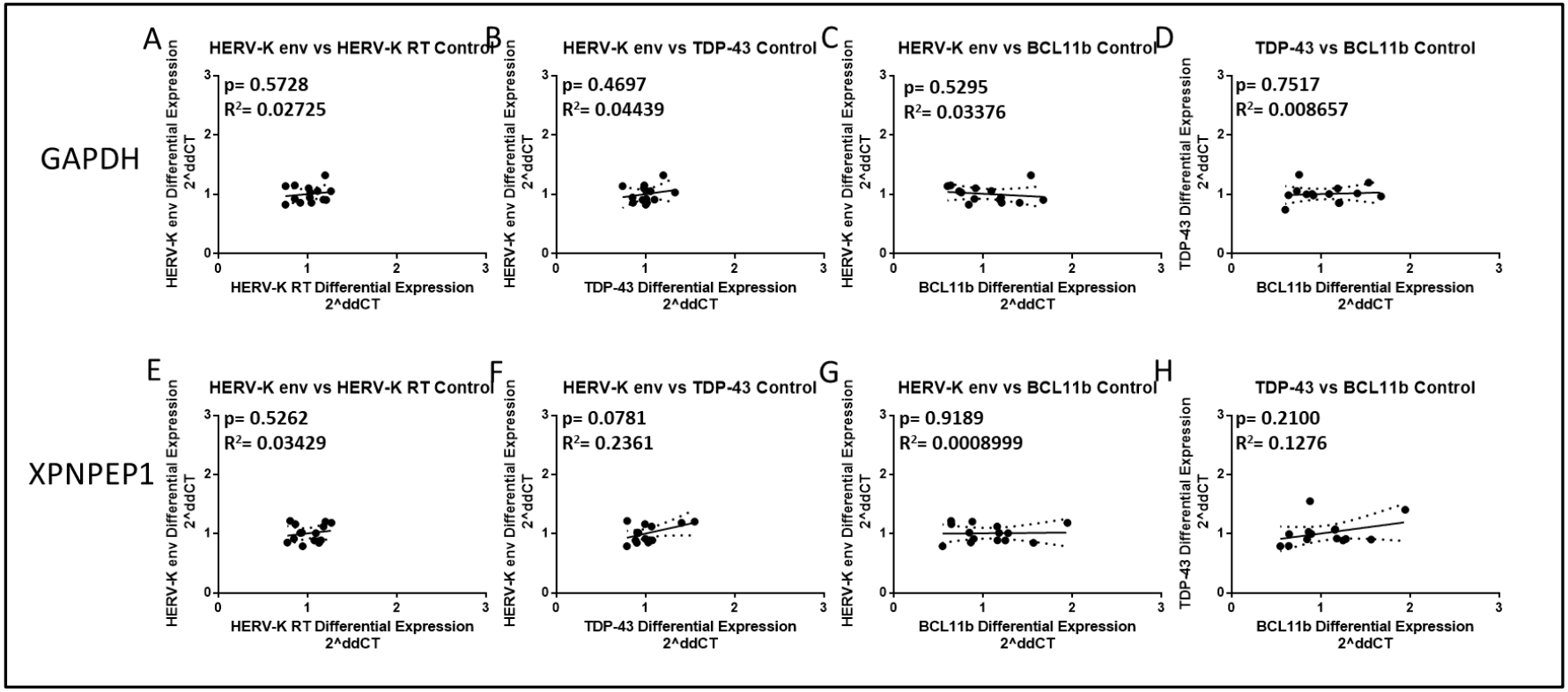


Figure S90. Correlation of HERV-K *env* & *RT*, TDP-43 and BCL11b Expression Data from n=14 Non-ALS Control Samples using ΔΔCt Method

The Figure above shows the correlation of HERV-K *env* & *RT*, TDP-43 and BCL11b differential expression data from n=19 ALS Cases when normalised to GAPDH or XPNPEP1. Graphs A & E shows comparison of HERV-K *env* & *RT* transcripts, B & F shows correlation of HERV-K *env* & TDP-43 transcripts, C & G shows data for the correlation of HERV-K *env* & BCL11b transcripts and D&H shows the correlation between TDP-43 and BCL11b. This data was generated using GraphPad v8.0.

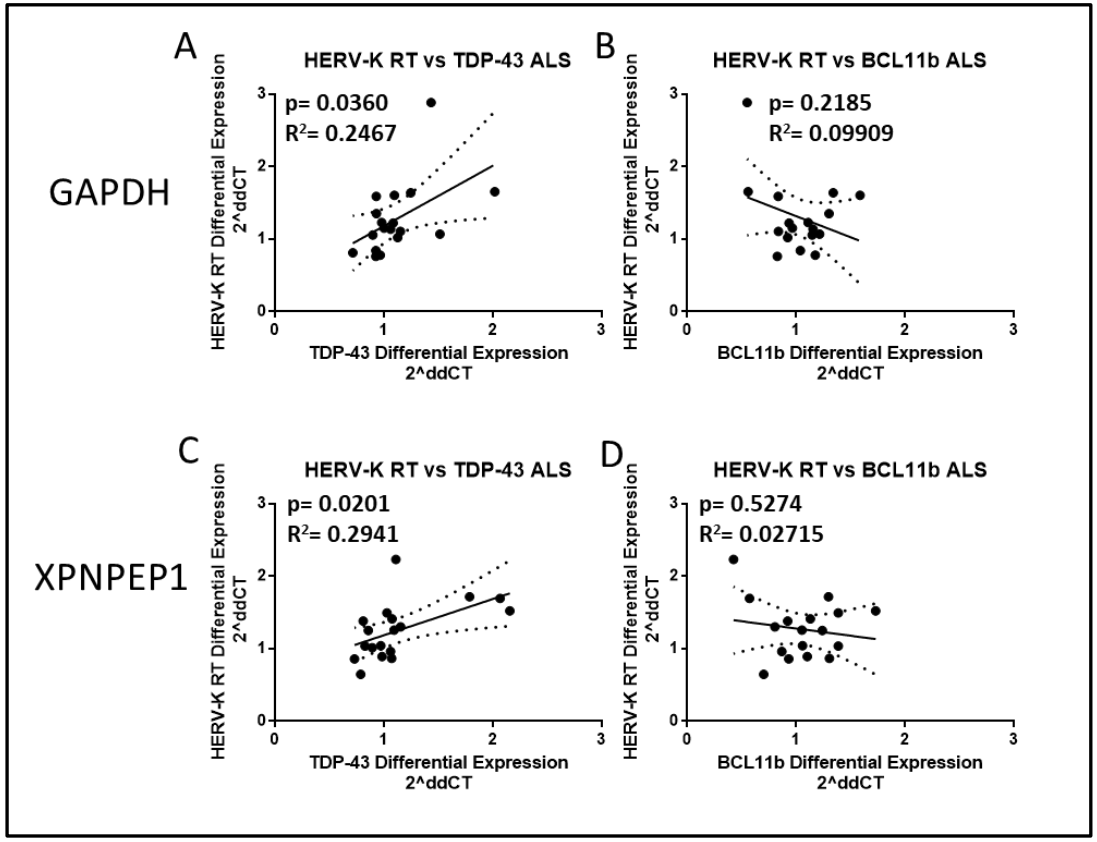


Figure S91. Correlation of HERV-K *RT*, BCL11b and TDP-43 Data from n=18 ALS Samples using ΔΔCt Method

The Figure above shows the correlation of HERV-K *RT* TDP-43 and BCL11b differential expression data from n=19 ALS Cases when normalised to GAPDH or XPNPEP1. Graphs A & C shows comparison of HERV-K *RT* and TDP-43 transcripts and B & D shows correlation of HERV-K *RT* & BCL11b. This data was generated using GraphPad v8.0.

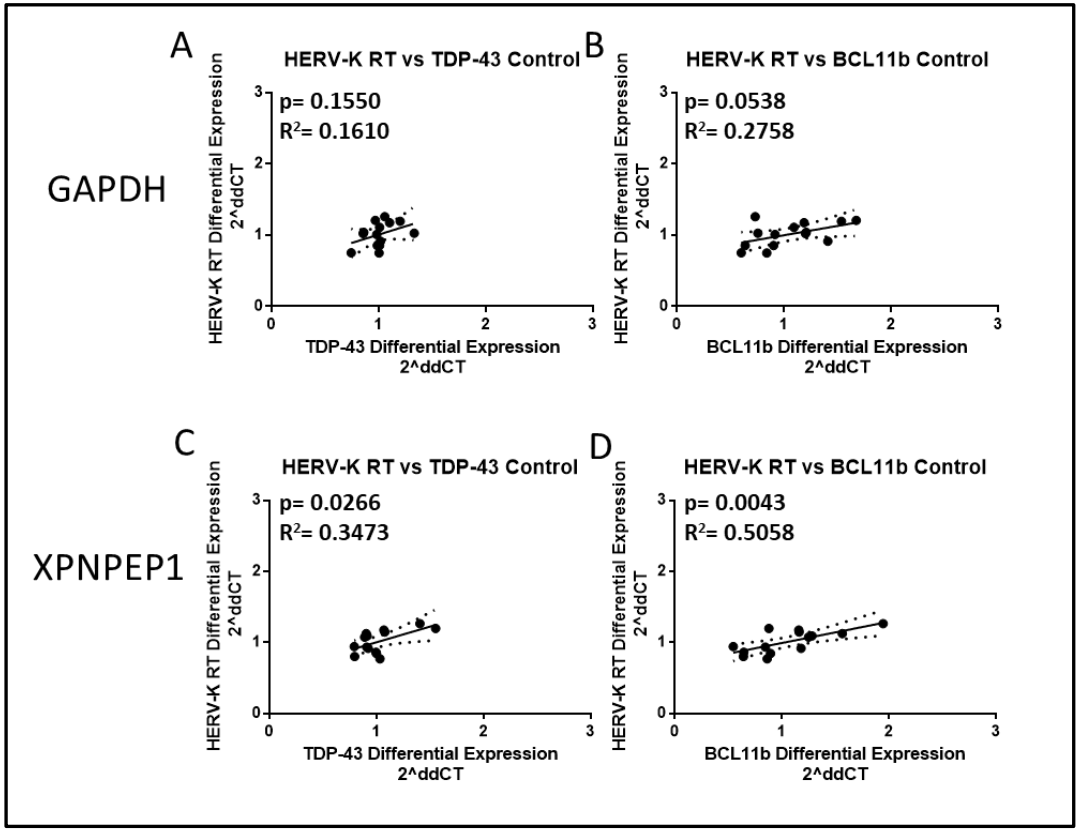


Figure S92. Correlation of HERV-K *RT*, BCL11b TDP-43 Expression Data from n=14 Non-ALS Control Samples using ΔΔCt Method

The Figure above shows the correlation of HERV-K *RT*, TDP-43 and BCL11b differential expression data from n=19 ALS Cases when normalised to GAPDH or XPNPEP1. Graphs A & C shows comparison of HERV-K *RT* and TDP-43 transcripts and B & D shows correlation of HERV-K *RT* & BCL11b. This data was generated using GraphPad v8.0.

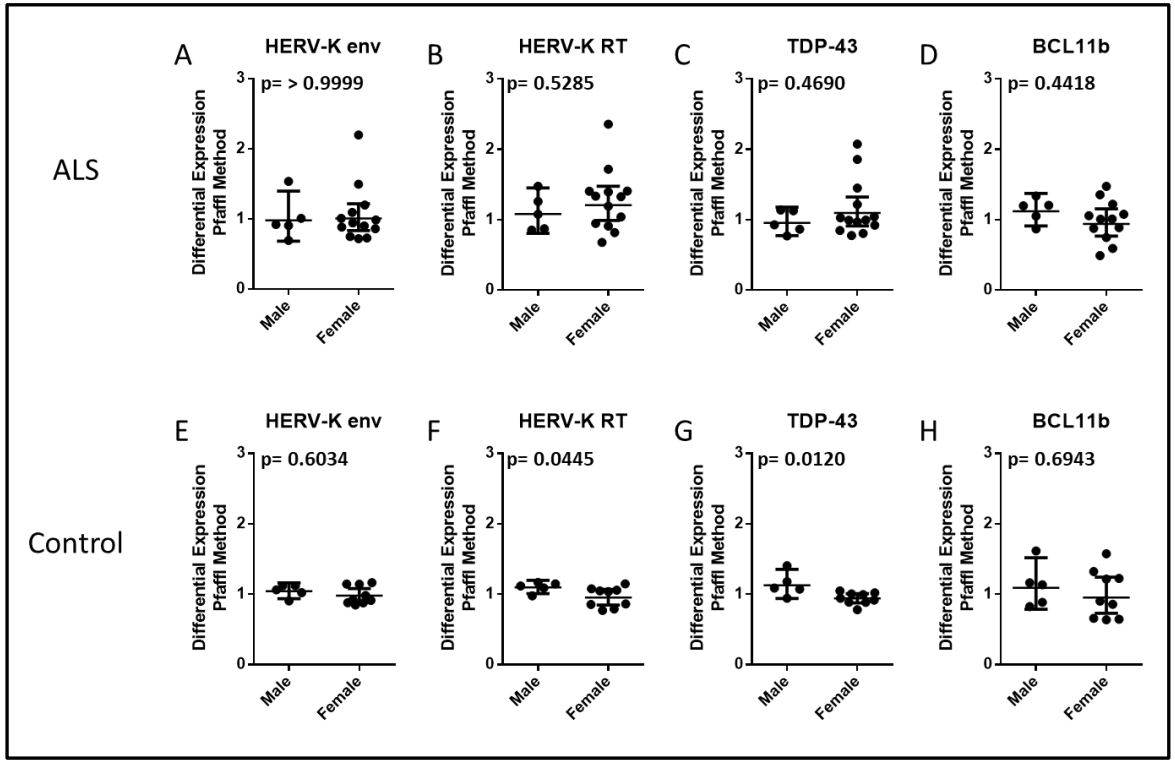


Figure S93. Effect of Sex on HERV-K *env* & *RT*, TDP-43 and BCL11b Expression Data from n=18 ALS and n=14 No-Cancer control Samples when normalised using Pfaffl Differential Expression Method.

The Figure above shows the differential expression data for the effect of gender in ALS and non-ALS Postmortem Premotor cortex tissue samples. Graphs A & E show data for HERV-K *gag* transcripts, B & F show data for HERV-K *pol* transcripts, C & G show data for HERV-K *env* transcripts and D & H show data for HERV-K *env* transcripts. This data was generated using GraphPad v8.0.

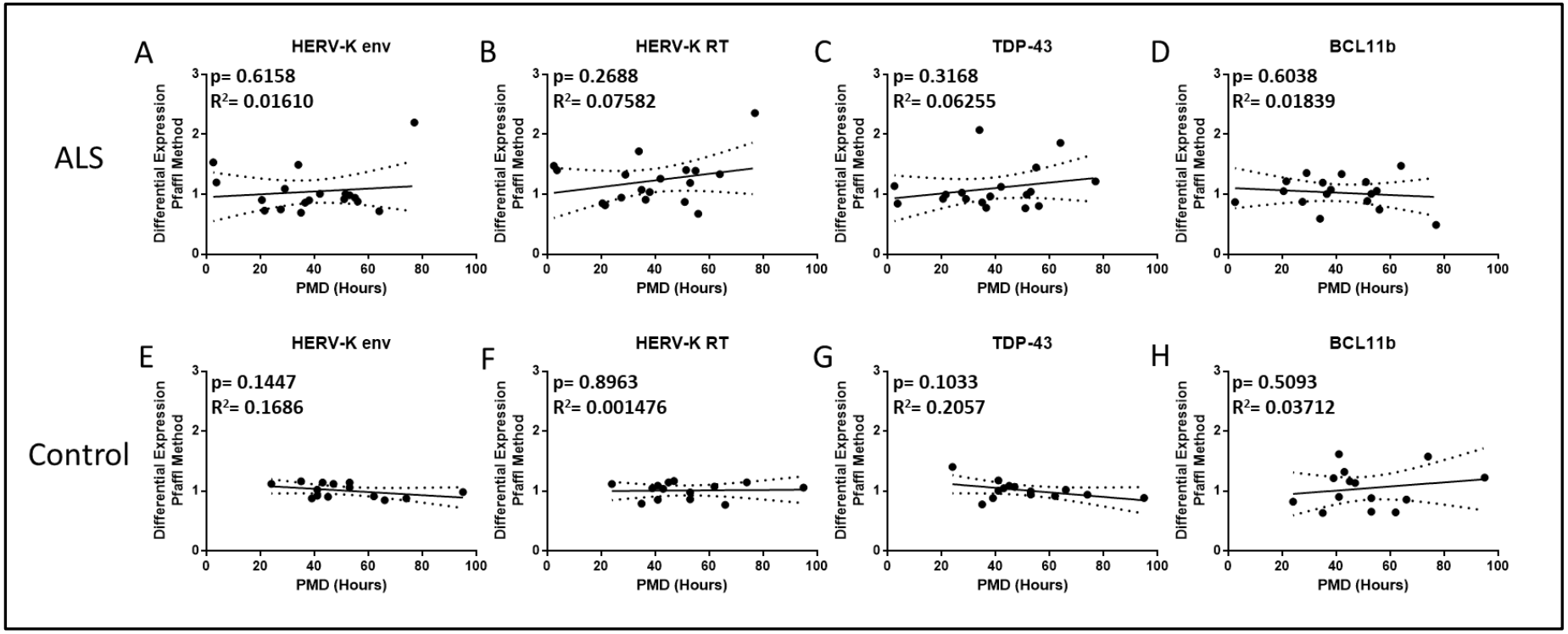


Figure S94. Effect of Postmortem Delay on HERV-K *env* & *RT*, TDP-43 and BCL11b Expression Data from n=18 ALS and n=14 No-Cancer control Samples when normalised using Pfaffl Differential Expression Method.

The Figure above shows the differential expression data for the effect of postmortem delay in ALS and non-ALS Postmortem Premotor cortex tissue samples. Graphs A & E show data for HERV-K *gag* transcripts, B & F show data for HERV-K *pol* transcripts, C & G show data for HERV-K *env* transcripts and D & H show data for HERV-K *env* transcripts. This data was generated using GraphPad v8.0.

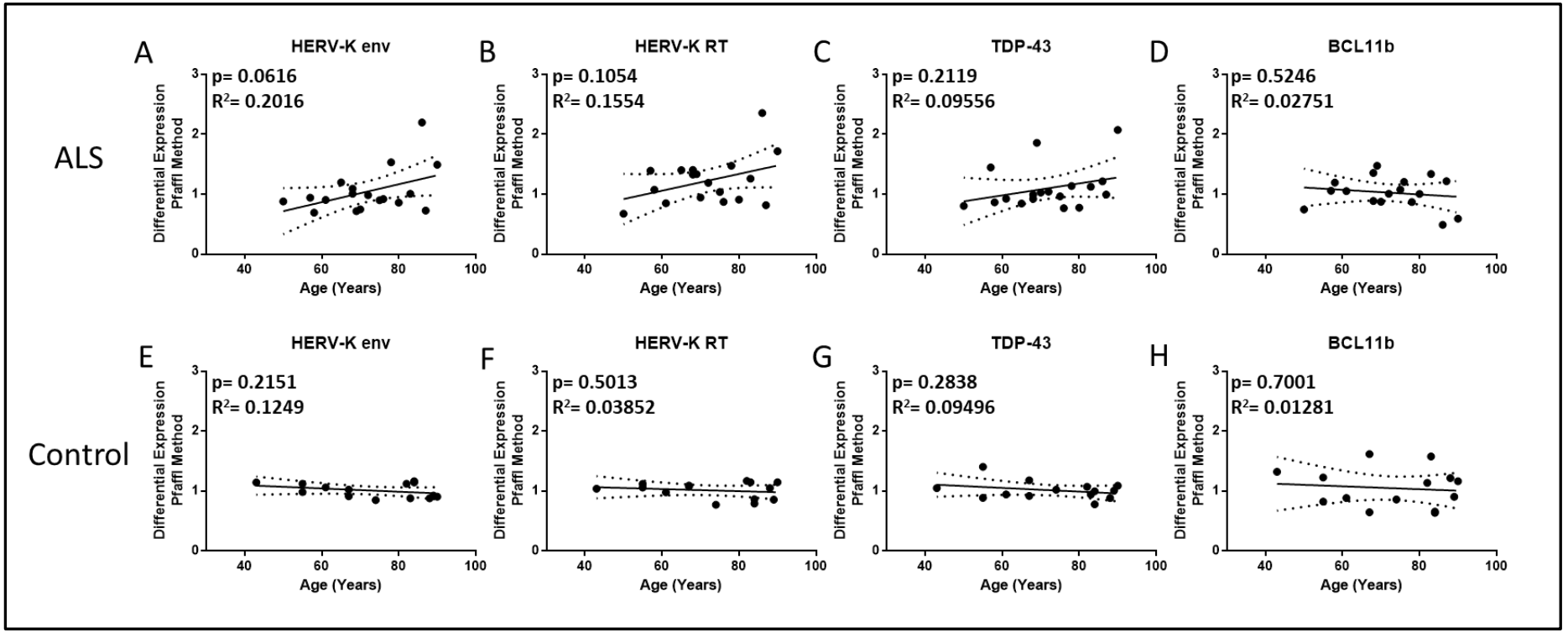


Figure S95. Effect of Age of Patient at time of Death on HERV-K *env* & *RT*, TDP-43 and BCL11b Expression Data from n=18 ALS and n=14 No-Cancer control Samples when normalised using Pfaffl Differential Expression Method.

The Figure above shows the differential expression data for the effect of patient age at time of death in ALS and non-ALS Postmortem Premotor cortex tissue samples. Graphs A & E show data for HERV-K *gag* transcripts, B & F show data for HERV-K *pol* transcripts, C & G show data for HERV-K *env* transcripts and D & H show data for HERV-K *env* transcripts. This data was generated using GraphPad v8.0.

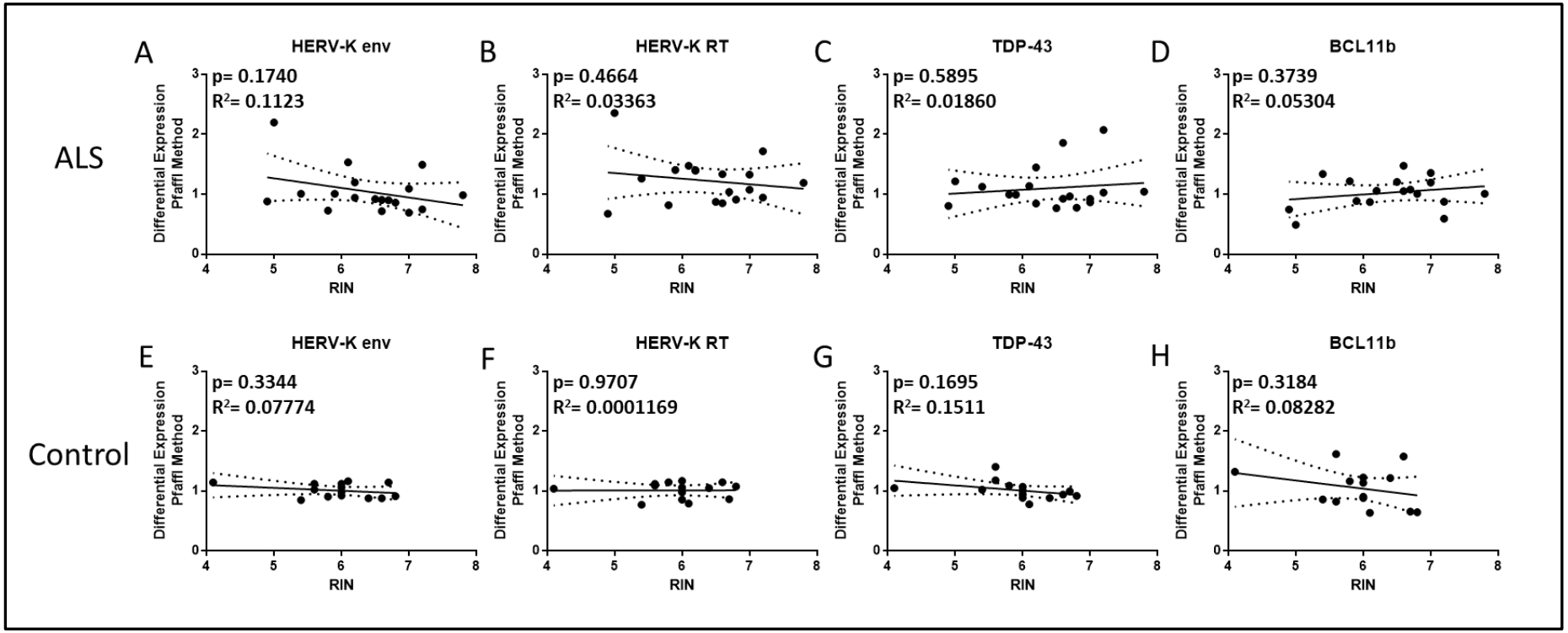


Figure S96. Effect of RNA Integrity Value on HERV-K *env* & *RT*, TDP-43 and BCL11b Expression Data from n=18 ALS and n=14 No-Cancer control Samples when normalised using Pfaffl Differential Expression Method.

The Figure above shows the differential expression data for the effect of patient age at time of death in ALS and non-ALS Postmortem Premotor cortex tissue samples. Graphs A & E show data for HERV-K *gag* transcripts, B & F show data for HERV-K *pol* transcripts, C & G show data for HERV-K *env* transcripts and D & H show data for HERV-K *env* transcripts. This data was generated using GraphPad v8.0.

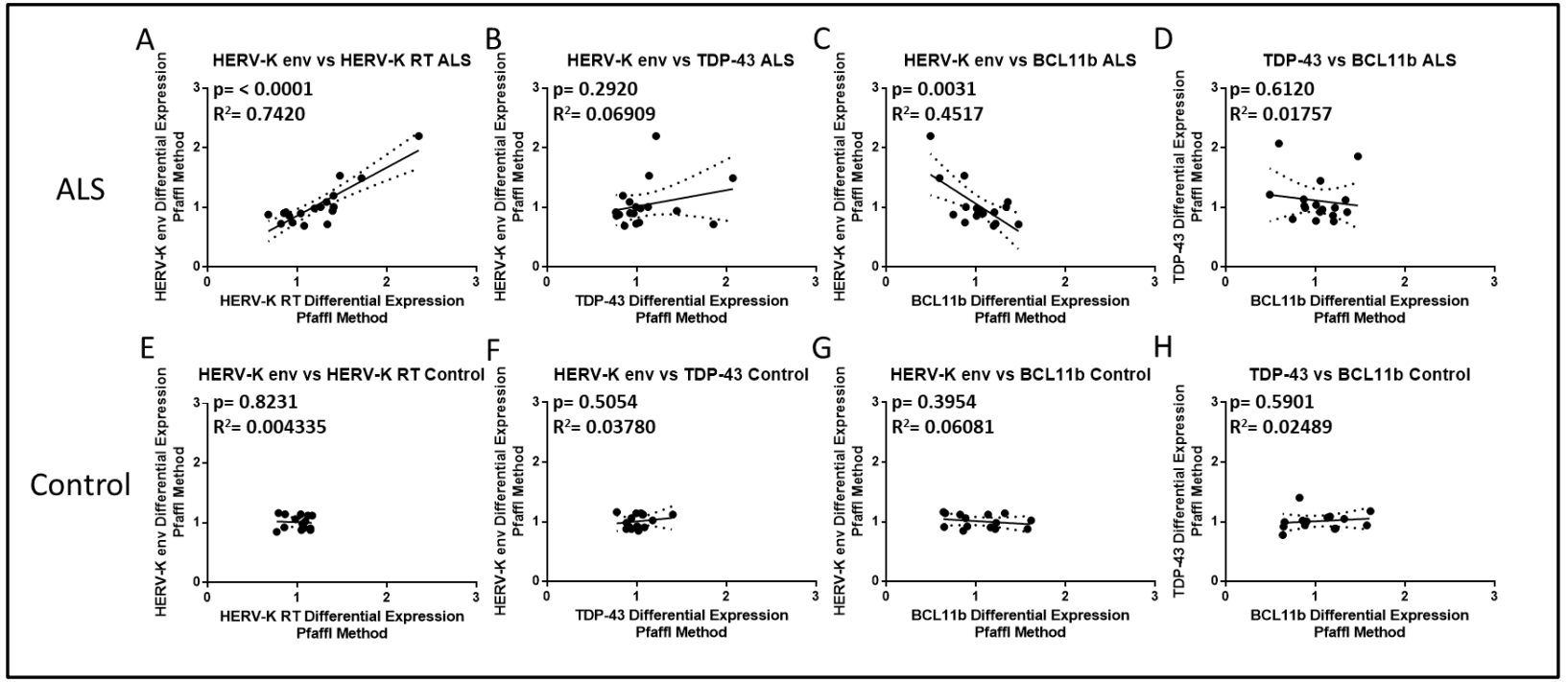


Figure S97. Correlation of HERV-K *env* & *RT*, TDP-43 and BCL11b Expression Data from n=18 ALS and n=14 No-Cancer control Samples when normalised using Pfaffl Differential Expression Method.

The Figure above shows the correlation of HERV-K *RT*, TDP-43 and BCL11b differential expression data in ALS and non-ALS Postmortem Premotor cortex tissue samples. Graphs A & E shows comparison of HERV-K *env* & *RT* transcripts, B & F shows correlation of HERV-K *env* & TDP-43 transcripts, C & G shows data for the correlation of HERV-K *env* & BCL11b transcripts and D&H shows the correlation between TDP-43 and BCL11b. This data was generated using GraphPad v8.0.

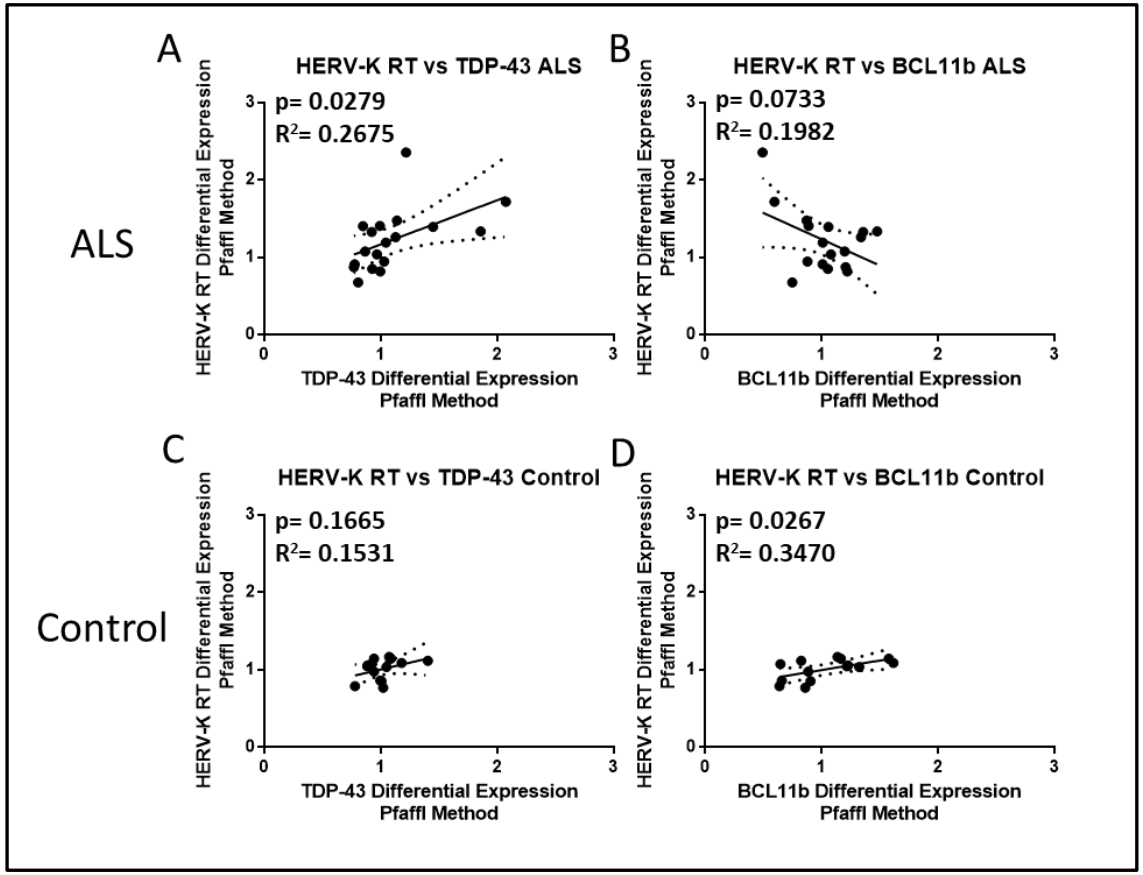


Figure S98. Correlation of HERV-K *RT*, BCL11b and TDP-43 Data from n=18 ALS and from n=14 Non-ALS Control Samples using Pfaffl Method

The Figure above shows the correlation of HERV-K *RT,* TDP-43 and BCL11b differential expression data from n=19 ALS Cases when normalised to GAPDH or XPNPEP1. Graphs A & C shows comparison of HERV-K *RT* and TDP-43 transcripts and B & D shows correlation of HERV-K *RT* & BCL11b. This data was generated using GraphPad v8.0.

Supplementary Table S4. Amplification efficiency Ct means for GAPDH, XPNPEP1, HERV-W *env* and HERV-K primer targets.

Efficiency for most primer sets was determined with 4-fold serial dilutions of cDNA using ALS (A151/10) and non-ALS control (A292/09) samples. Dilutions in Red were not included in efficiency calculations due to high standard deviation (SD) values (greater than 0.3). Efficiency was determined in Microsoft Excel using the slope of the generated standard curve and the following equation: =(10^(-1/Slope)-1)*100

Primer Gene Target	Premotor Cortex Tissue Sample ID	Clinical Status	Undiluted cDNA (Mean Ct)	1:4 Diluted cDNA (Mean Ct)	1:16 Diluted cDNA (Mean Ct)	1:64 Diluted cDNA (Mean Ct)	1:256 Diluted cDNA (Mean Ct)	1:1024 Diluted cDNA (Mean Ct)
XPNPEP1	A151/10	ALS	22.992	25.105	26.988	29.090	31.357	32.840
XPNPEP1	A292/09	Non-ALS Control	23.398	25.363	27.466	29.804	31.566	34.304
GAPDH	A151/10	ALS	16.234	18.520	20.584	22.708	24.582	26.950
GAPDH	A292/09	Non-ALS Control	15.404	17.570	19.888	22.011	24.042	26.206
HERV-K *gag*	A292/09	Non-ALS Control	23.530	25.807	27.919	29.839	32.056	34.208
HERV-K *pol*	A151/10	ALS	22.061	24.641	26.744	28.764	30.708	32.485
HERV-K *pol*	A292/09	Non-ALS Control	21.971	24.233	26.258	28.386	30.436	32.403
HERV-K *env*	A151/10	ALS	21.749	24.047	26.462	28.380	30.311	33.649
HERV-K *env*	A292/09	Non-ALS Control	22.675	24.998	27.083	29.013	31.363	33.123
HERV-K *RT*	A151/10	ALS	21.923	24.008	26.041	28.183	30.281	32.659
HERV-K *RT*	A292/09	Non-ALS Control	22.223	24.263	26.571	28.611	30.961	31.775
HERV-W *env*	A151/10	ALS	25.270	28.521	31.023	32.696	33.877	34.043
HERV-W *env*	A292/09	Non-ALS Control	26.020	28.887	31.657	33.308	33.897	34.359
TDP-43	A151/10	ALS	21.617	23.587	25.750	27.929	30.098	32.342
TDP-43	A292/09	Non-ALS Control	21.414	23.239	25.276	27.368	29.619	31.858
BCL11b	A151/10	ALS	24.372	26.627	28.616	30.715	32.906	33.390
BCL11b	A292/09	Non-ALS Control	24.769	27.174	29.515	31.229	33.108	33.543

Supplementary Table S5. Amplification efficiency Ct means for HERV-K *gag* primer target using a known ALS sample (A203/11).

Efficiency was determined with 2-fold serial dilutions to generate standard curve. Efficiency was determined in Microsoft Excel using the slope of the generated standard curve and the following equation: =(10^(-1/Slope)-1)*100

Primer Gene Target	Tissue Sample ID	Clinical Status	Undiluted cDNA (Mean Ct)	1:2 Diluted cDNA (Mean Ct)	1:4 Diluted cDNA (Mean Ct)	1:8 Diluted cDNA (Mean Ct)	1:16 Diluted cDNA (Mean Ct)	1:32 Diluted cDNA (Mean Ct)
HERV-K *gag*	A203/11	ALS	25.519	26.615	27.711	28.709	29.694	30.438

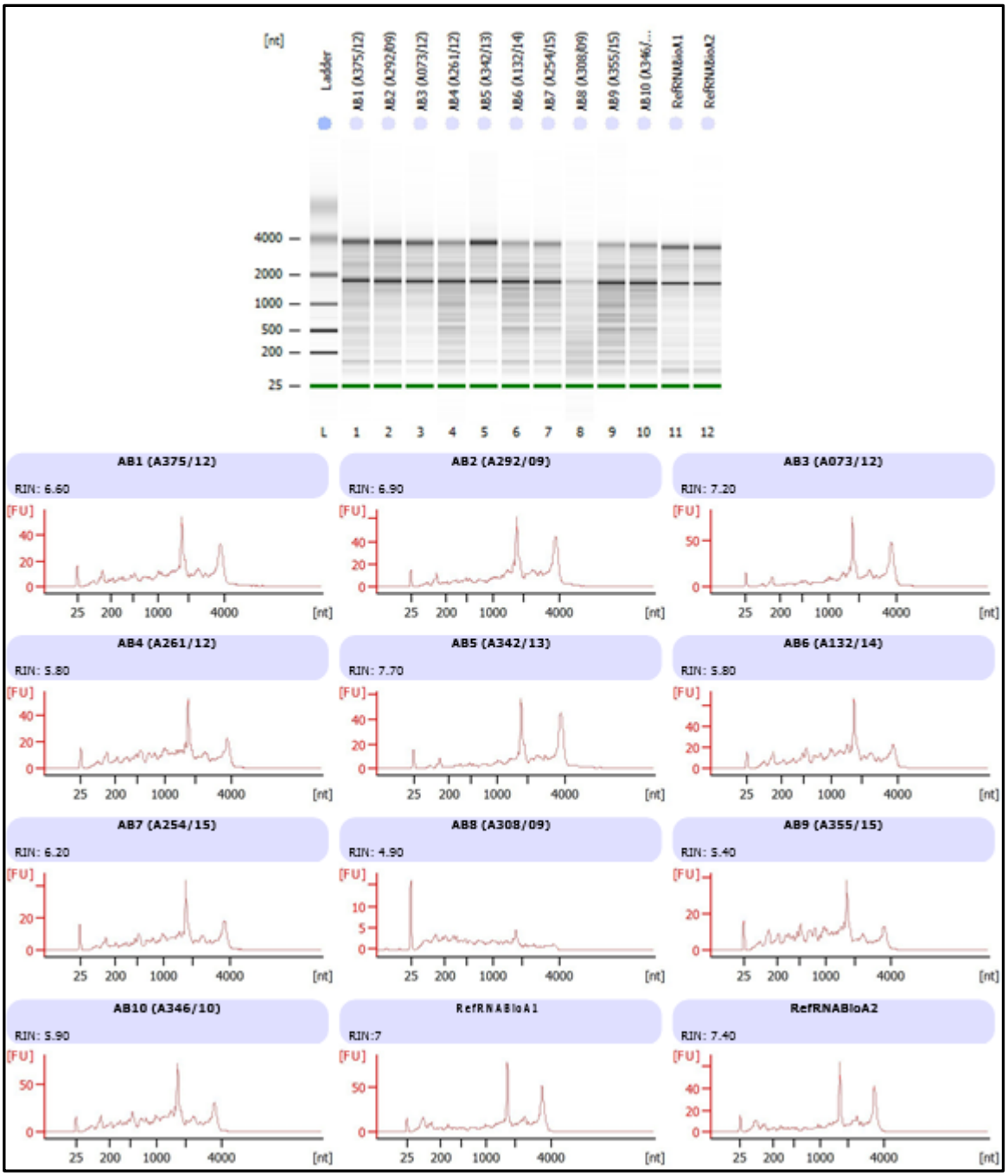


Figure S99. Agilent 2100 Bioanalyser results for samples AB1-AB10 showing computer generated gel image based on sample electrophoretogram, electrophoretograms and RIN.

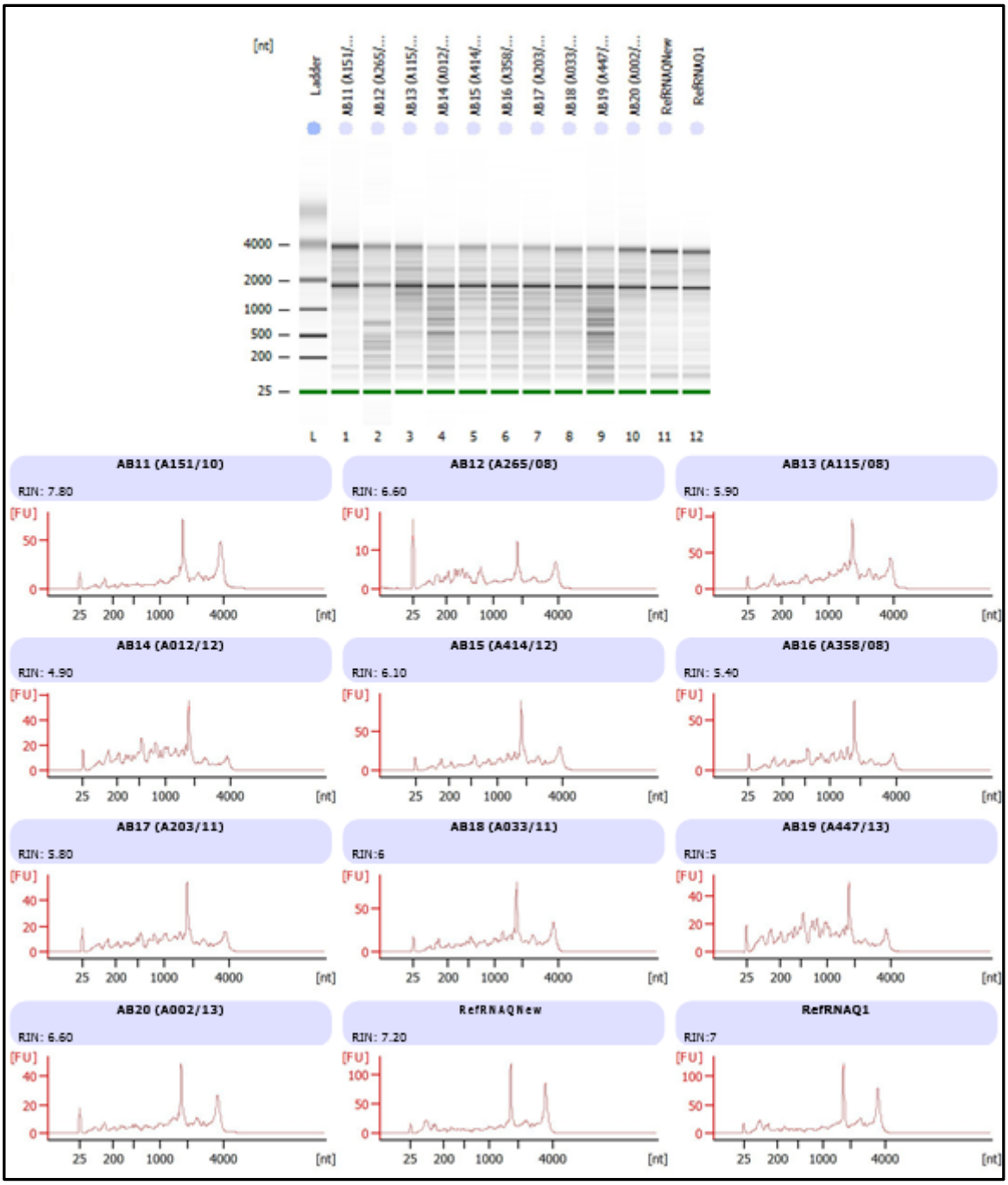


Figure S100. Agilent 2100 Bioanalyser results for samples AB11-AB20 showing computer generated gel image based on sample electrophoretogram, electrophoretograms and RIN.

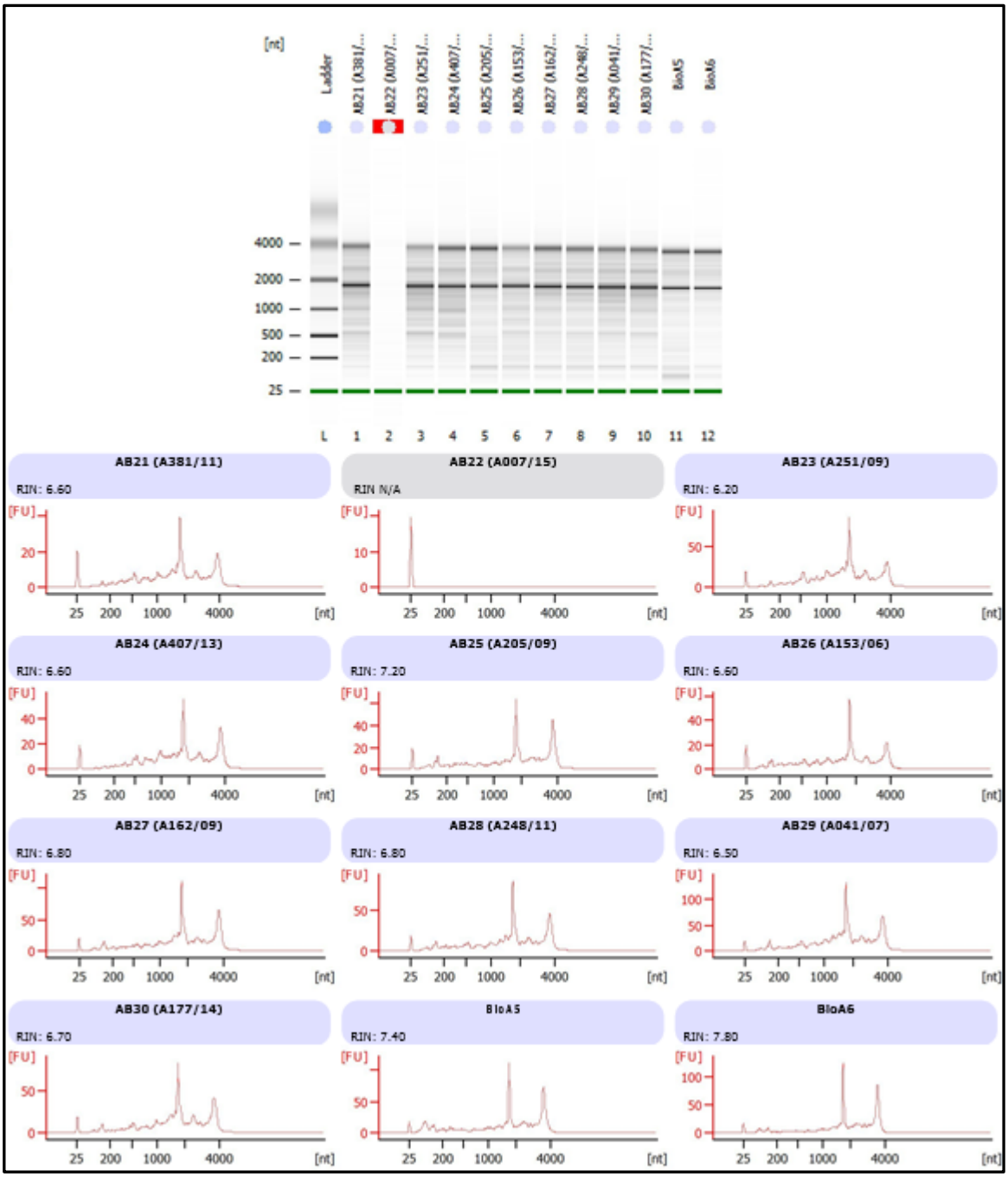


Figure S101. Agilent 2100 Bioanalyser results for samples AB21-AB30 showing computer generated gel image based on sample electrophoretogram, electrophoretograms and RIN.

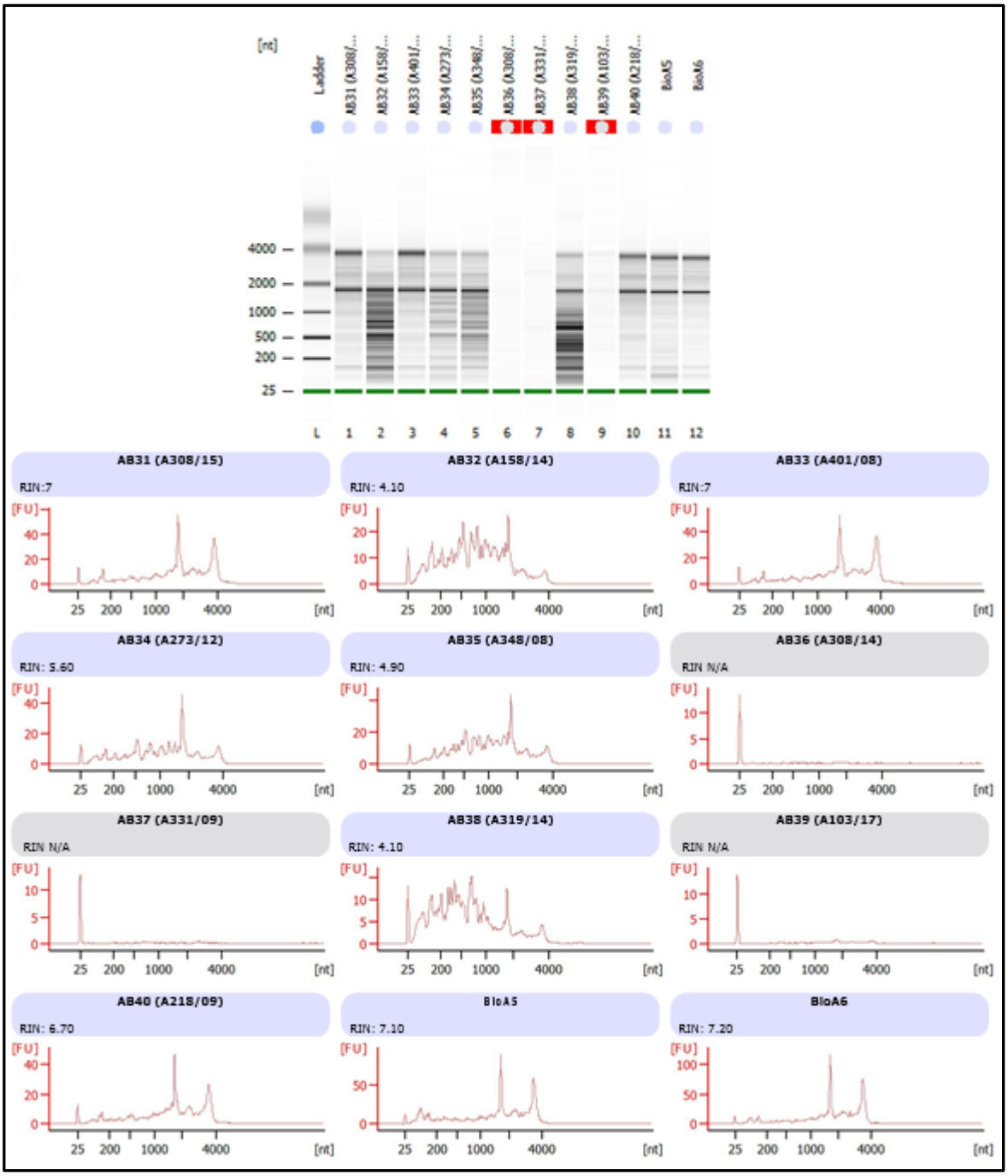


Figure S102. Agilent 2100 Bioanalyser results for samples AB31-AB40 showing computer generated gel image based on sample electrophoretogram, electrophoretograms and RIN.

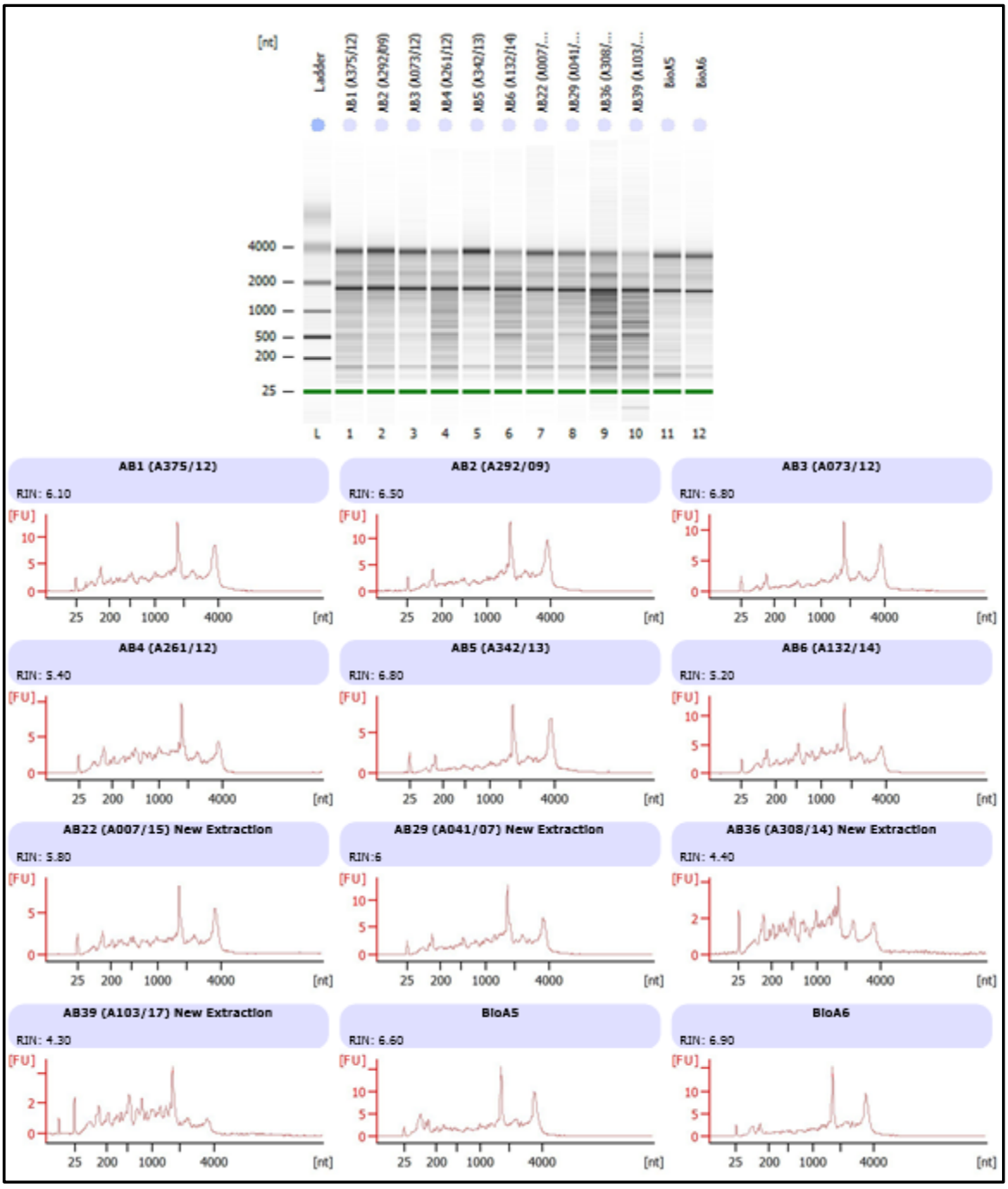


Figure S103. Agilent 2100 Bioanalyser results for repeat samples AB1-AB6
These samples were included as a comparison for newly extracted samples AB22, AB29, AB36 and AB39. These repeat extractions were necessary as the previous experiment yielded samples with too low an RNA concentration to be measured. The figure includes computer generated gel image based on sample electrophoretogram, electrophoretograms and RIN for each sample.

S3. Additional Information for Chapters 6.0 & 7.0

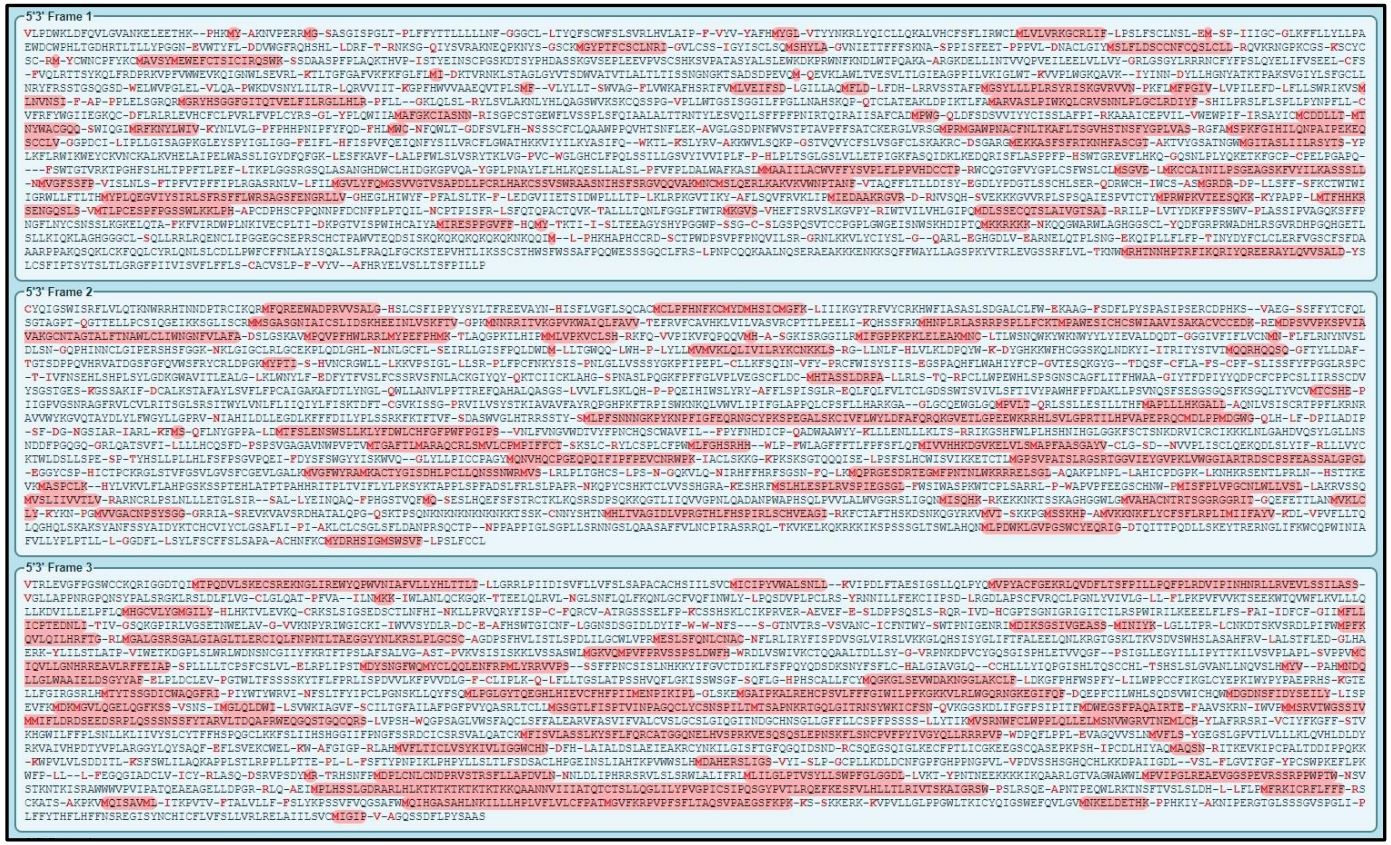


Figure S104. Single Letter FASTA Protein Sequence Translated from ERVmap ID 1412 Nucleotide Sequence Coding for Endogenous Retrovirus HERV4_I

The figure above shows translated single letter FASTA protein sequence for ERVmap region 1412. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

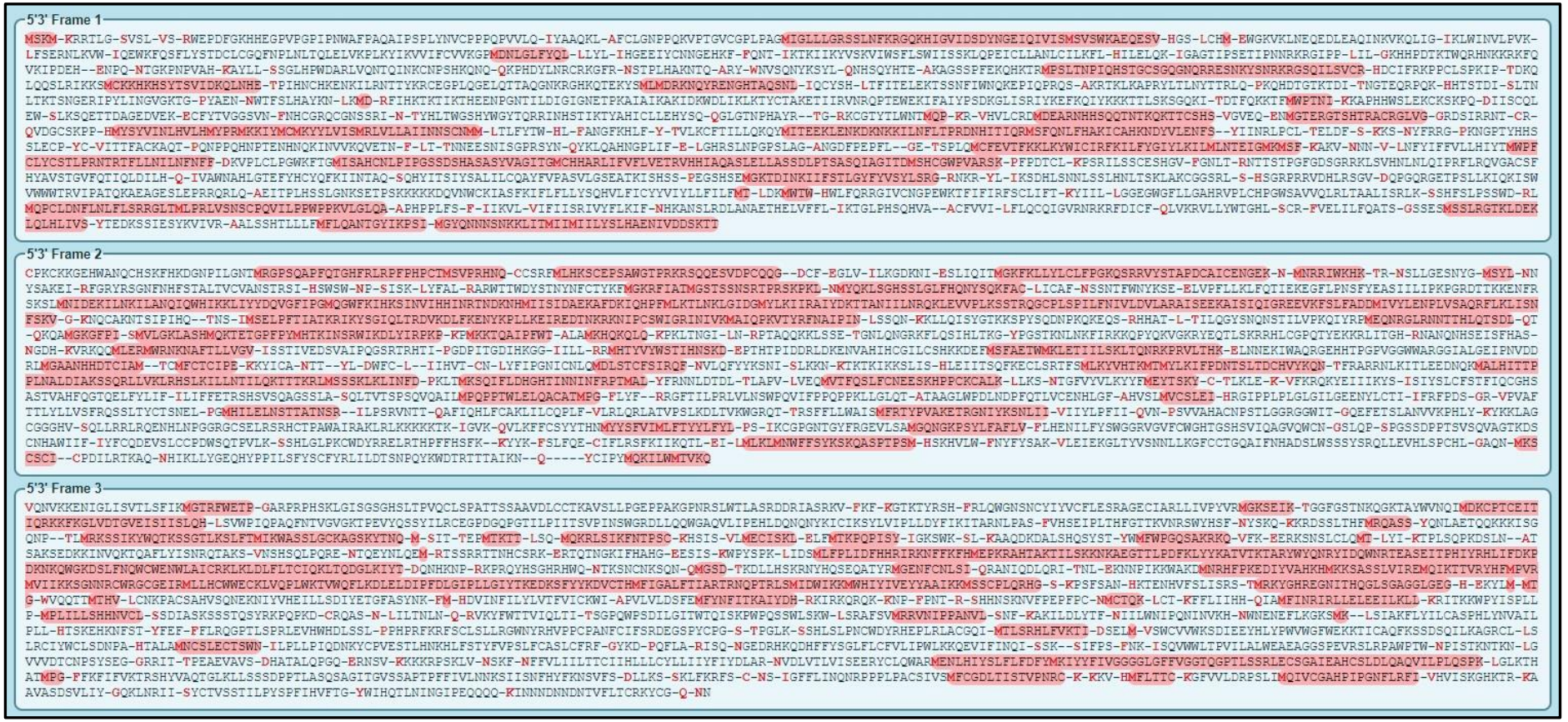


Figure S105. Single Letter FASTA Protein Sequence Translated from ERVmap ID 4160 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-K9

The figure above shows translated single letter FASTA protein sequence for ERVmap region 4160. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

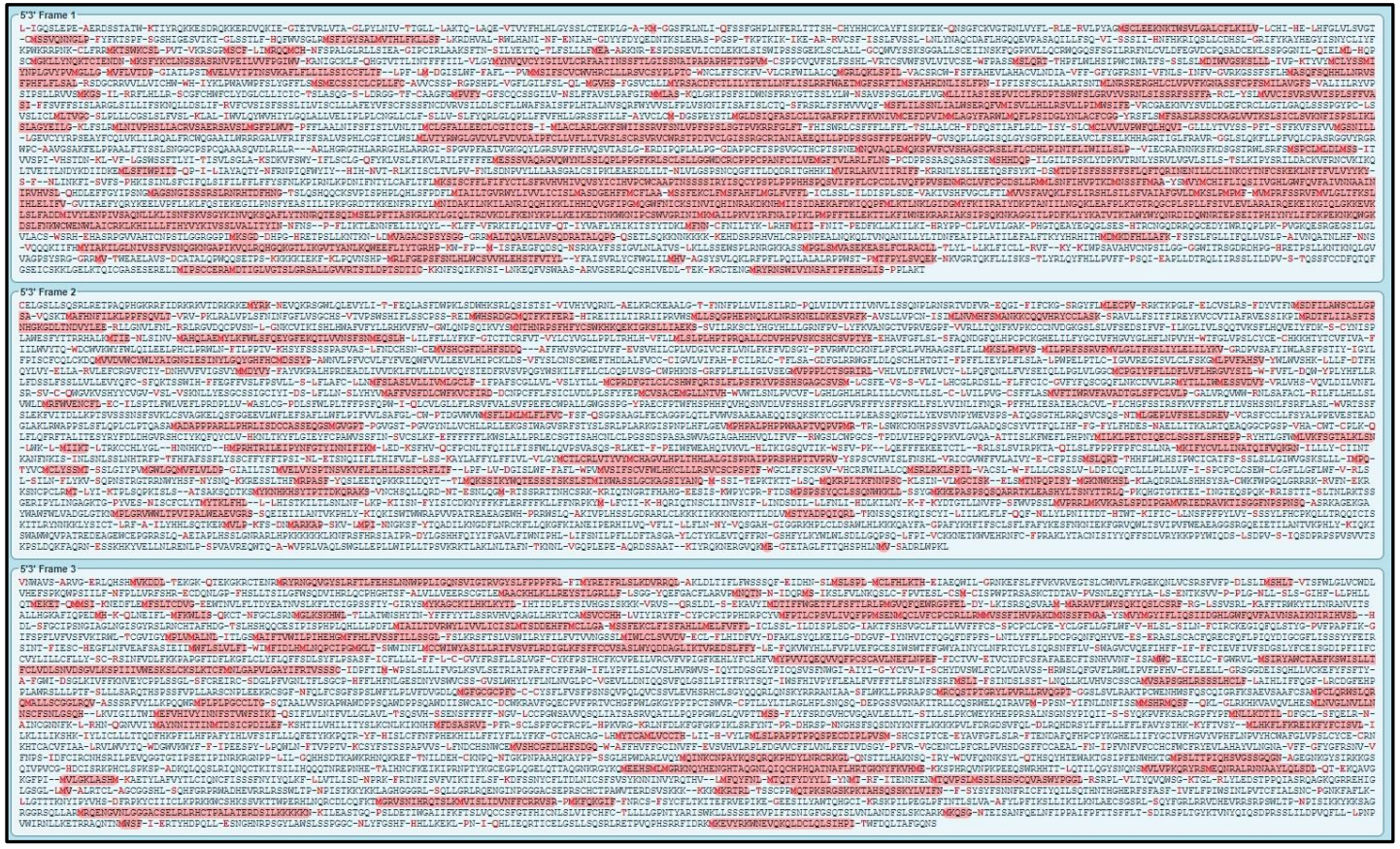


Figure S106. Single Letter FASTA Protein Sequence Translated from ERVmap ID 2458 Nucleotide Sequence Coding for Endogenous Retrovirus MER57

The figure above shows translated single letter FASTA protein sequence for ERVmap region 2458. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single “-“.

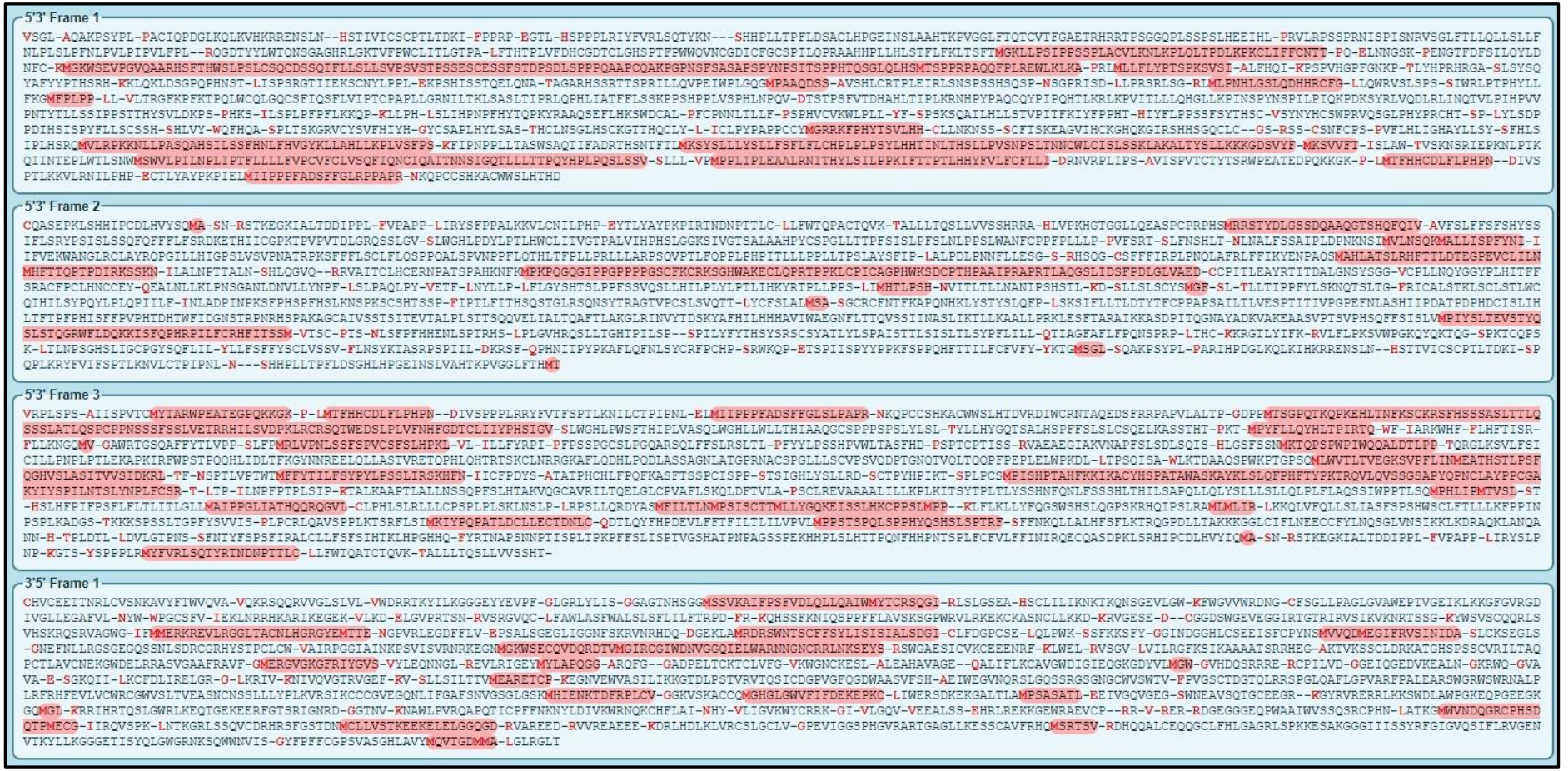


Figure S107. Single Letter FASTA Protein Sequence Translated from ERVmap ID 2658 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-H

The figure above shows translated single letter FASTA protein sequence for ERVmap region 2658. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single “-“.

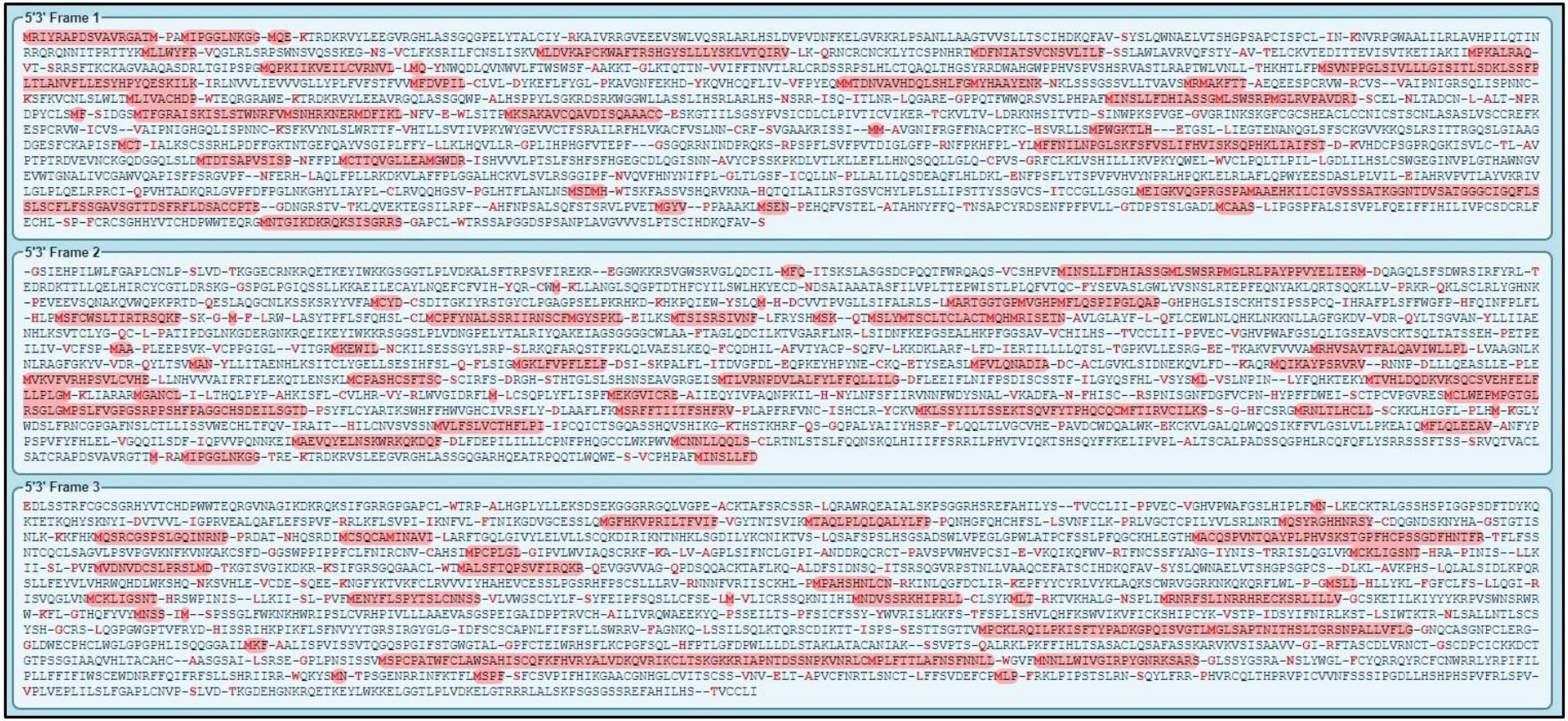


Figure S108. Single Letter FASTA Protein Sequence Translated from ERVmap ID 4757 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-K3

The figure above shows translated single letter FASTA protein sequence for ERVmap region 4757. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S109. Single Letter FASTA Protein Sequence Translated from ERVmap ID 4506 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-L

The figure above shows translated single letter FASTA protein sequence for ERVmap region 4506. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

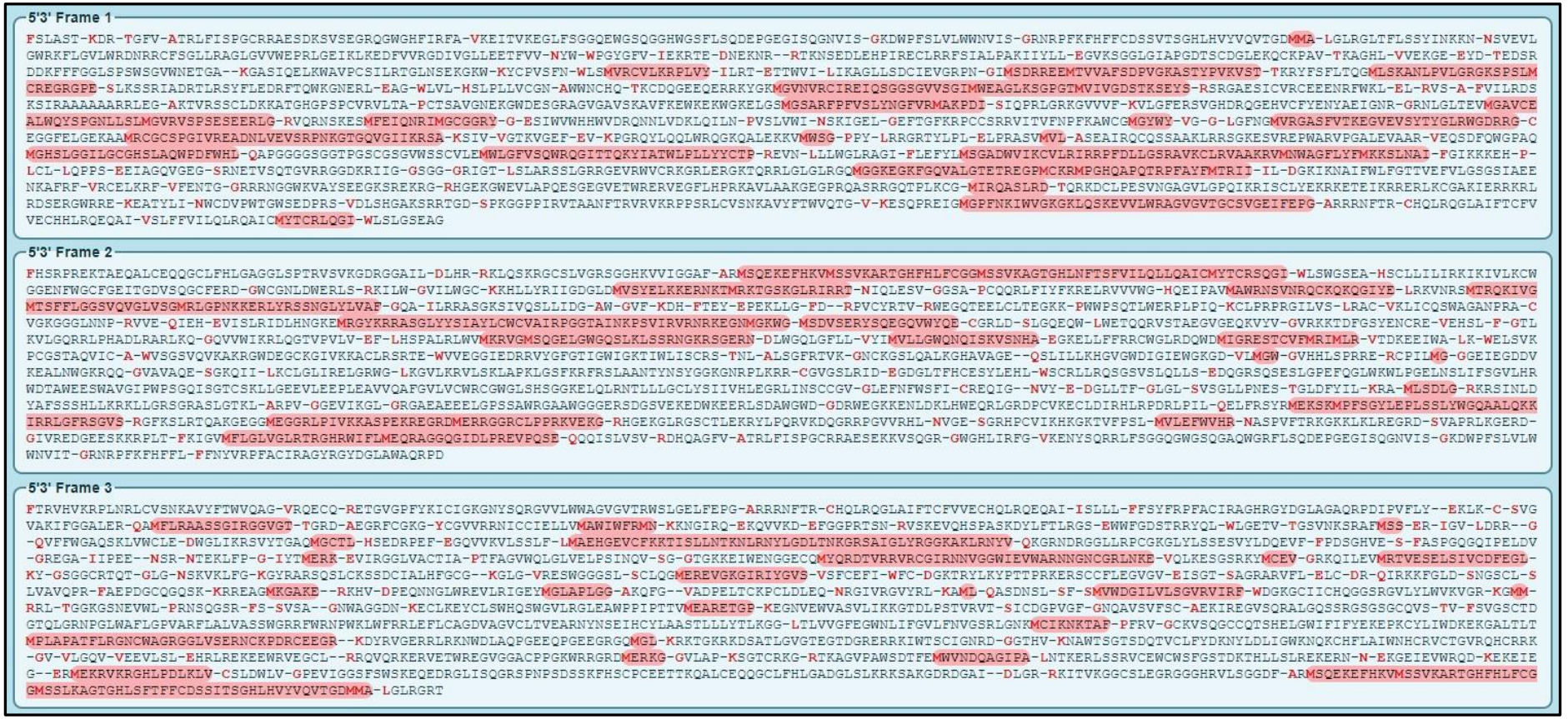


Figure S110. Single Letter FASTA Protein Sequence Translated from ERVmap ID 4864 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-H

The figure above shows translated single letter FASTA protein sequence for ERVmap region 4864. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

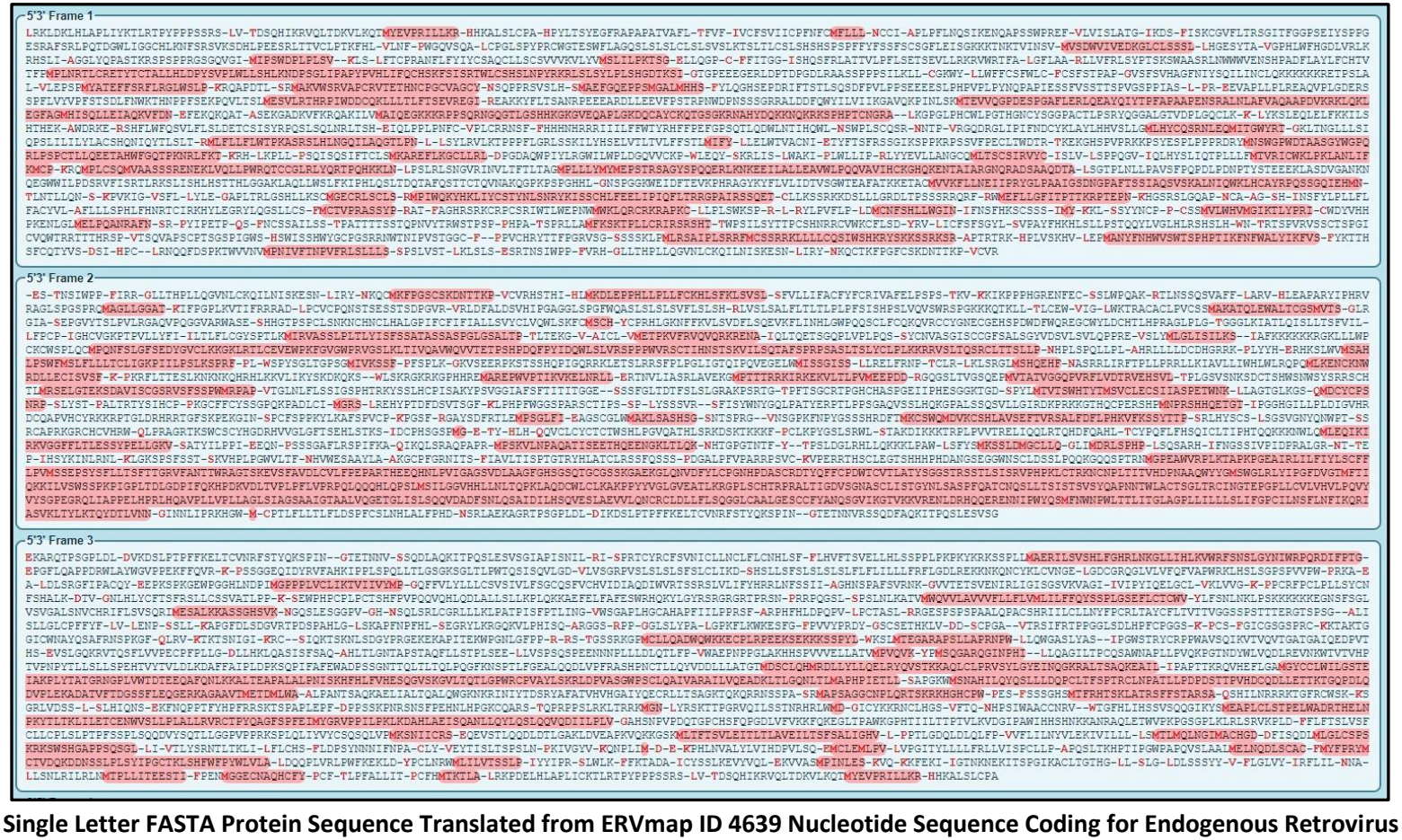


Figure S111. Single Letter FASTA Protein Sequence Translated from ERVmap ID 4639 Nucleotide Sequence Coding for Endogenous Retrovirus HERVS71

The figure above shows translated single letter FASTA protein sequence for ERVmap region 4639. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

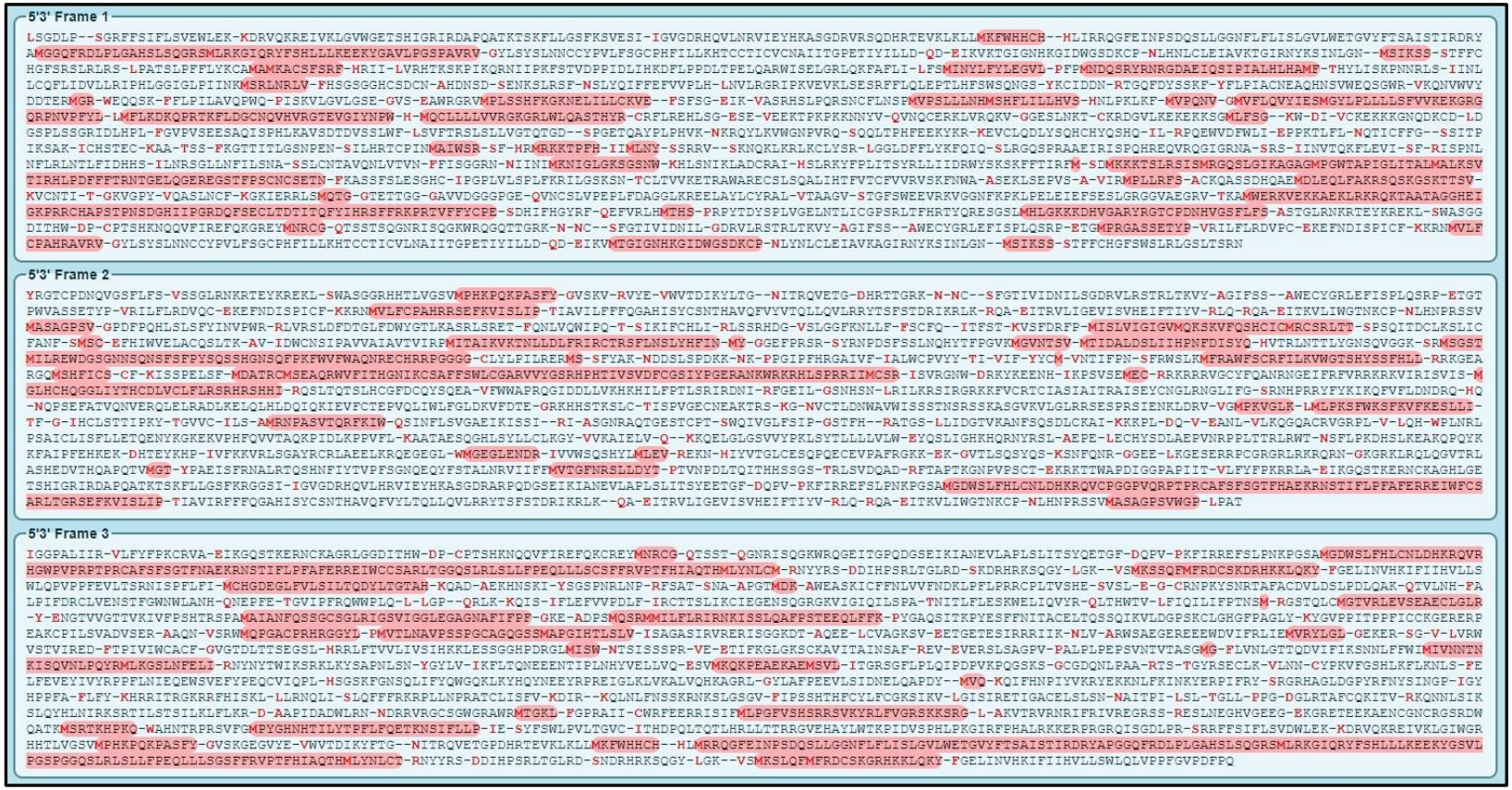


Figure S112. Single Letter FASTA Protein Sequence Translated from ERVmap ID 2699 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-K11

The figure above shows translated single letter FASTA protein sequence for ERVmap region 2699. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S113. Single Letter FASTA Protein Sequence Translated from ERVmap ID 2010 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-H

The figure above shows translated single letter FASTA protein sequence for ERVmap region 2010. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S114. Single Letter FASTA Protein Sequence Translated from ERVmap ID 935 Nucleotide Sequence Coding for Endogenous Retrovirus HERV9

The figure above shows translated single letter FASTA protein sequence for ERVmap region 935. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

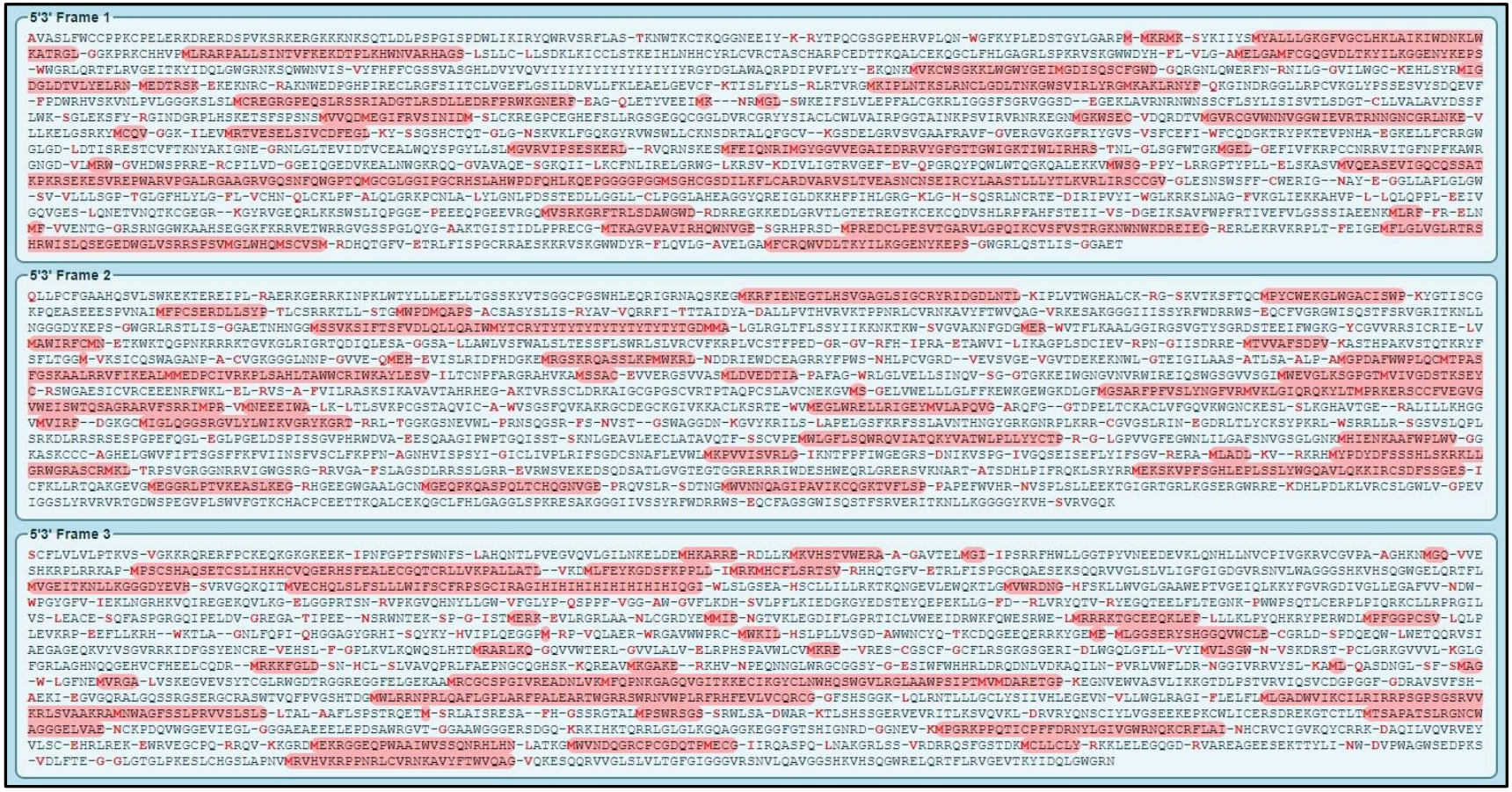


Figure S115. Single Letter FASTA Protein Sequence Translated from ERVmap ID 2294 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-H

The figure above shows translated single letter FASTA protein sequence for ERVmap region 2294. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

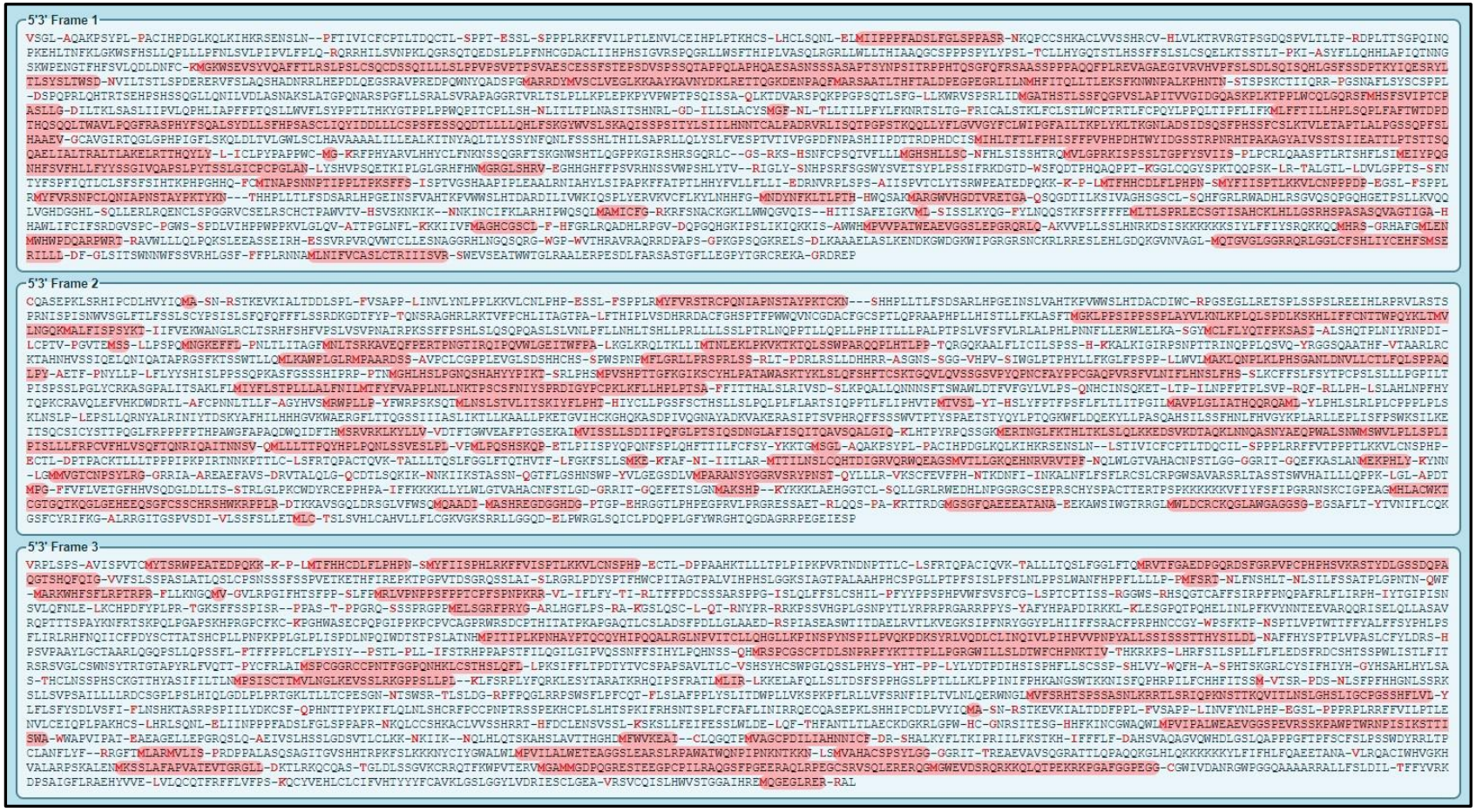


Figure S116. Single Letter FASTA Protein Sequence Translated from ERVmap ID 4843 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-H

The figure above shows translated single letter FASTA protein sequence for ERVmap region 4843. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S117. Single Letter FASTA Protein Sequence Translated from ERVmap ID 2548 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-L

The figure above shows translated single letter FASTA protein sequence for ERVmap region 2548. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S118. Single Letter FASTA Protein Sequence Translated from ERVmap ID W-92 Nucleotide Sequence Coding for Endogenous Retrovirus HERV17

The figure above shows translated single letter FASTA protein sequence for ERVmap region W-92. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

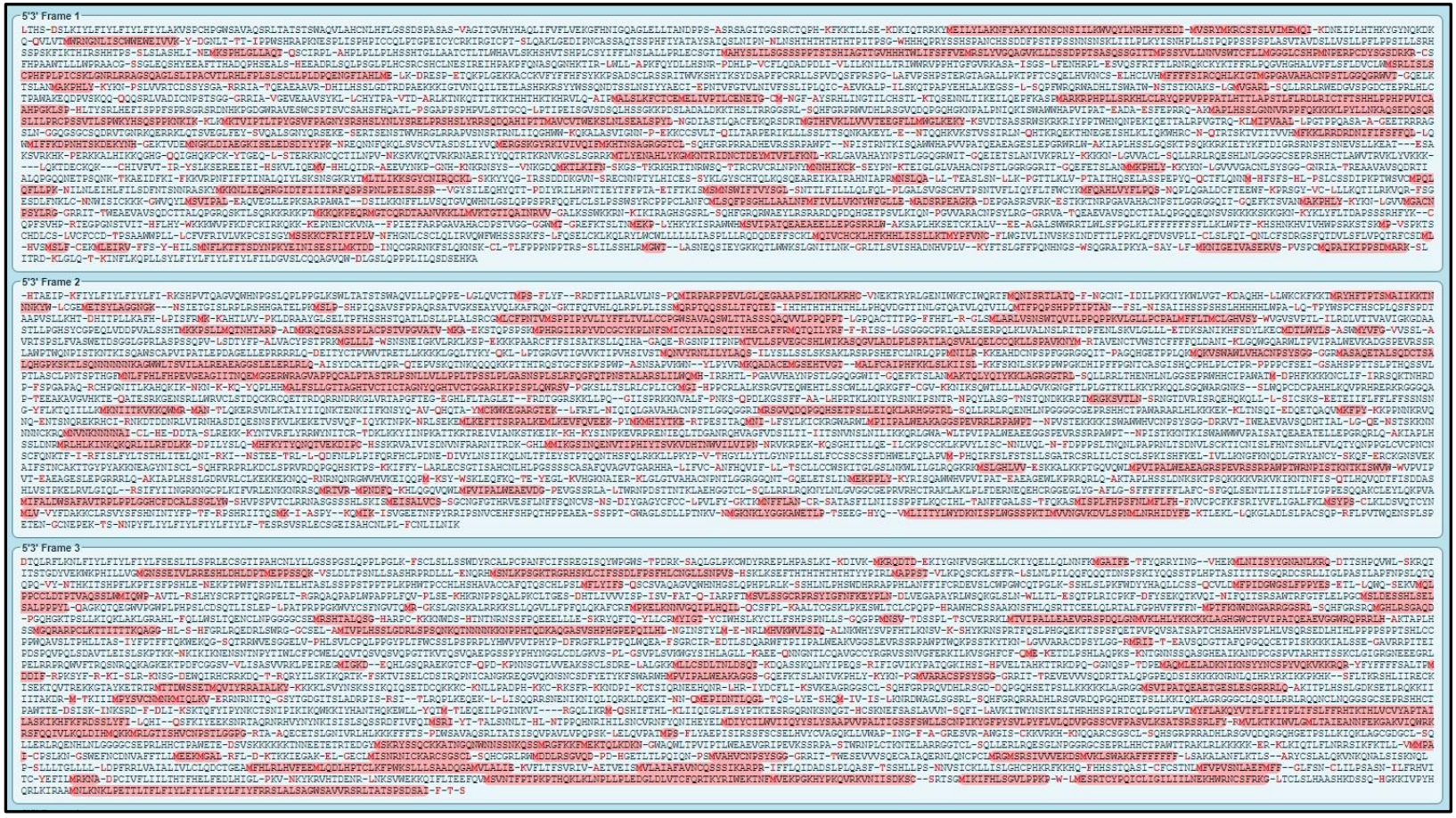


Figure S119. Single Letter FASTA Protein Sequence Translated from ERVmap ID 2049 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-L

The figure above shows translated single letter FASTA protein sequence for ERVmap region 2049. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S120. Single Letter FASTA Protein Sequence Translated from ERVmap ID 1143 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-H

The figure above shows translated single letter FASTA protein sequence for ERVmap region 1143. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single “-“.

Figure S121. Single Letter FASTA Protein Sequence Translated from ERVmap ID 4861 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-L

The figure above shows translated single letter FASTA protein sequence for ERVmap region 4861. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

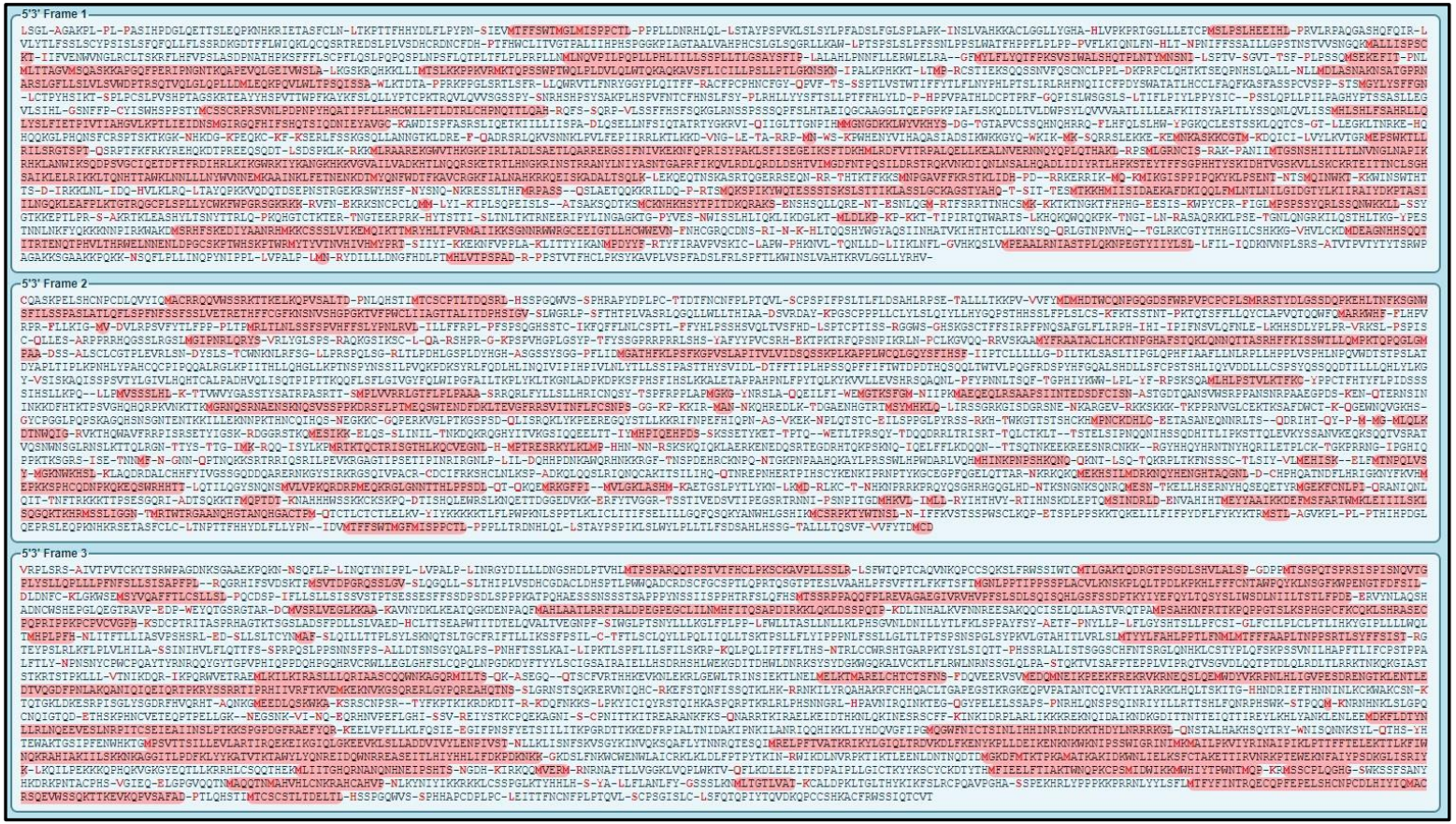


Figure S122. Single Letter FASTA Protein Sequence Translated from ERVmap ID 6195 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-H

The figure above shows translated single letter FASTA protein sequence for ERVmap region 6195. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S123. Single Letter FASTA Protein Sequence Translated from ERVmap ID 1115 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-K9

The figure above shows translated single letter FASTA protein sequence for ERVmap region 1115. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S124. Single Letter FASTA Protein Sequence Translated from ERVmap ID 1739 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-H

The figure above shows translated single letter FASTA protein sequence for ERVmap region 1739. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S125. Single Letter FASTA Protein Sequence Translated from ERVmap ID 4152 Nucleotide Sequence Coding for Endogenous Retrovirus PRIMA4

The figure above shows translated single letter FASTA protein sequence for ERVmap region 4152. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S126. Single Letter FASTA Protein Sequence Translated from ERVmap ID 2223 Nucleotide Sequence Coding for Endogenous Retrovirus HERV4_I

The figure above shows translated single letter FASTA protein sequence for ERVmap region 2223. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S127. Single Letter FASTA Protein Sequence Translated from ERVmap ID 1643 Nucleotide Sequence Coding for Endogenous Retrovirus HERVP71A

The figure above shows translated single letter FASTA protein sequence for ERVmap region 1643. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

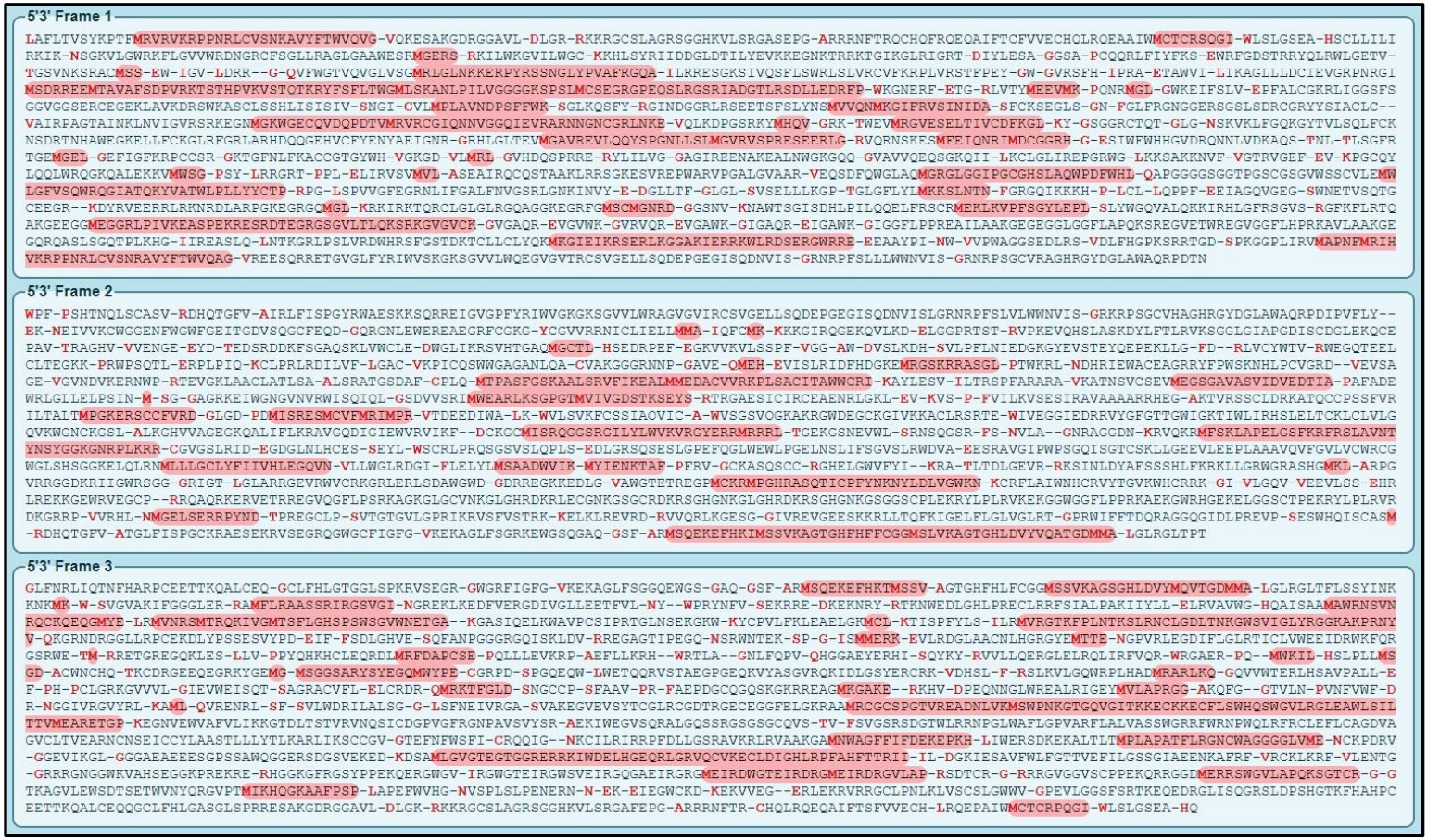


Figure S128. Single Letter FASTA Protein Sequence Translated from ERVmap ID 2724 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-H

The figure above shows translated single letter FASTA protein sequence for ERVmap region 2724. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S129. Single Letter FASTA Protein Sequence Translated from ERVmap ID 3388 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-K13
The figure above shows translated single letter FASTA protein sequence for ERVmap region 3388. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

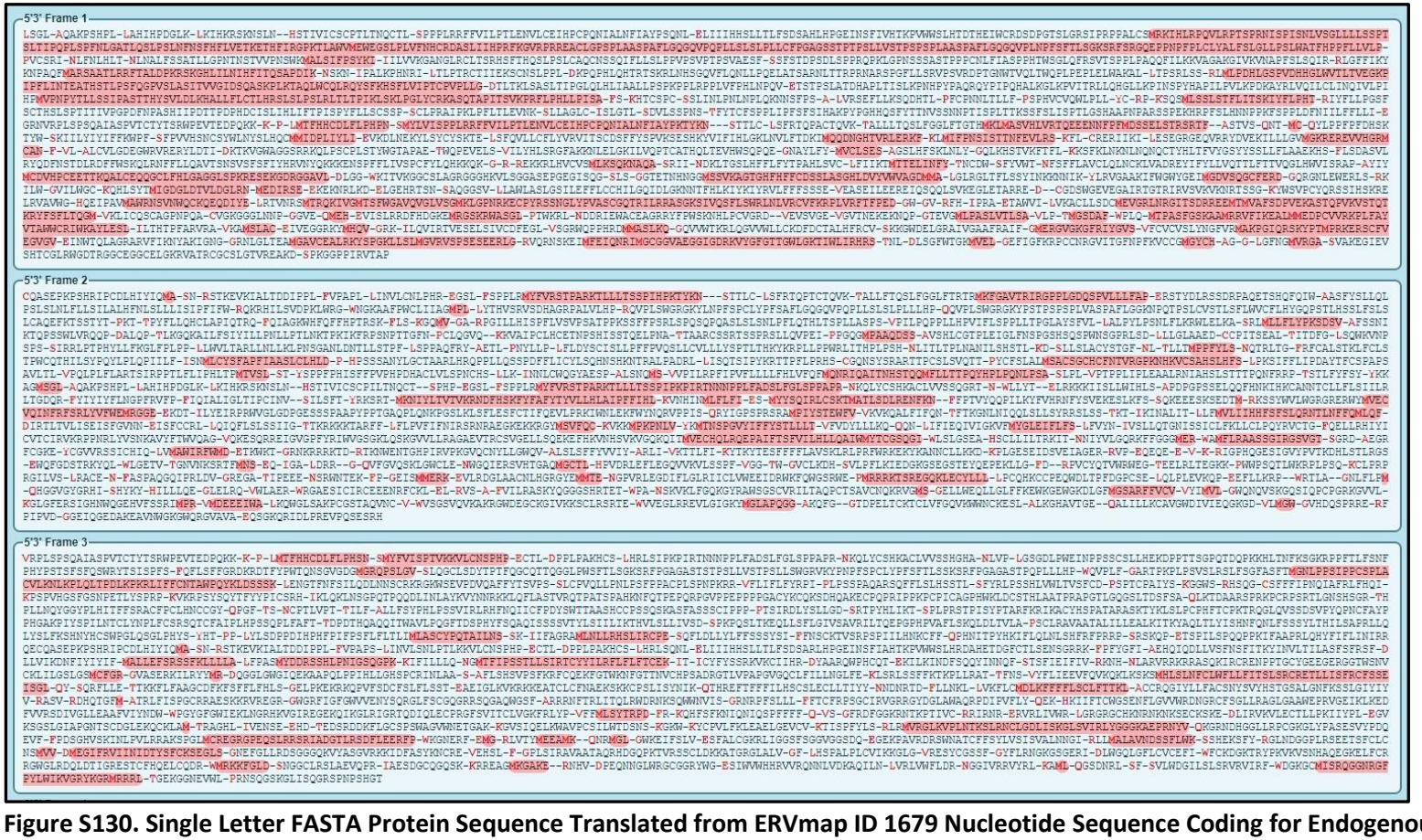


Figure S130. Single Letter FASTA Protein Sequence Translated from ERVmap ID 1679 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-H

The figure above shows translated single letter FASTA protein sequence for ERVmap region 1679. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S131. Single Letter FASTA Protein Sequence Translated from ERVmap ID 2621 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-H

The figure above shows translated single letter FASTA protein sequence for ERVmap region 2621. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

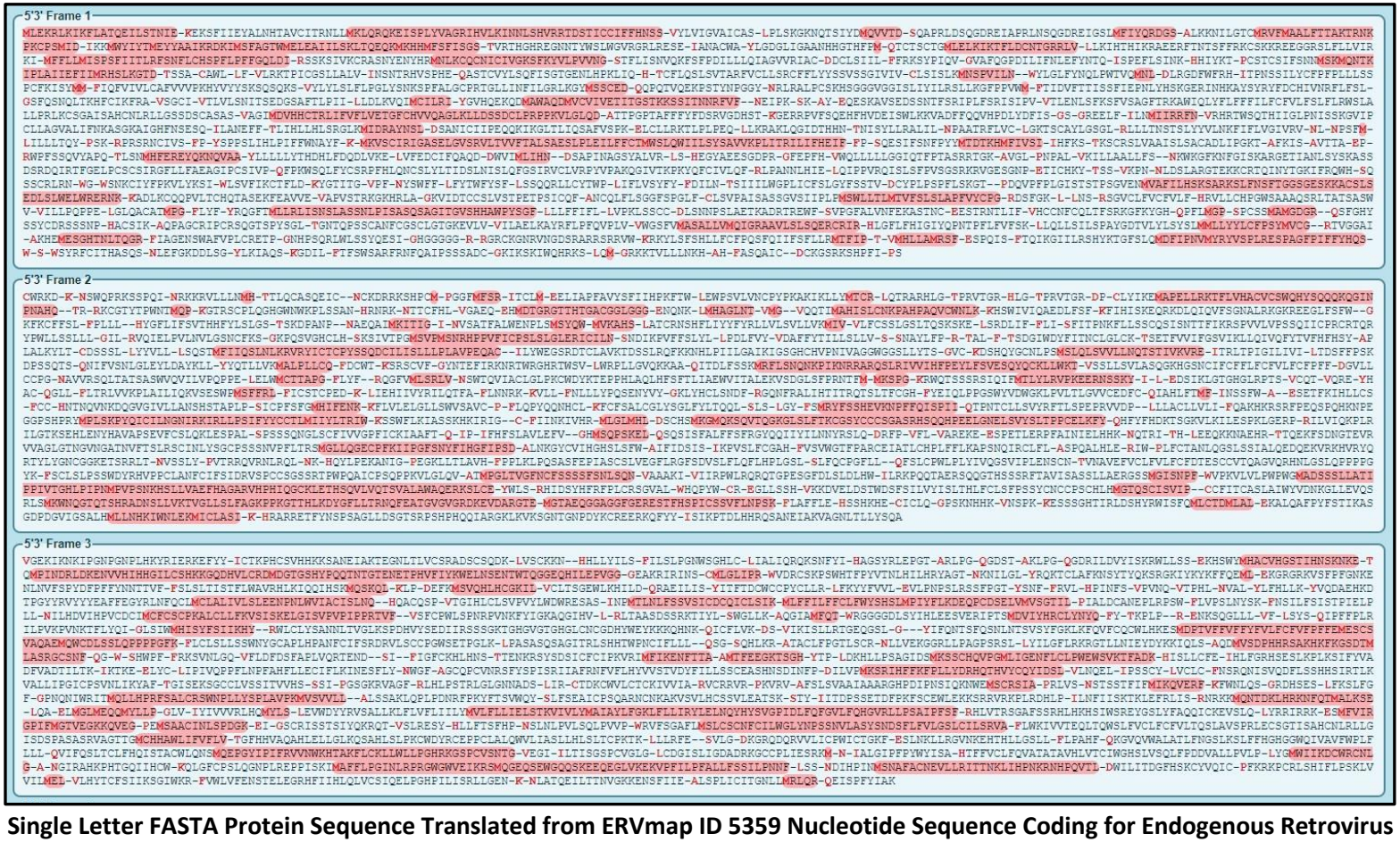


Figure S132. Single Letter FASTA Protein Sequence Translated from ERVmap ID 5359 Nucleotide Sequence Coding for Endogenous Retrovirus MER89

The figure above shows translated single letter FASTA protein sequence for ERVmap region 5359. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S133. Single Letter FASTA Protein Sequence Translated from ERVmap ID 1379 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-K3

The figure above shows translated single letter FASTA protein sequence for ERVmap region 1379. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

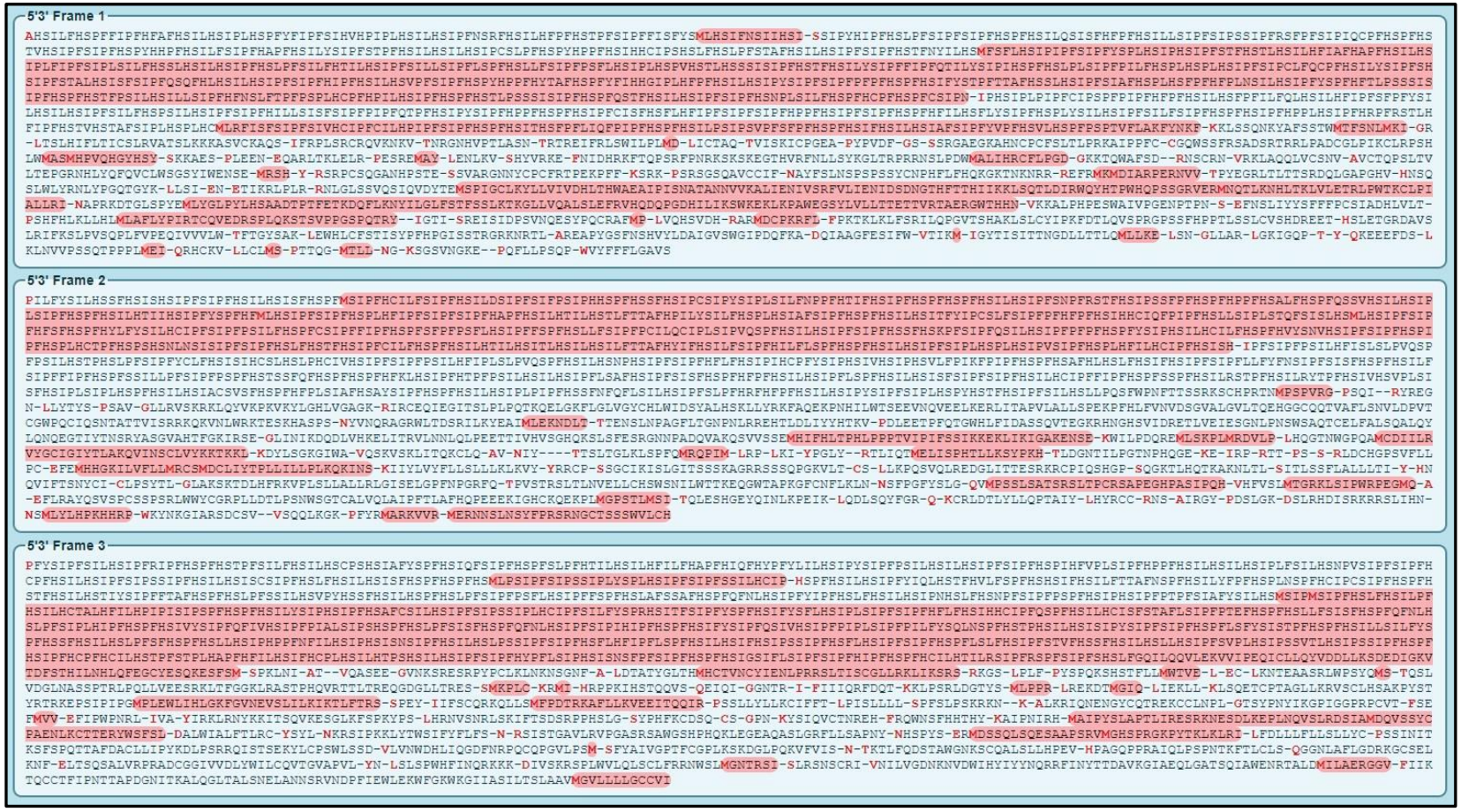


Figure S134. Single Letter FASTA Protein Sequence Translated from ERVmap ID 3200 Nucleotide Sequence Coding for Endogenous Retrovirus HERVIP1-F

The figure above shows translated single letter FASTA protein sequence for ERVmap region 3200. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

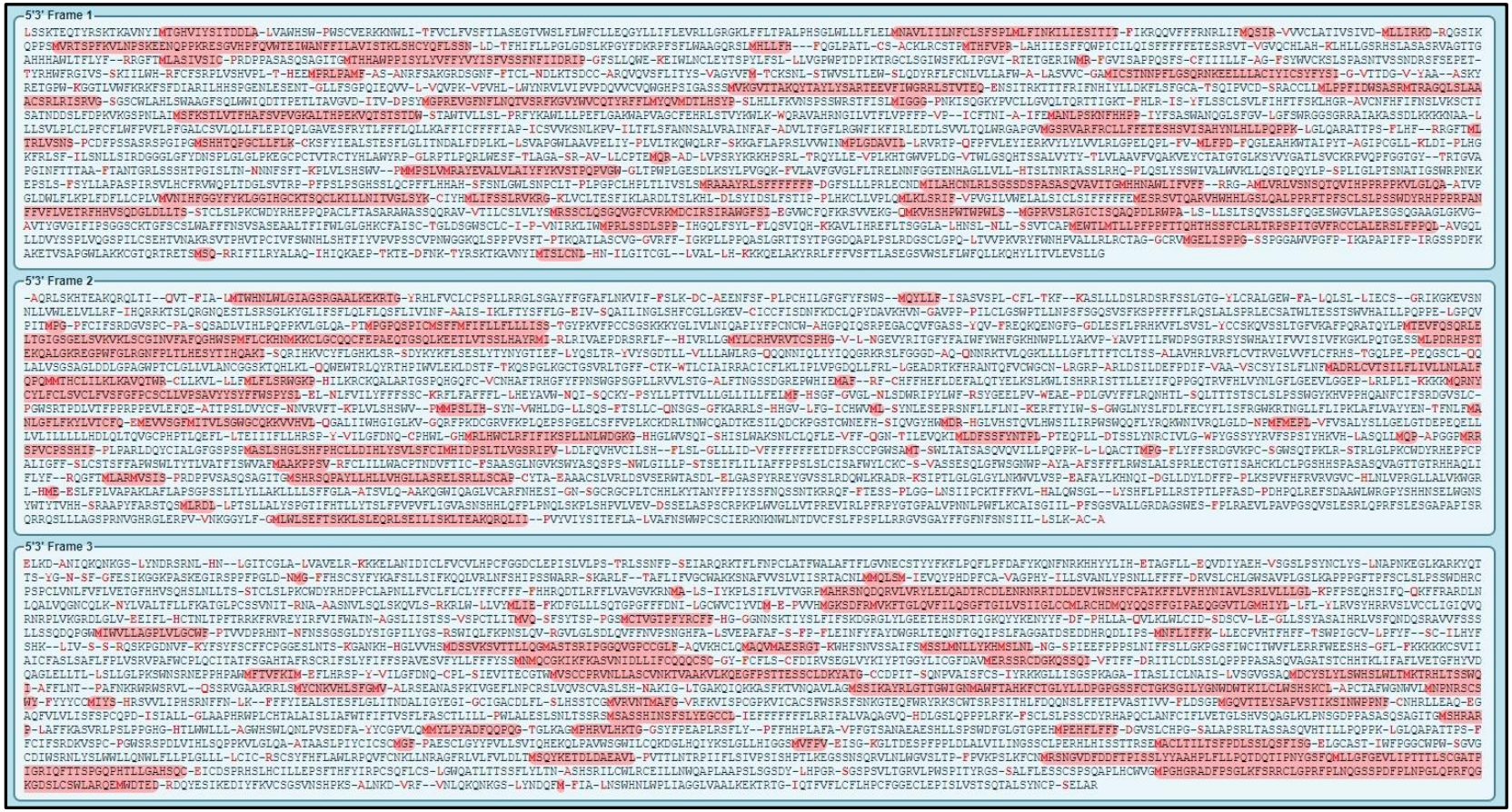


Figure S135. Single Letter FASTA Protein Sequence Translated from ERVmap ID 4656 Nucleotide Sequence Coding for Endogenous Retrovirus HERV3

The figure above shows translated single letter FASTA protein sequence for ERVmap region 4656. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S136. Single Letter FASTA Protein Sequence Translated from ERVmap ID 2334 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-H

The figure above shows translated single letter FASTA protein sequence for ERVmap region 2334. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S137. Single Letter FASTA Protein Sequence Translated from ERVmap ID 2360 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-17

The figure above shows translated single letter FASTA protein sequence for ERVmap region 2360. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

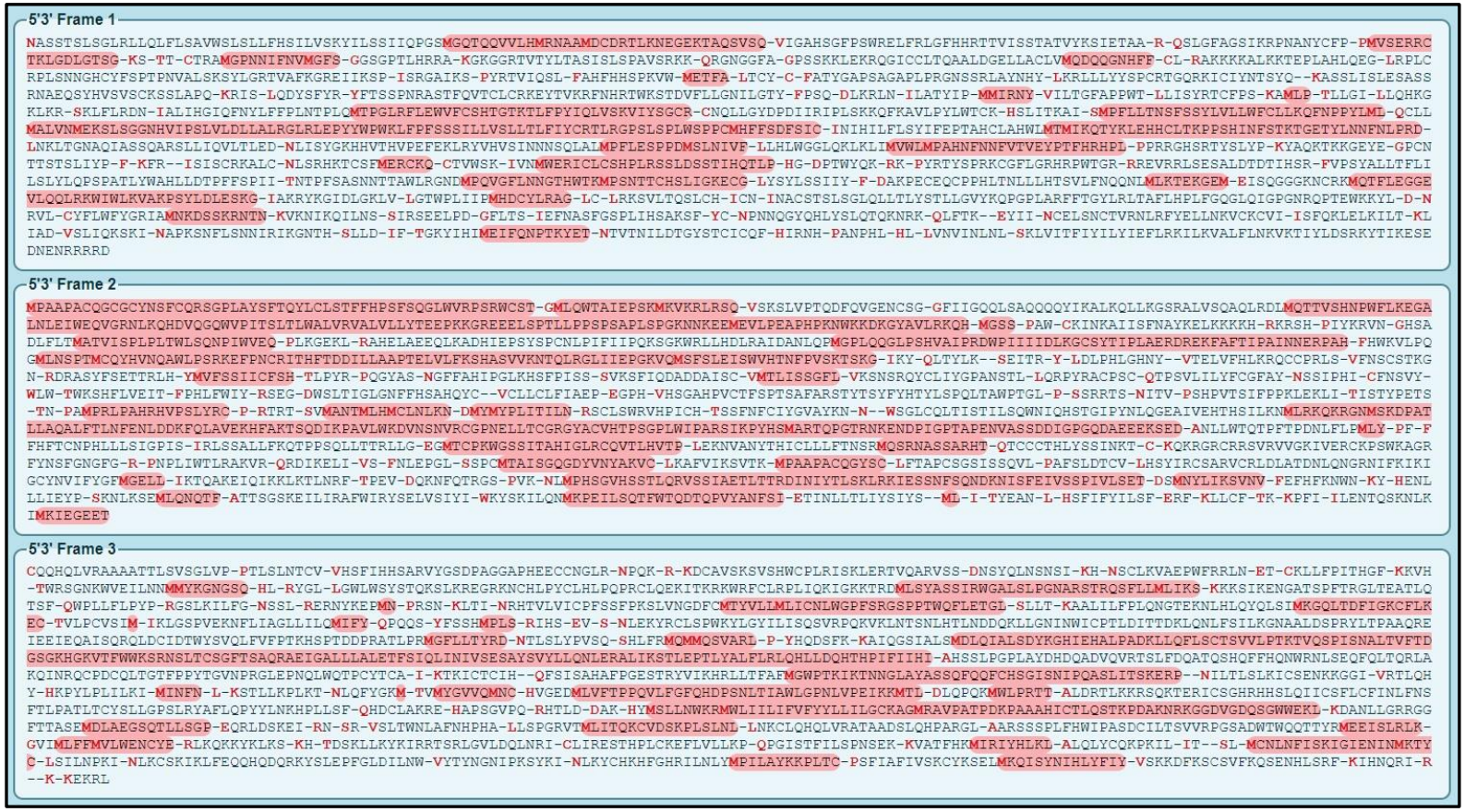


Figure S138. Single Letter FASTA Protein Sequence Translated from ERVmap ID 1549 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-K22

The figure above shows translated single letter FASTA protein sequence for ERVmap region 1549. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S139. Single Letter FASTA Protein Sequence Translated from ERVmap ID 1797 Nucleotide Sequence Coding for Endogenous Retrovirus HERV9

The figure above shows translated single letter FASTA protein sequence for ERVmap region 1797. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

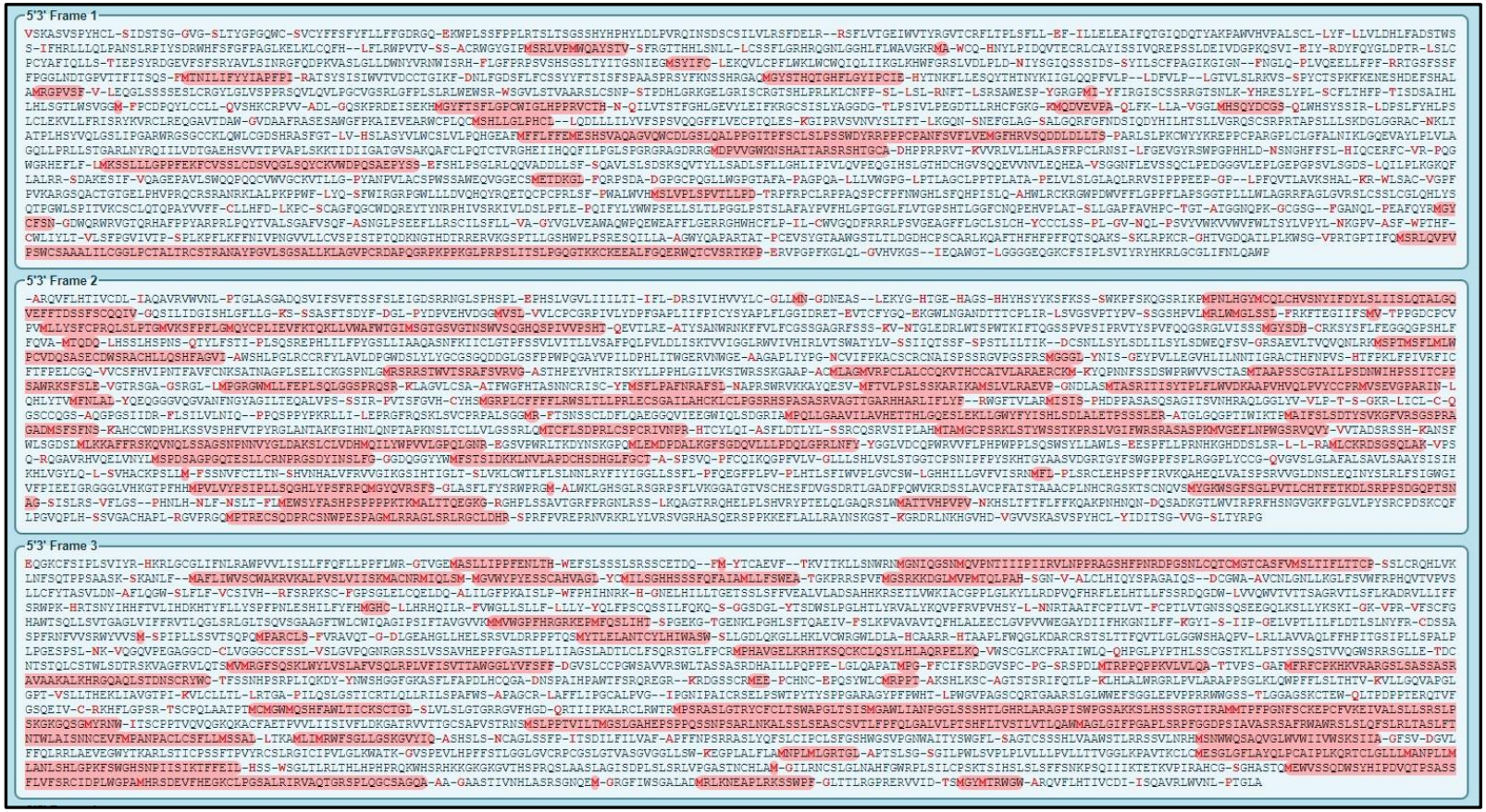


Figure S140. Single Letter FASTA Protein Sequence Translated from ERVmap ID 5361 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-E

The figure above shows translated single letter FASTA protein sequence for ERVmap region 5361. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S141. Single Letter FASTA Protein Sequence Translated from ERVmap ID 4340 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-K9

The figure above shows translated single letter FASTA protein sequence for ERVmap region 4340. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

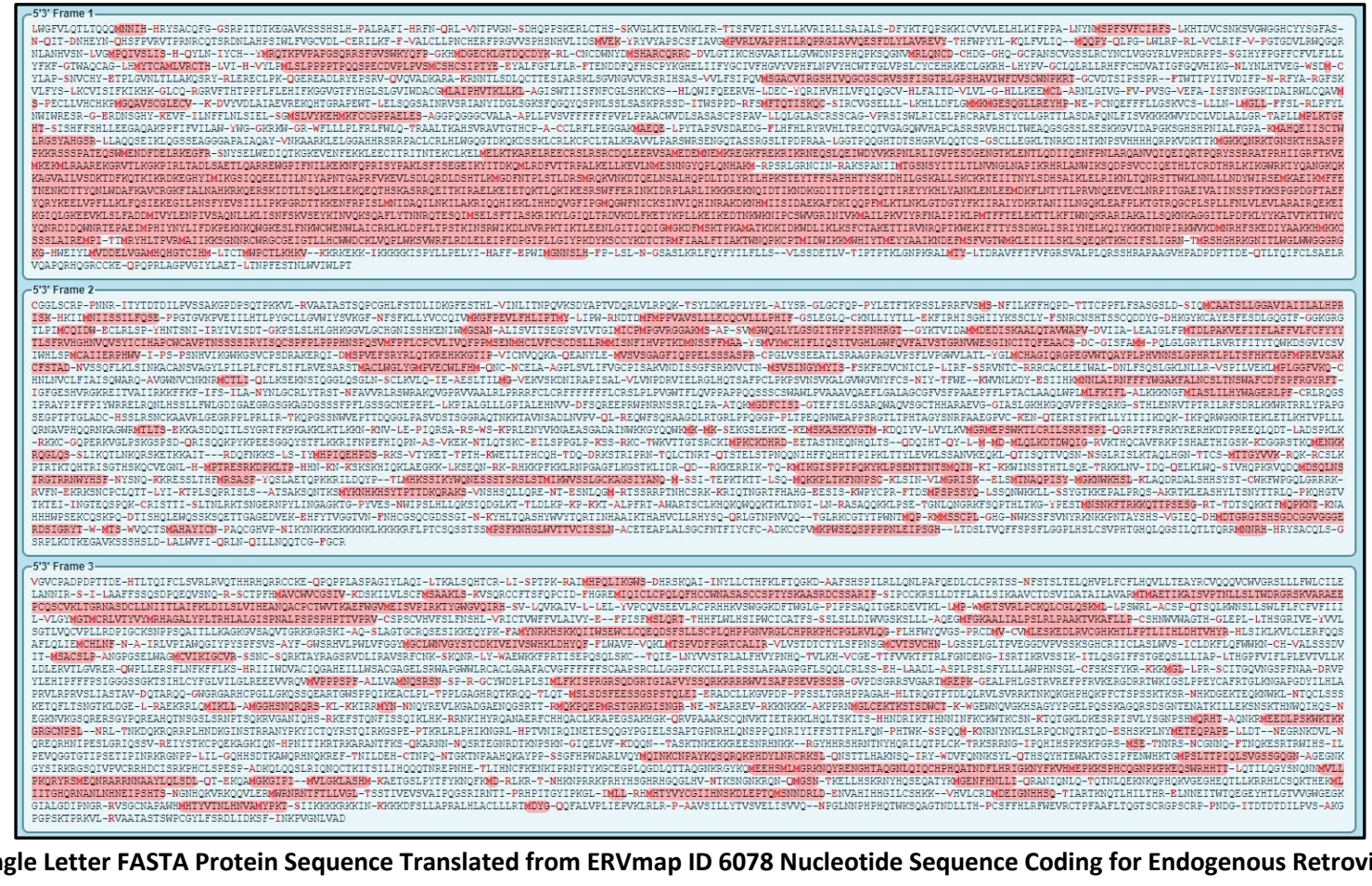


Figure S142. Single Letter FASTA Protein Sequence Translated from ERVmap ID 6078 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-K22

The figure above shows translated single letter FASTA protein sequence for ERVmap region 6078. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S143. Single Letter FASTA Protein Sequence Translated from ERVmap ID 2307 Nucleotide Sequence Coding for Endogenous Retrovirus HERV30

The figure above shows translated single letter FASTA protein sequence for ERVmap region 2307. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S144. Single Letter FASTA Protein Sequence Translated from ERVmap ID 2306 Nucleotide Sequence Coding for Endogenous Retrovirus HERVIP10B3
The figure above shows translated single letter FASTA protein sequence for ERVmap region 2306. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

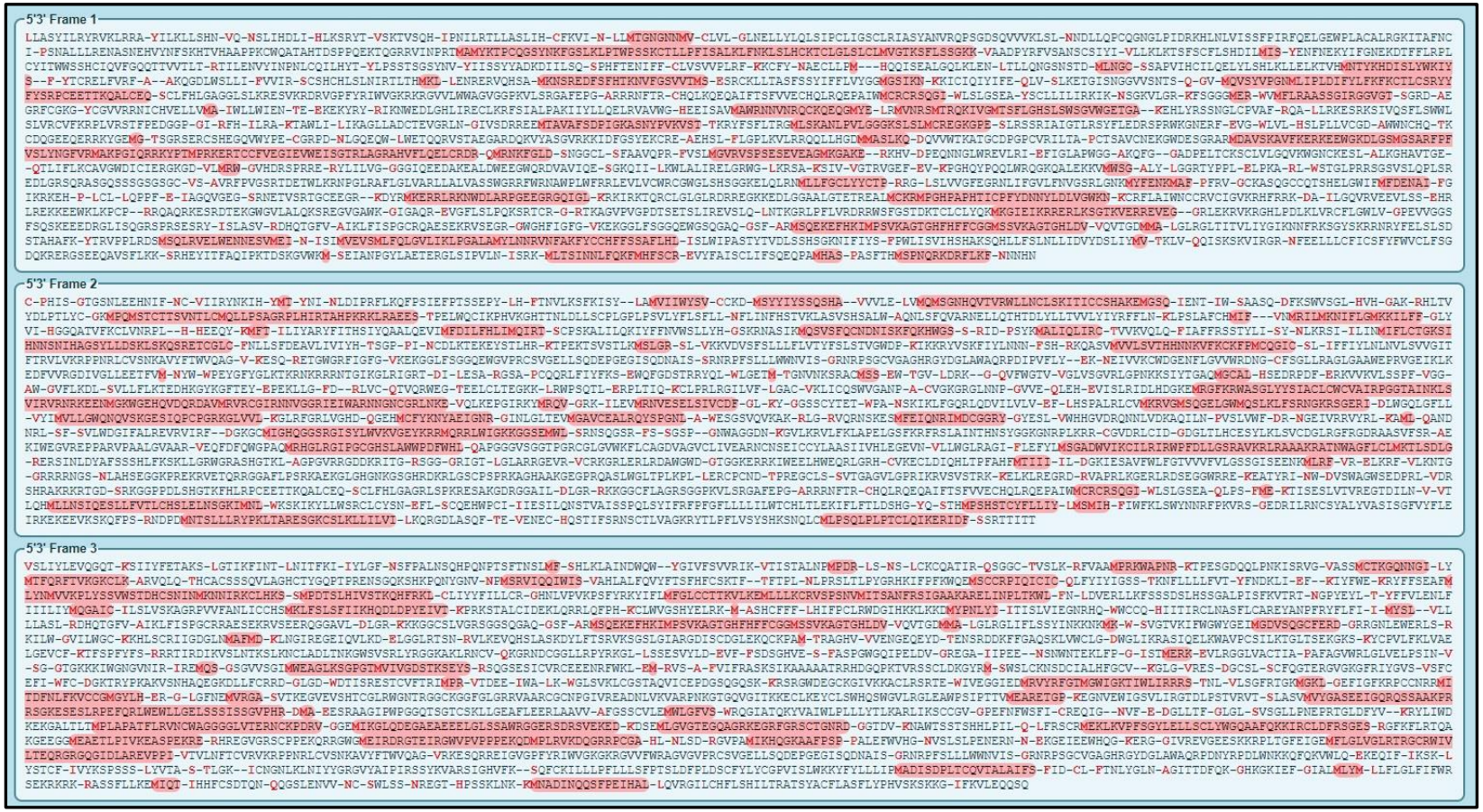


Figure S145. Single Letter FASTA Protein Sequence Translated from ERVmap ID 570 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-H

The figure above shows translated single letter FASTA protein sequence for ERVmap region 570. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S146. Single Letter FASTA Protein Sequence Translated from ERVmap ID 5947 Nucleotide Sequence Coding for Endogenous Retrovirus MER101

The figure above shows translated single letter FASTA protein sequence for ERVmap region 5947. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S147. Single Letter FASTA Protein Sequence Translated from ERVmap ID 909 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-H

The figure above shows translated single letter FASTA protein sequence for ERVmap region 909. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S148. Single Letter FASTA Protein Sequence Translated from ERVmap ID 2305 Nucleotide Sequence Coding for Endogenous Retrovirus HUERS-P3

The figure above shows translated single letter FASTA protein sequence for ERVmap region 2305. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S149. Single Letter FASTA Protein Sequence Translated from ERVmap ID 765 Nucleotide Sequence Coding for Endogenous Retrovirus HERVL18

The figure above shows translated single letter FASTA protein sequence for ERVmap region 765. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S150. Single Letter FASTA Protein Sequence Translated from ERVmap ID 3704 Nucleotide Sequence Coding for Endogenous Retrovirus HERV15

The figure above shows translated single letter FASTA protein sequence for ERVmap region 3704. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

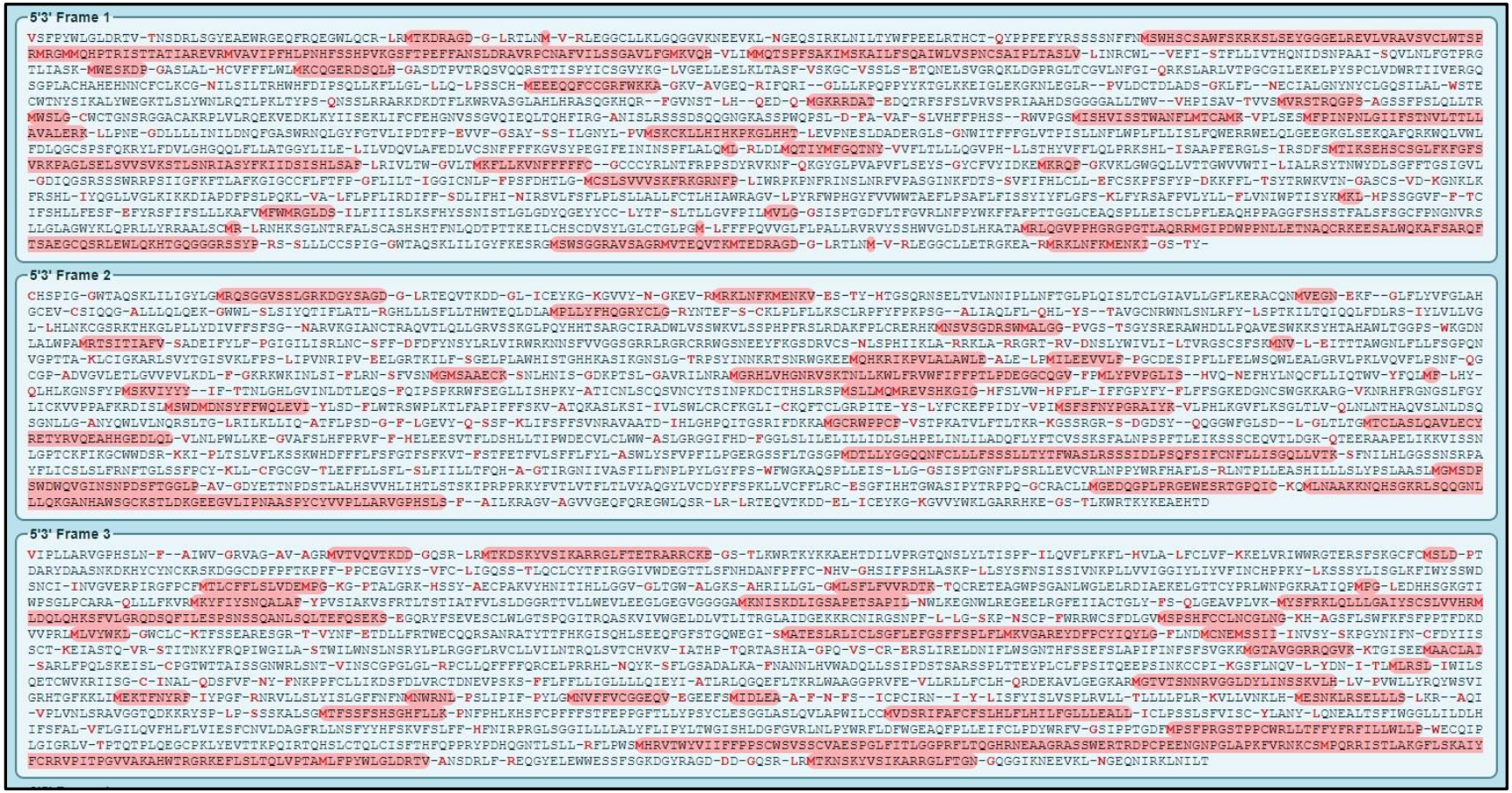


Figure S151. Single Letter FASTA Protein Sequence Translated from ERVmap ID 4249 Nucleotide Sequence Coding for Endogenous Retrovirus HERVIP10F

The figure above shows translated single letter FASTA protein sequence for ERVmap region 4249. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

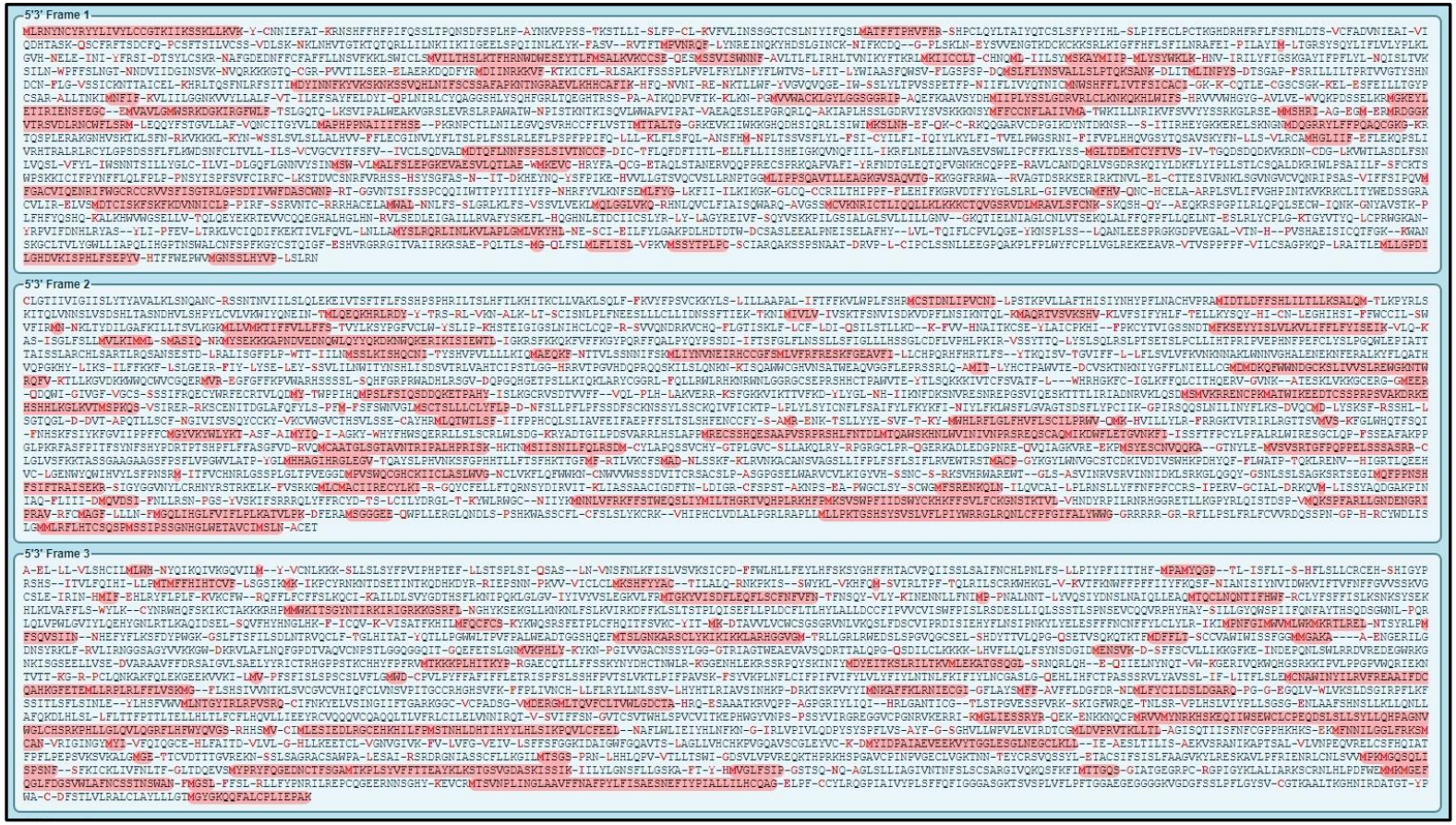


Figure S152. Single Letter FASTA Protein Sequence Translated from ERVmap ID 3866 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-K22

The figure above shows translated single letter FASTA protein sequence for ERVmap region 3866. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S153. Single Letter FASTA Protein Sequence Translated from ERVmap ID 4444 Nucleotide Sequence Coding for Endogenous Retrovirus Harlequin

The figure above shows translated single letter FASTA protein sequence for ERVmap region 4444. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

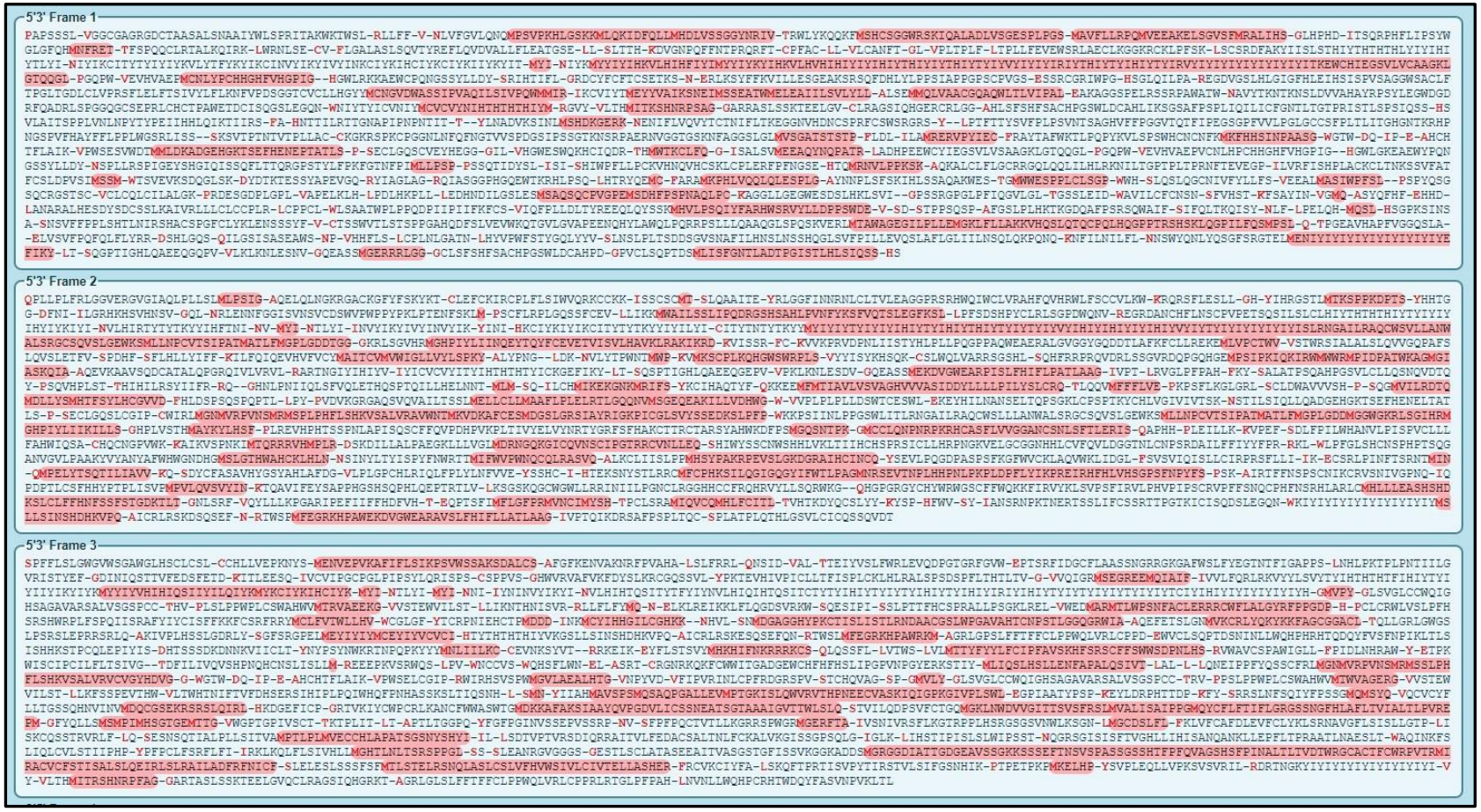


Figure S154. Single Letter FASTA Protein Sequence Translated from ERVmap ID 5633 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-L

The figure above shows translated single letter FASTA protein sequence for ERVmap region 5633. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

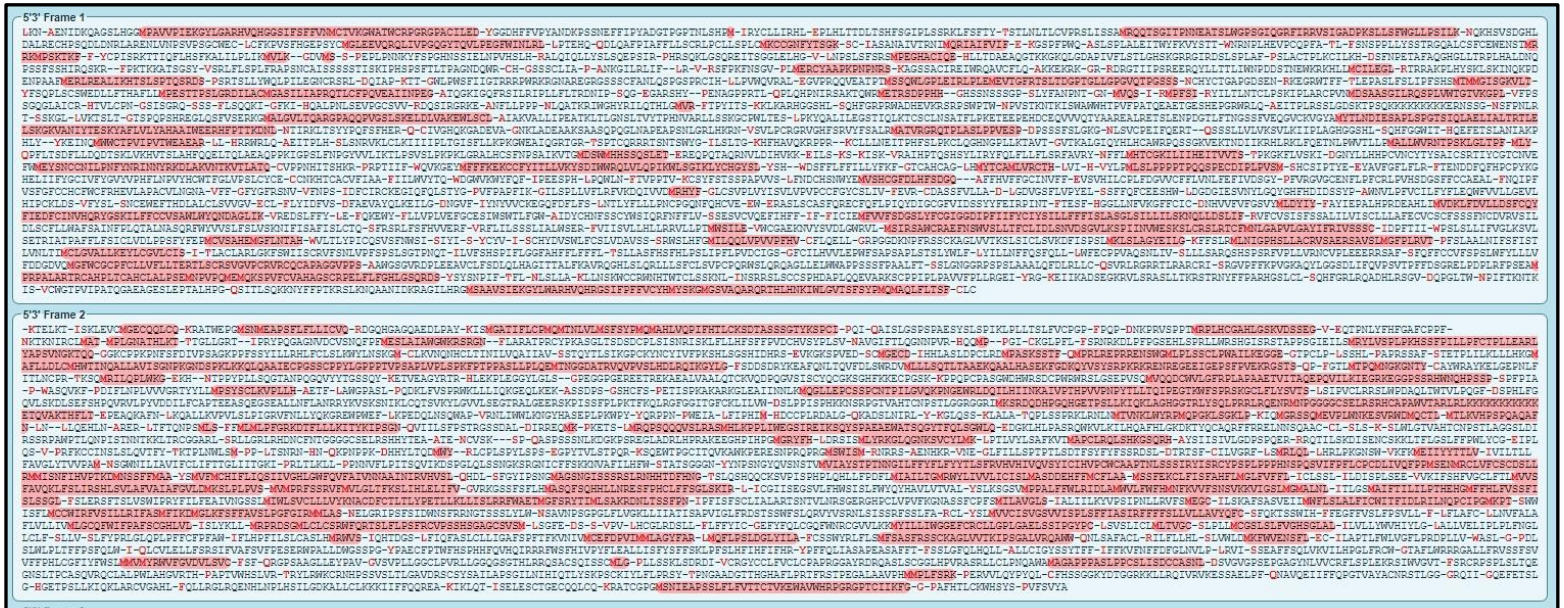


Figure S155. Single Letter FASTA Protein Sequence Translated from ERVmap ID 3547 Nucleotide Sequence Coding for Endogenous Retrovirus HUERS-P3
The figure above shows translated single letter FASTA protein sequence for ERVmap region 3547. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

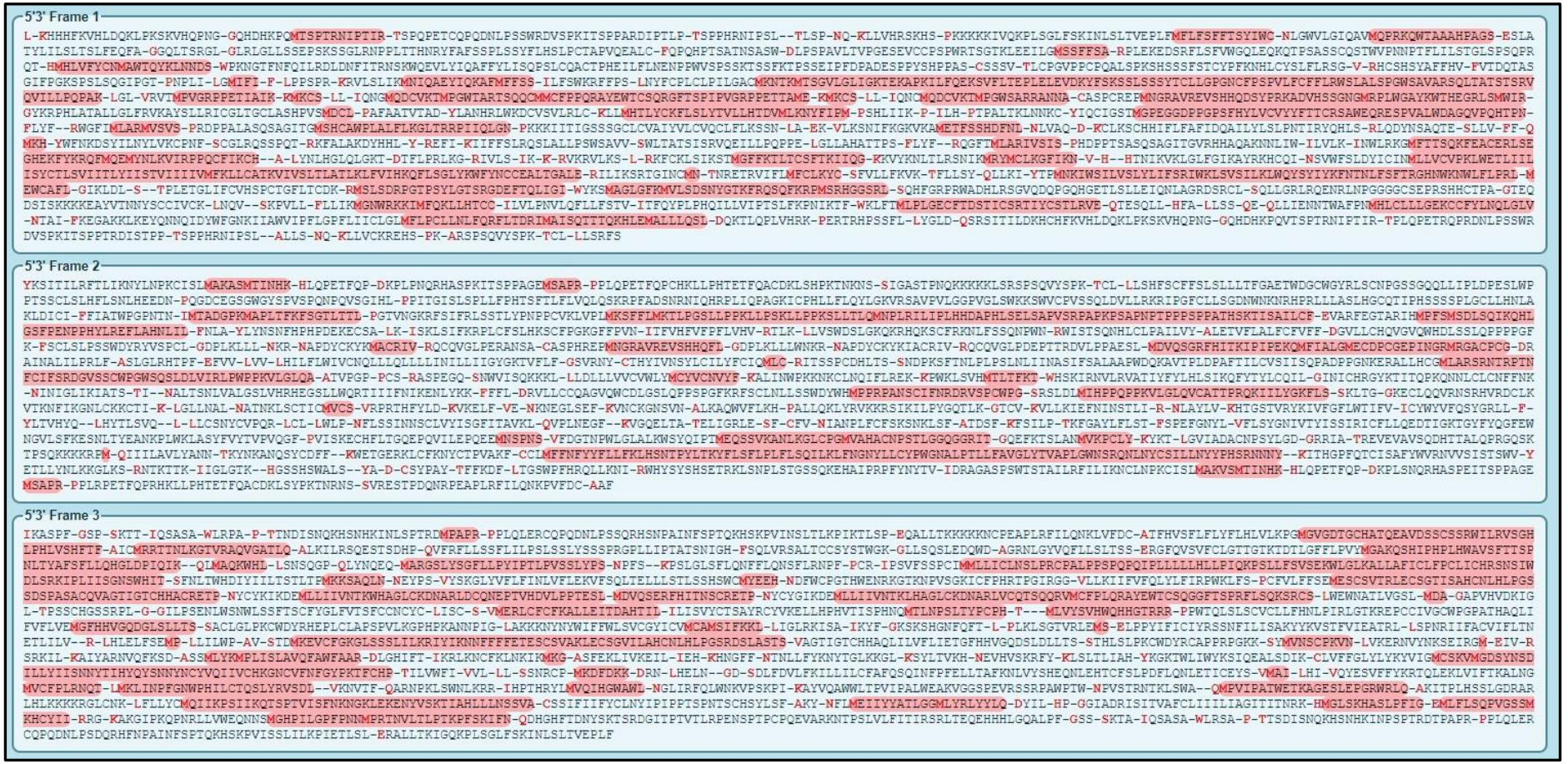


Figure S156. Single Letter FASTA Protein Sequence Translated from ERVmap ID 4180 Nucleotide Sequence Coding for Endogenous Retrovirus LTR19

The figure above shows translated single letter FASTA protein sequence for ERVmap region 4180. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

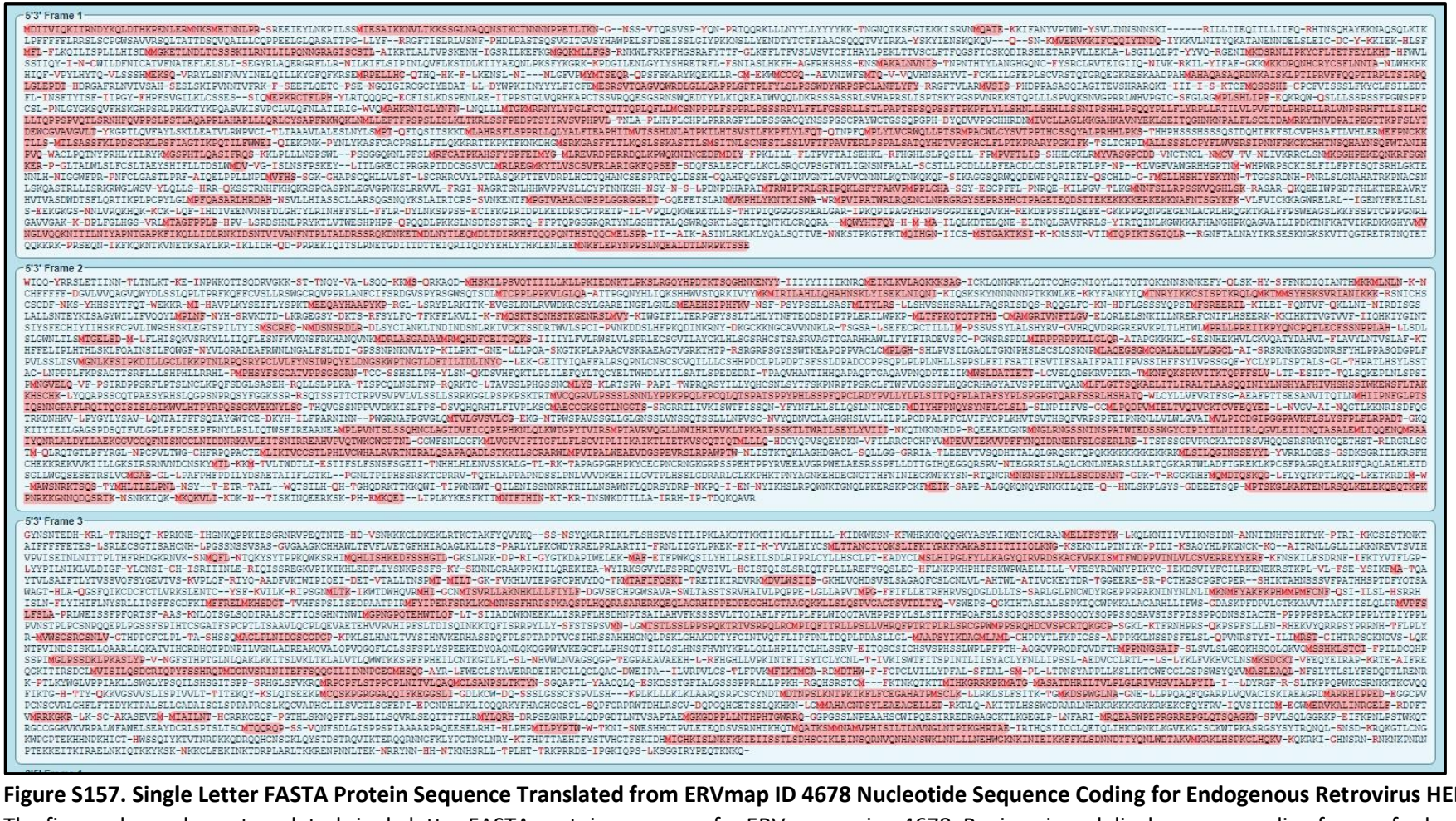


Figure S157. Single Letter FASTA Protein Sequence Translated from ERVmap ID 4678 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-H

The figure above shows translated single letter FASTA protein sequence for ERVmap region 4678. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S158. Single Letter FASTA Protein Sequence Translated from ERVmap ID 3776 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-K22

The figure above shows translated single letter FASTA protein sequence for ERVmap region 3776. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

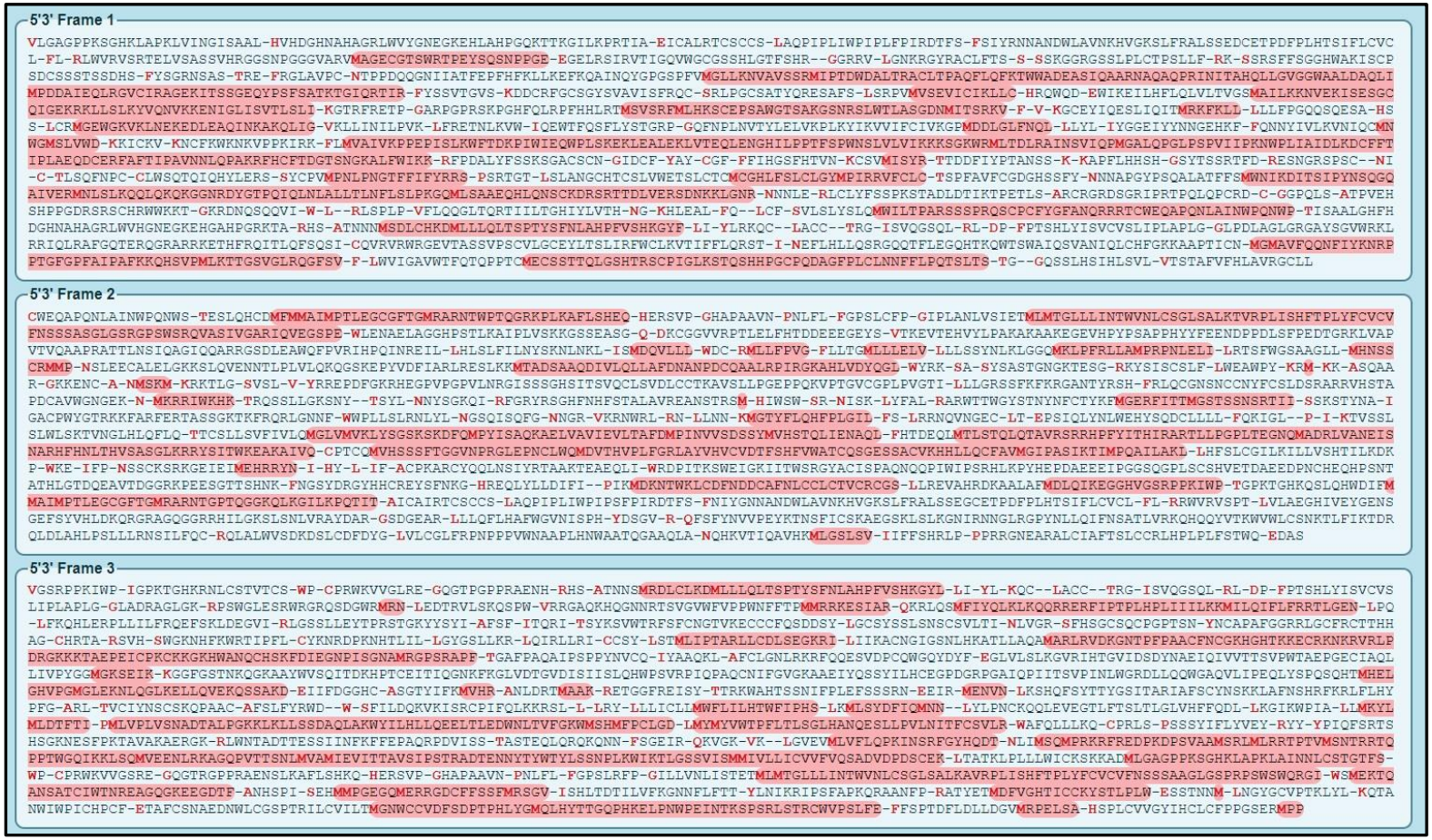


Figure S159. Single Letter FASTA Protein Sequence Translated from ERVmap ID 3167 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-K9

The figure above shows translated single letter FASTA protein sequence for ERVmap region 3167. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single “-“.

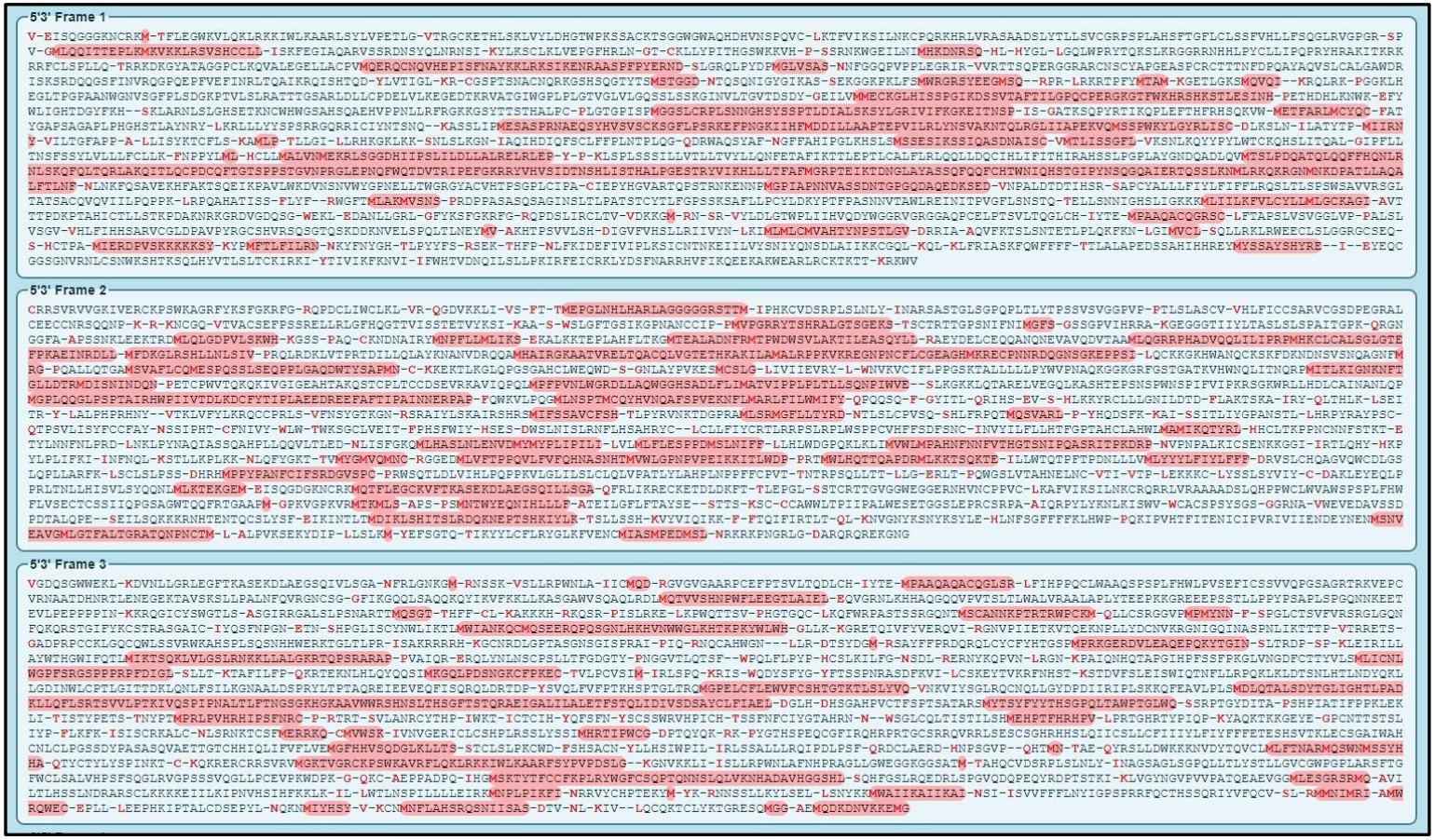


Figure S160. Single Letter FASTA Protein Sequence Translated from ERVmap ID 673 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-K22

The figure above shows translated single letter FASTA protein sequence for ERVmap region 673. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

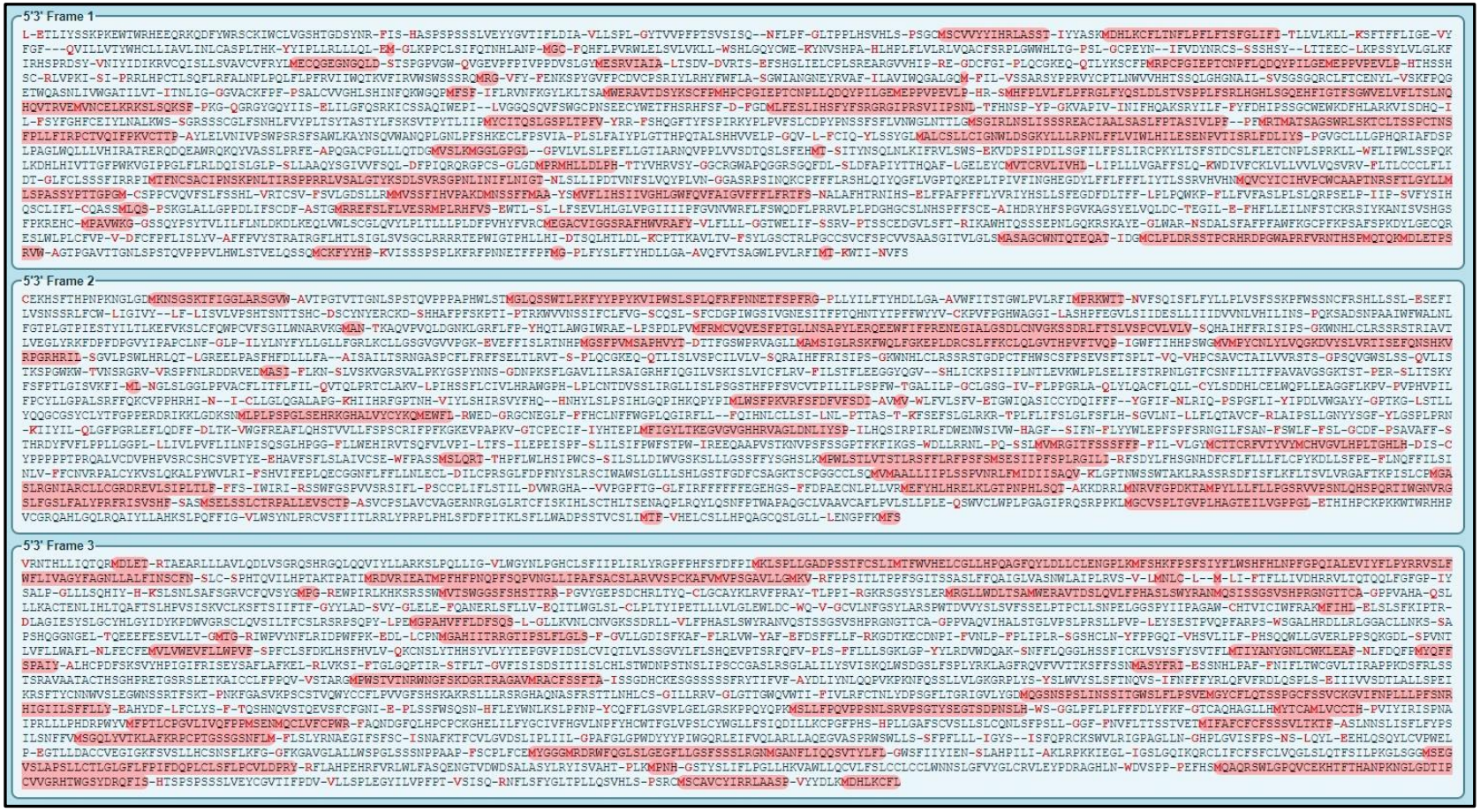


Figure S161. Single Letter FASTA Protein Sequence Translated from ERVmap ID 857 Nucleotide Sequence Coding for Endogenous Retrovirus HERVP71A
The figure above shows translated single letter FASTA protein sequence for ERVmap region 857. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S162. Single Letter FASTA Protein Sequence Translated from ERVmap ID 4060 Nucleotide Sequence Coding for Endogenous Retrovirus HERV9

The figure above shows translated single letter FASTA protein sequence for ERVmap region 4060. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

Figure S163. Single Letter FASTA Protein Sequence Translated from ERVmap ID 5446 Nucleotide Sequence Coding for Endogenous Retrovirus HERVK9

The figure above shows translated single letter FASTA protein sequence for ERVmap region 5446. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

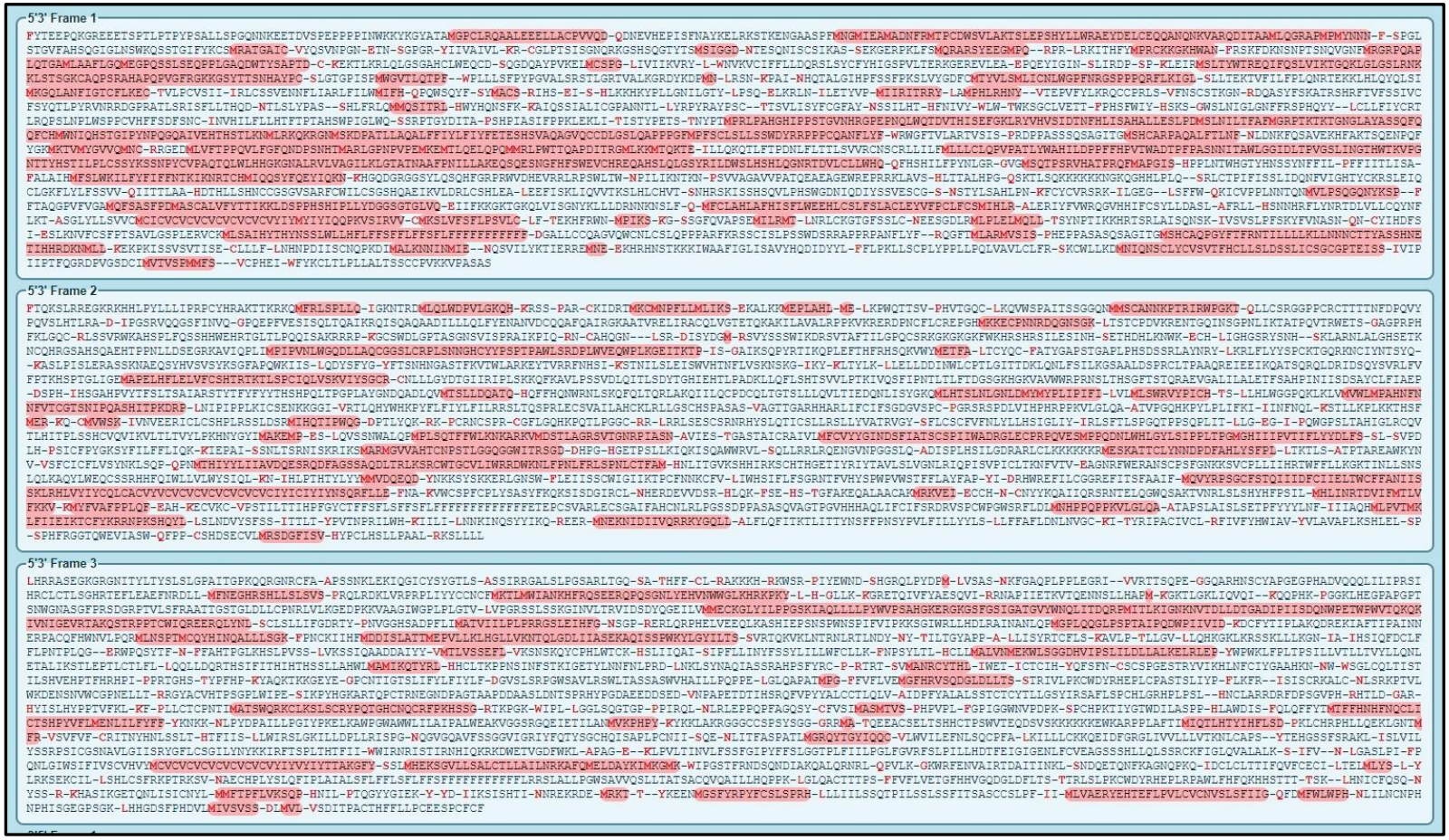


Figure S164. Single Letter FASTA Protein Sequence Translated from ERVmap ID 3606 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-K22

The figure above shows translated single letter FASTA protein sequence for ERVmap region 3606. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

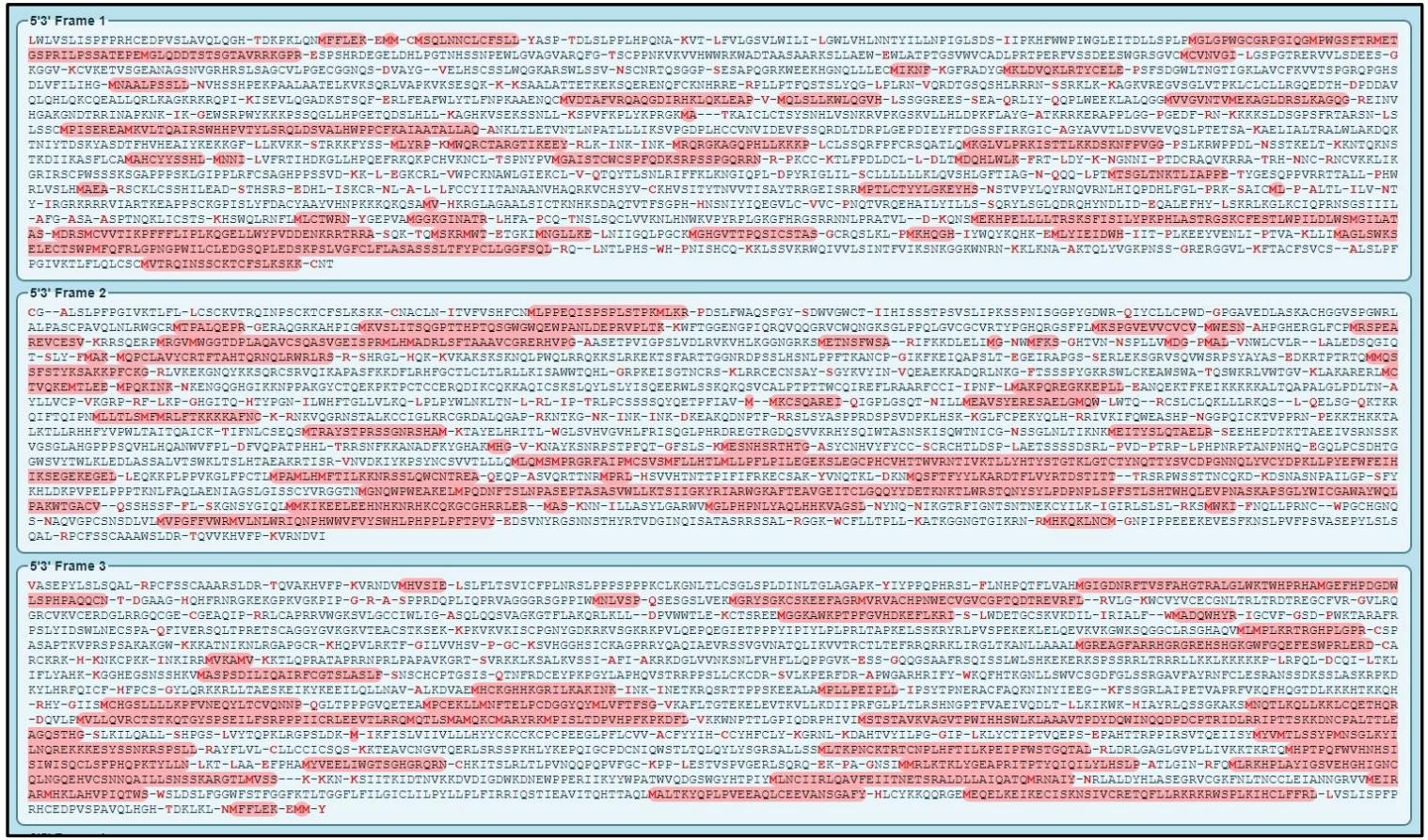


Figure S165. Single Letter FASTA Protein Sequence Translated from ERVmap ID 4352 Nucleotide Sequence Coding for Endogenous Retrovirus HERV15

The figure above shows translated single letter FASTA protein sequence for ERVmap region 4352. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

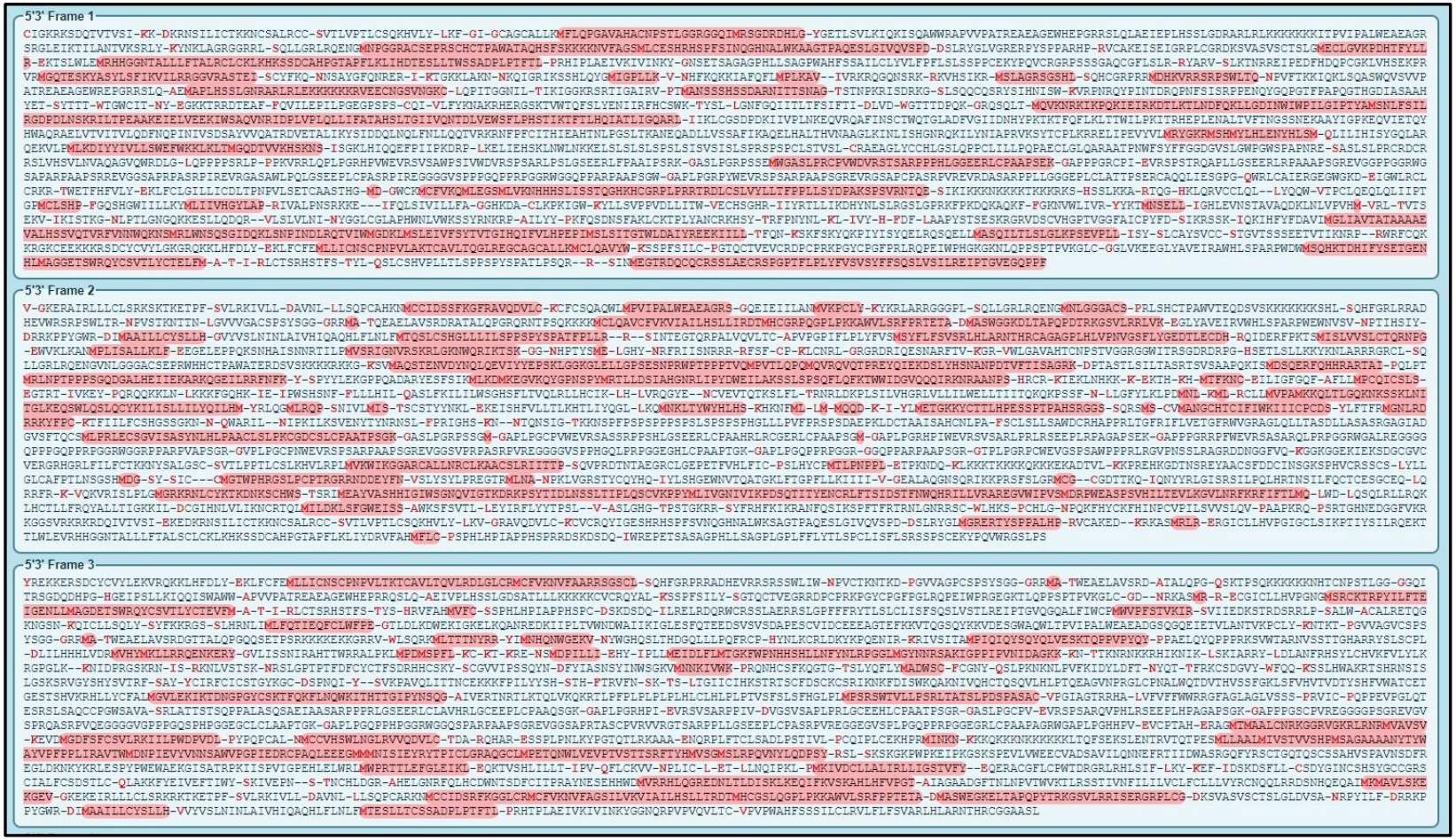


Figure S166. Single Letter FASTA Protein Sequence Translated from ERVmap ID K-46 Nucleotide Sequence Coding for Endogenous Retrovirus HERV-K47

The figure above shows translated single letter FASTA protein sequence for ERVmap region K-46. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

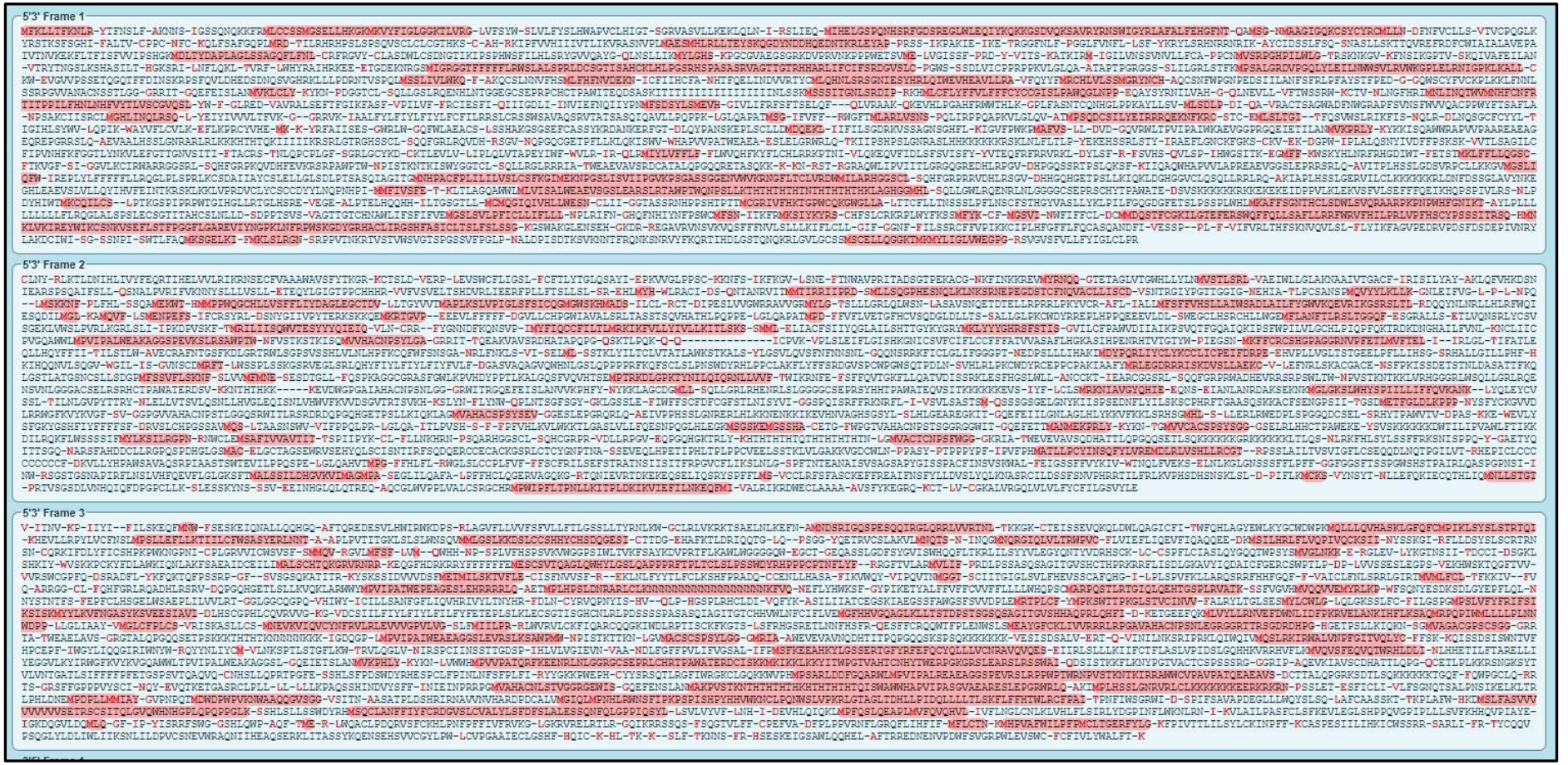


Figure S167. Single Letter FASTA Protein Sequence Translated from ERVmap ID 2916 Nucleotide Sequence Coding for Endogenous Retrovirus MER57A

The figure above shows translated single letter FASTA protein sequence for ERVmap region 2916. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

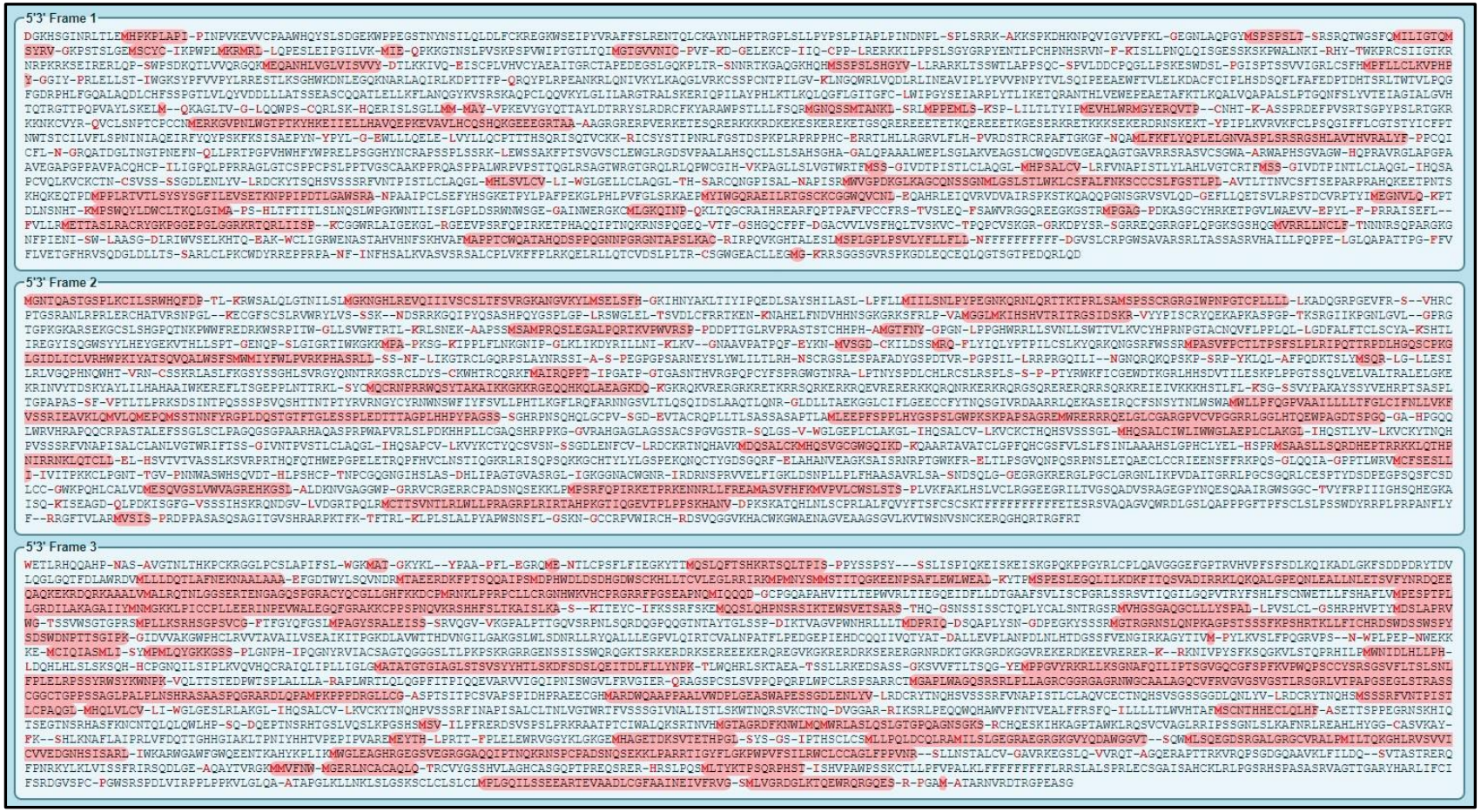


Figure S168. Single Letter FASTA Protein Sequence Translated from ERVmap ID 1728 Nucleotide Sequence Coding for Endogenous Retrovirus HERV9NC

The figure above shows translated single letter FASTA protein sequence for ERVmap region 1728. Regions in red display open reading frames for hypothetical proteins starting at a Methionine and ending with a stop codon, represented by a single "-".

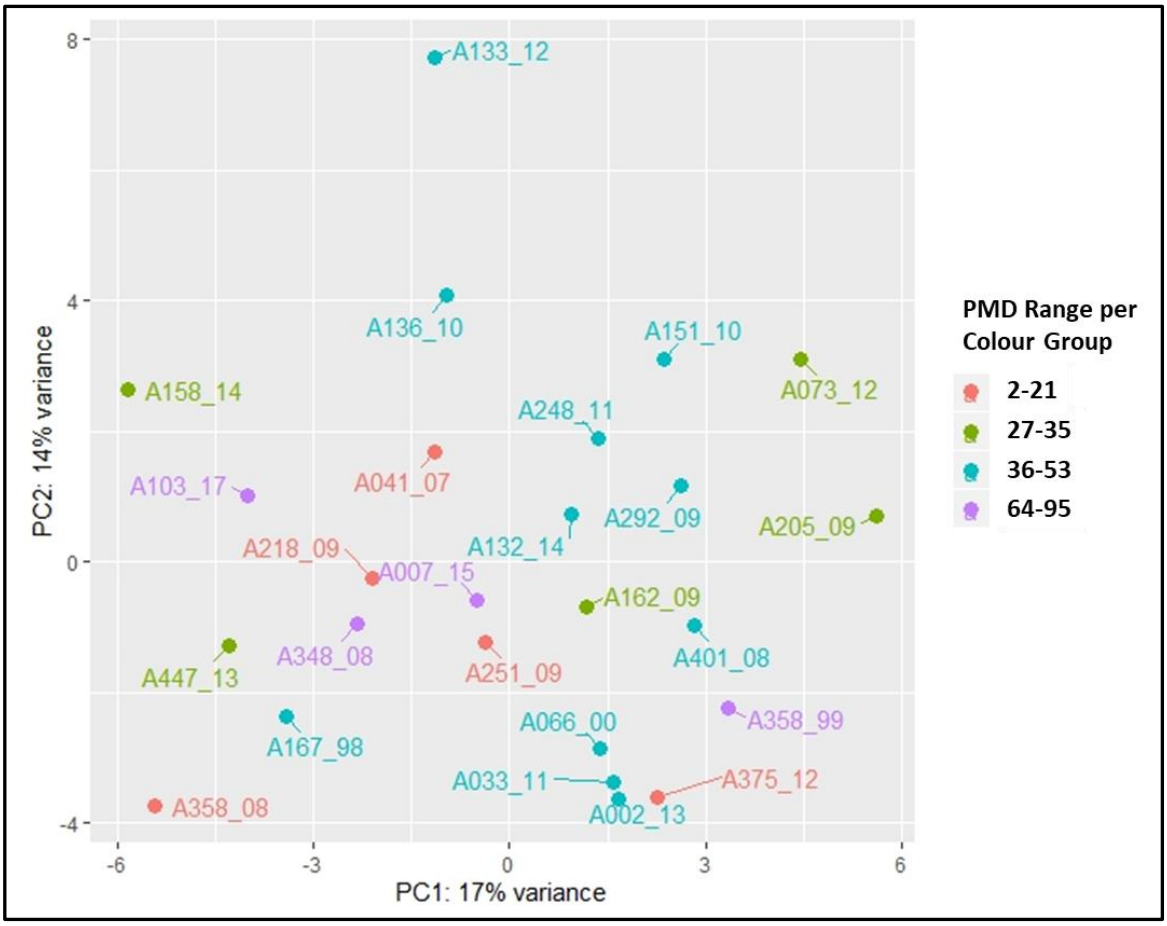


Figure S169. PCA Plot of ALS and non-ALS Control Postmortem Primary Motor Cortex Tissue Samples Coloured by Postmortem Delay in Hours.

The principle component analysis (PCA) shown in the figure above plots variation between all ERVs in each sample using normalised counts for each gene per patient sample. These principle component values are then plotted against each other to group samples together based on similar expression profiles of ERVs.

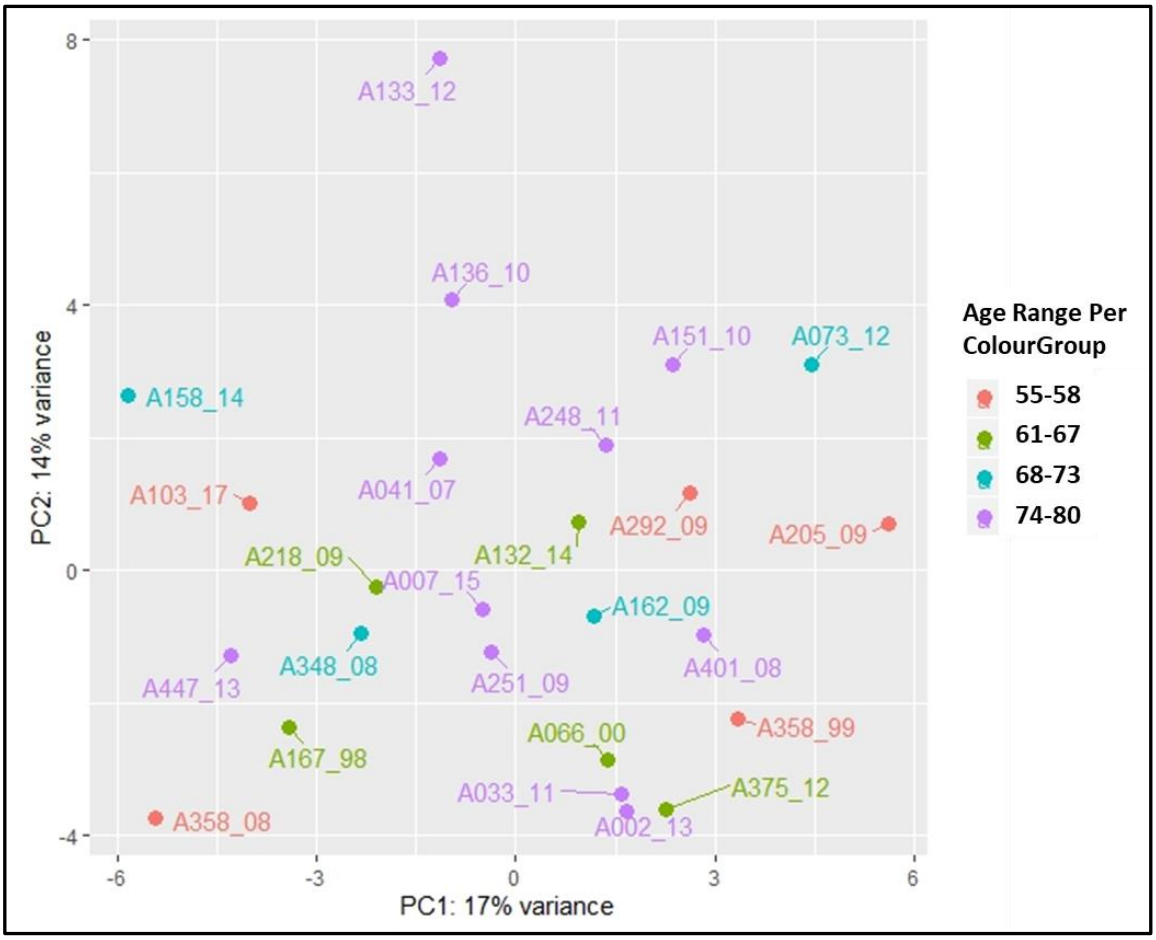


Figure S170. PCA Plot of ALS and non-ALS Control Postmortem Primary Motor Cortex Tissue Samples Coloured by Patient Age at time of Death.

The principle component analysis (PCA) shown in the figure above plots variation between all ERVs in each sample using normalised counts for each gene per patient sample. These principle component values are then plotted against each other to group samples together based on similar expression profiles of ERVs.

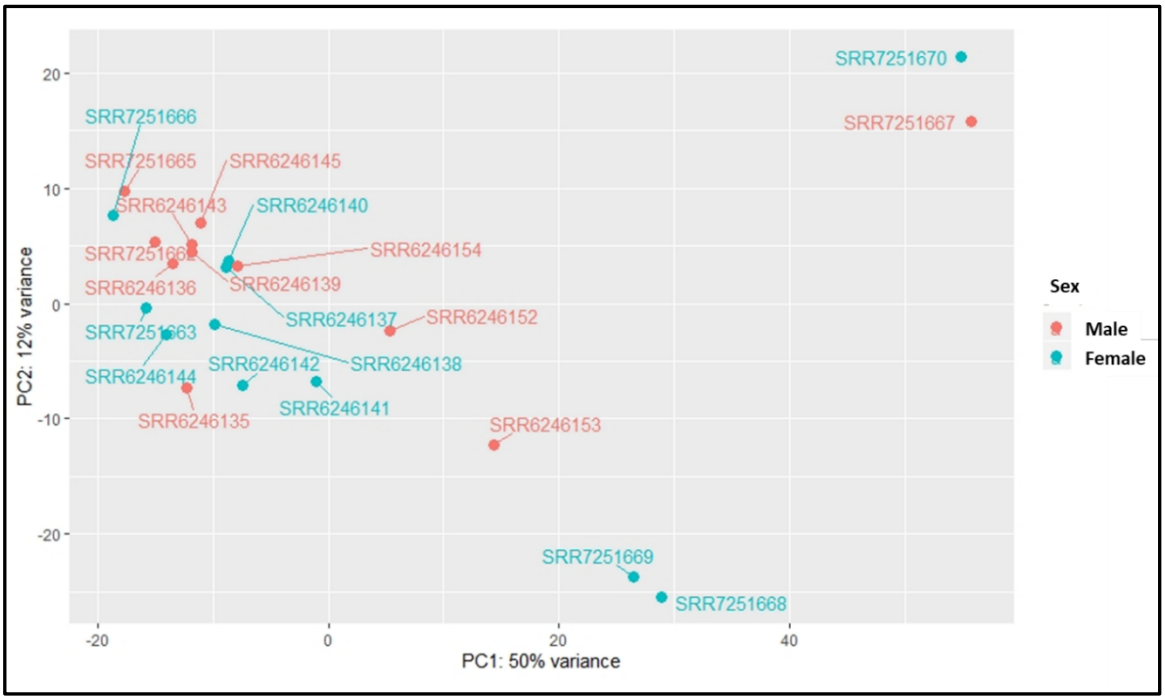


Figure S171. PCA Plot of ALS and non-ALS Control Peripheral Blood Mononuclear Cell Samples Coloured by Patient Sex.

The principle component analysis (PCA) shown in the figure above plots variation between all ERVs in each sample using normalised counts for each gene per patient sample. These principle component values are then plotted against each other to group samples together based on similar expression profiles of ERVs.

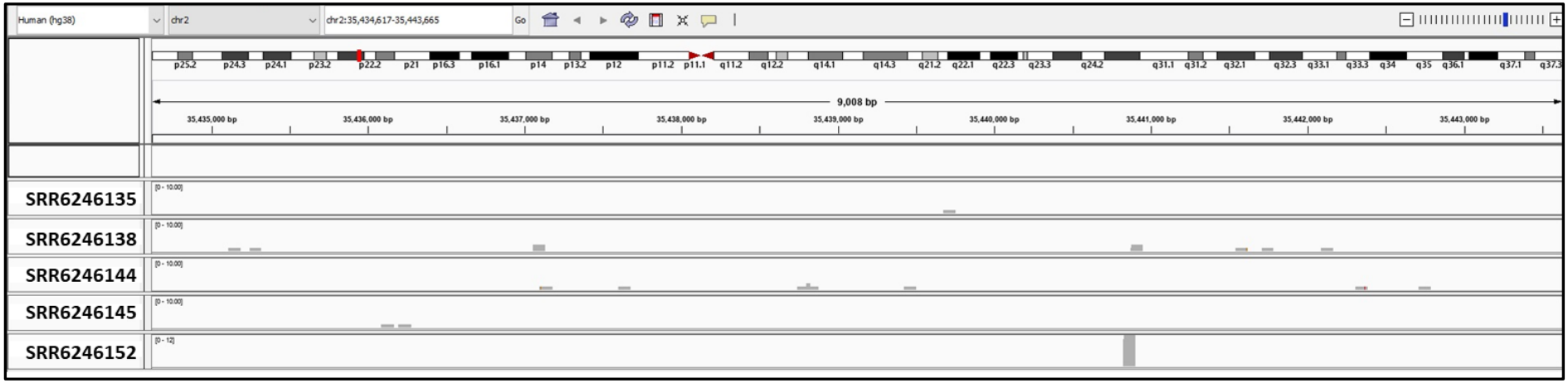


Figure S172. Read Alignment Coverage for ERVMap 570 (HERV-H)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 570 (Chromosome 2, locus p22.3), identified as HERV-H. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

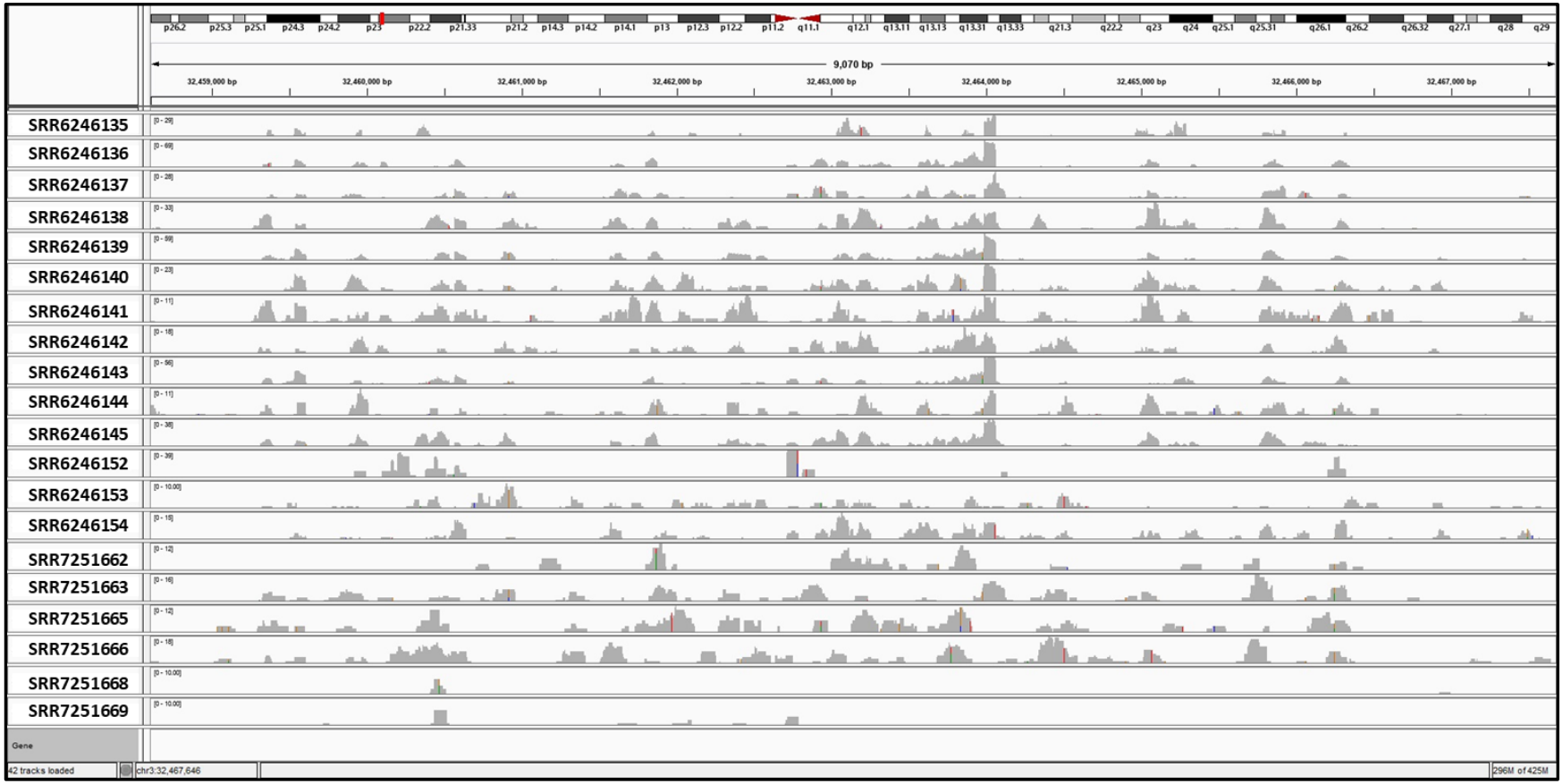


Figure S173. Read Alignment Coverage for ERVMap 909 (HERV-H)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 909 (Chromosome 3, locus p22.3), identified as HERV-H. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

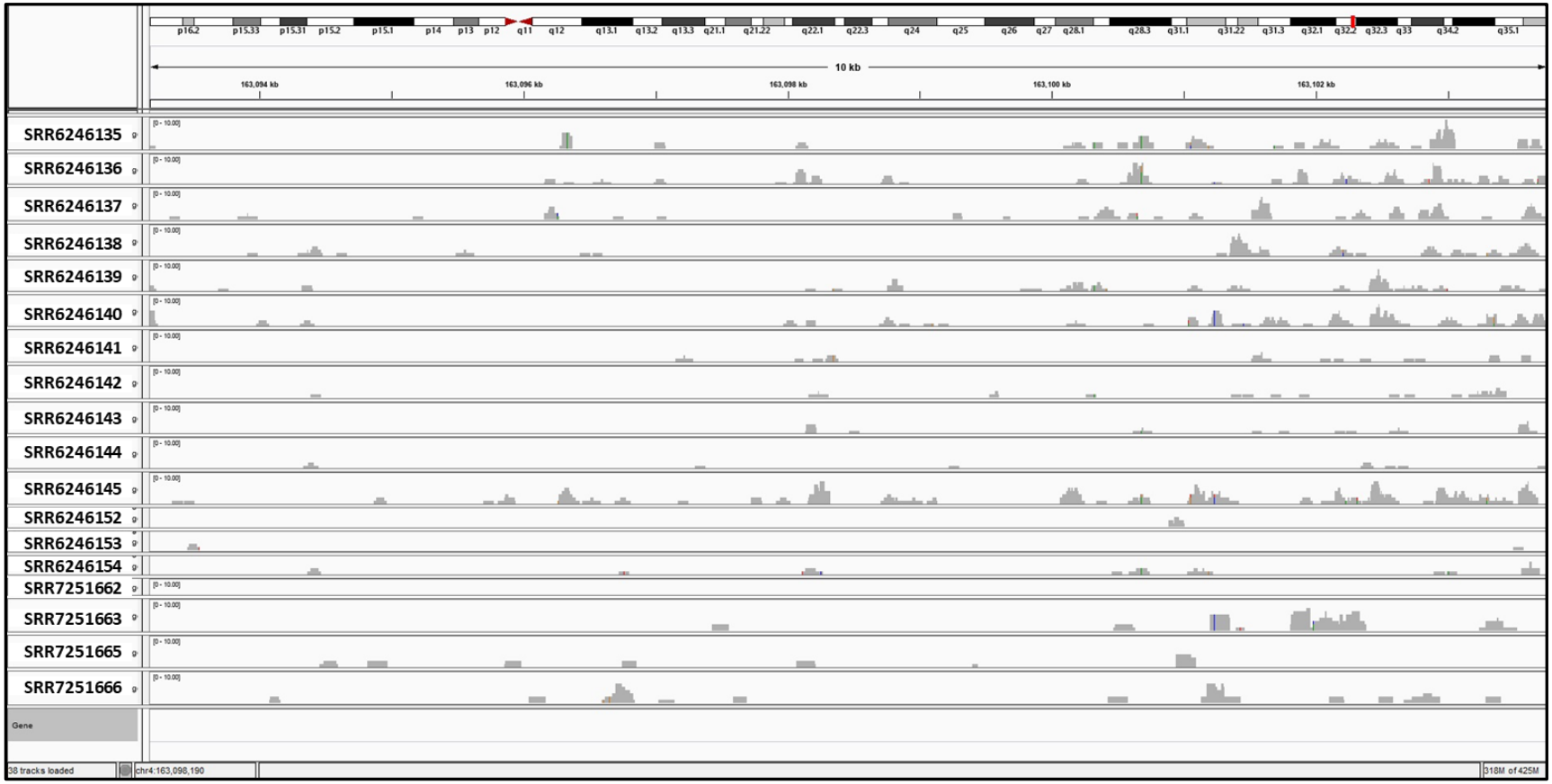


Figure S174. Read Alignment Coverage for ERVMap 1679 (HERV-H)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 1679 (Chromosome 4, locus q32.2), identified as HERV-H. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

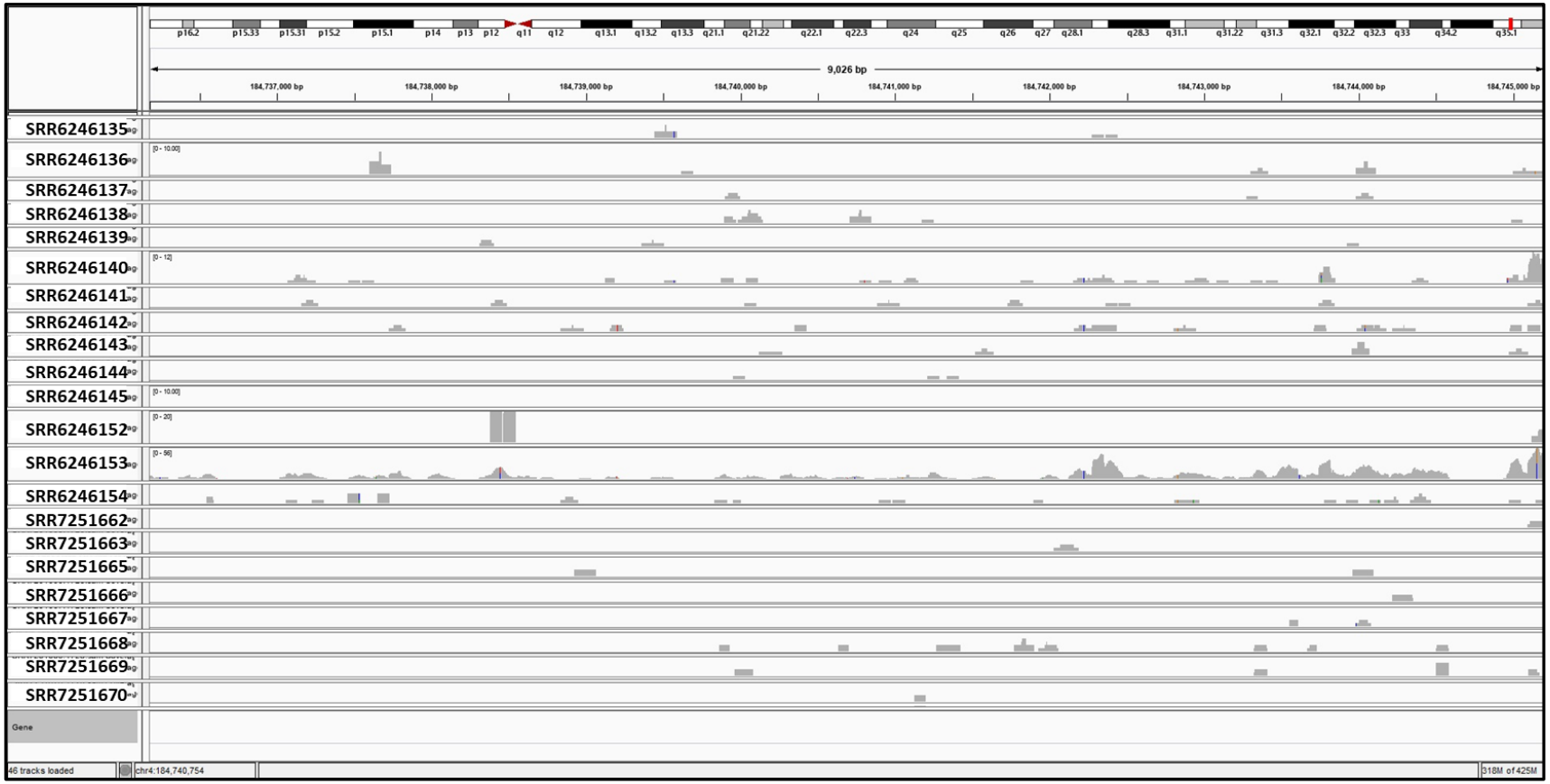


Figure S175. Read Alignment Coverage for ERVMap 1728 (HERV9NC)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 1728 (Chromosome 4, locus q35.1), identified as HERV9NC. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

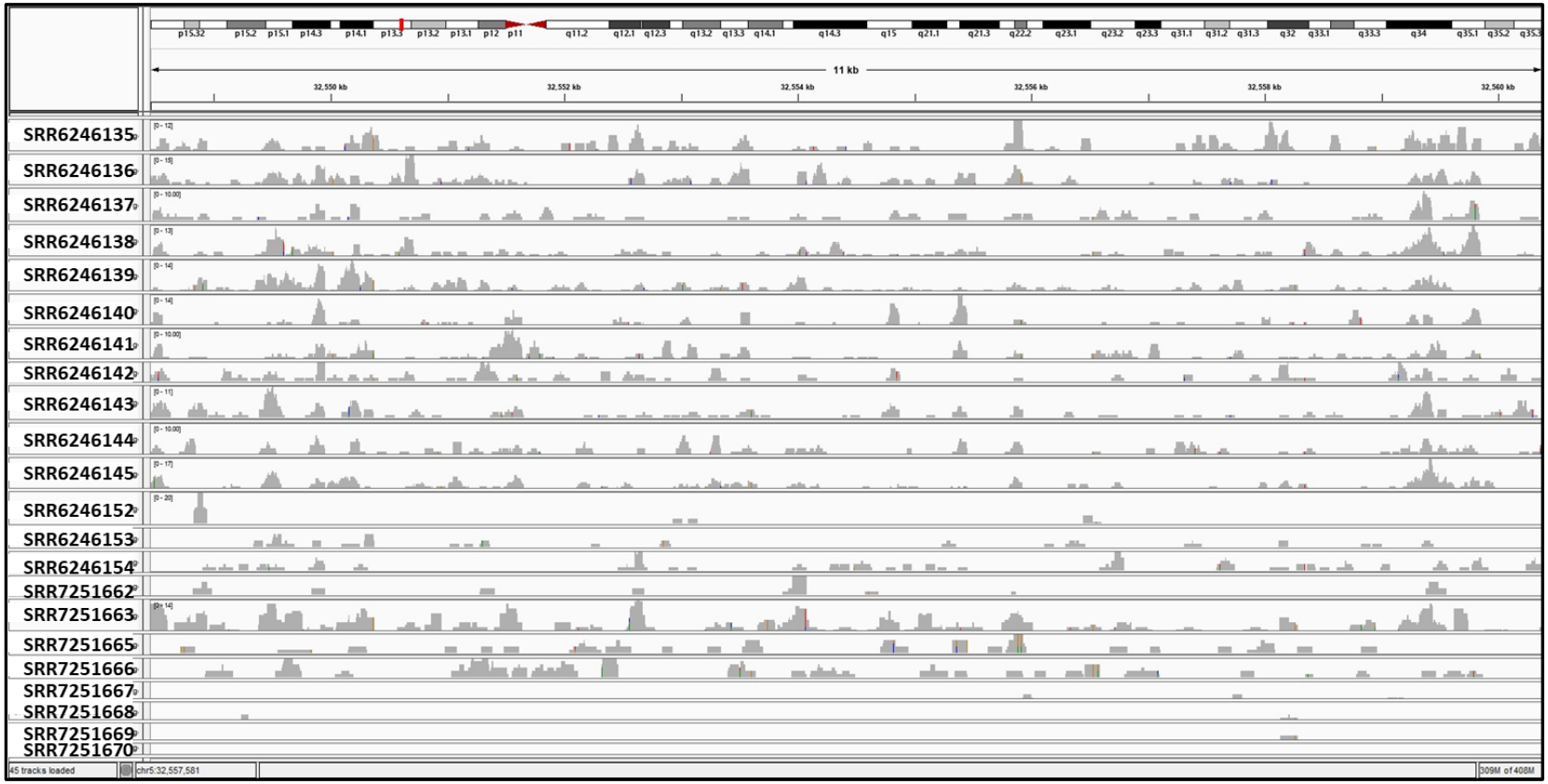


Figure S176. Read Alignment Coverage for ERVMap 1797 (HERV9)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 1797 (Chromosome 5, locus p13.3), identified as HERV9. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

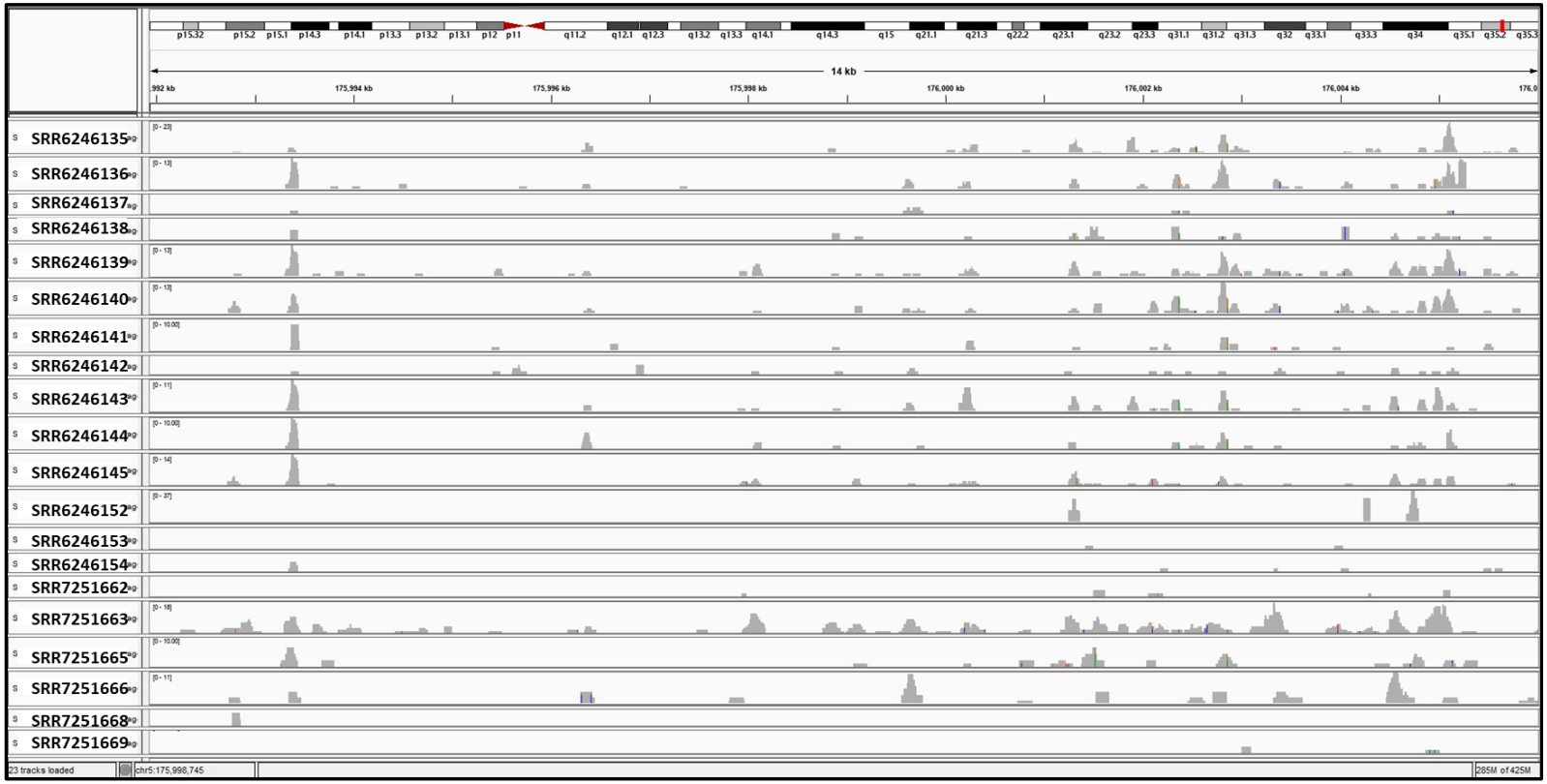


Figure S177. Read Alignment Coverage for ERVMap 2049 (HERV-L)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 2049 (Chromosome 5, locus q35.2), identified as HERV-L. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

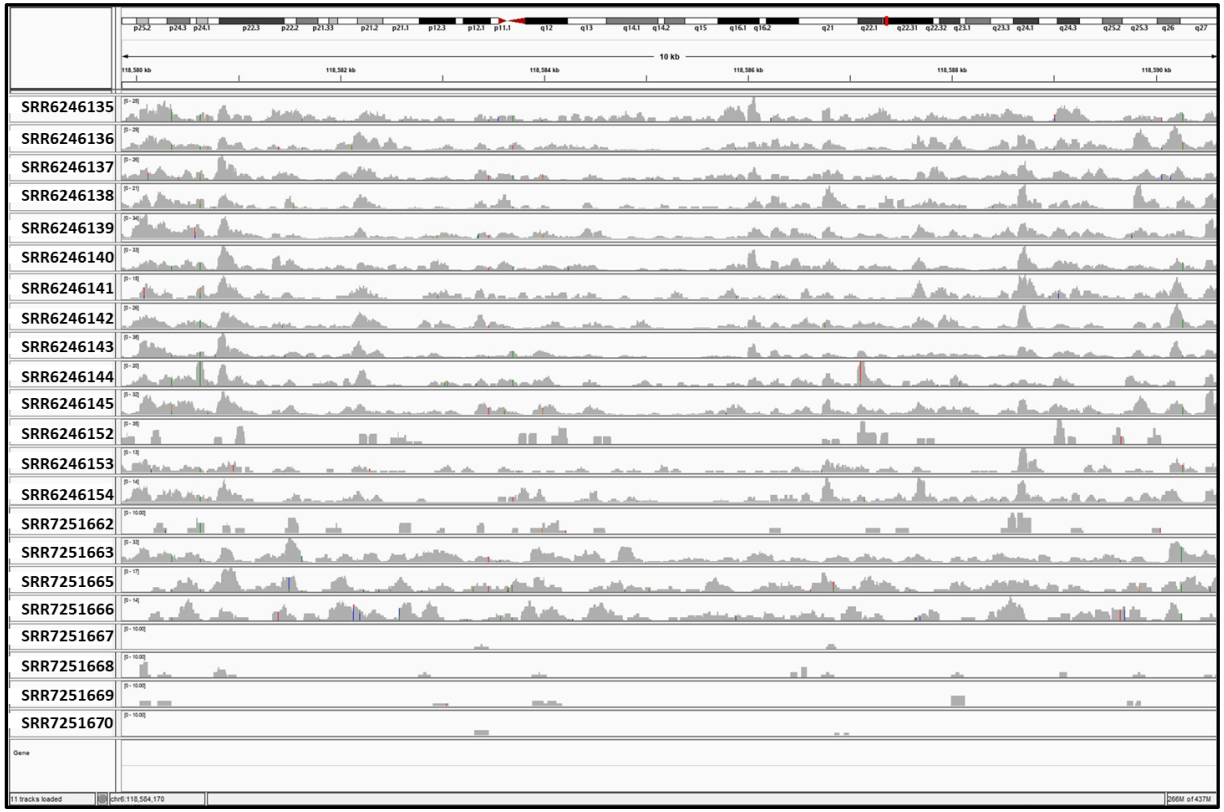


Figure S178. Read Alignment Coverage for ERVMap 2305 (HUERS-P3)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 2305 (Chromosome 6, locus q22.31), identified as HUERS-P3. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

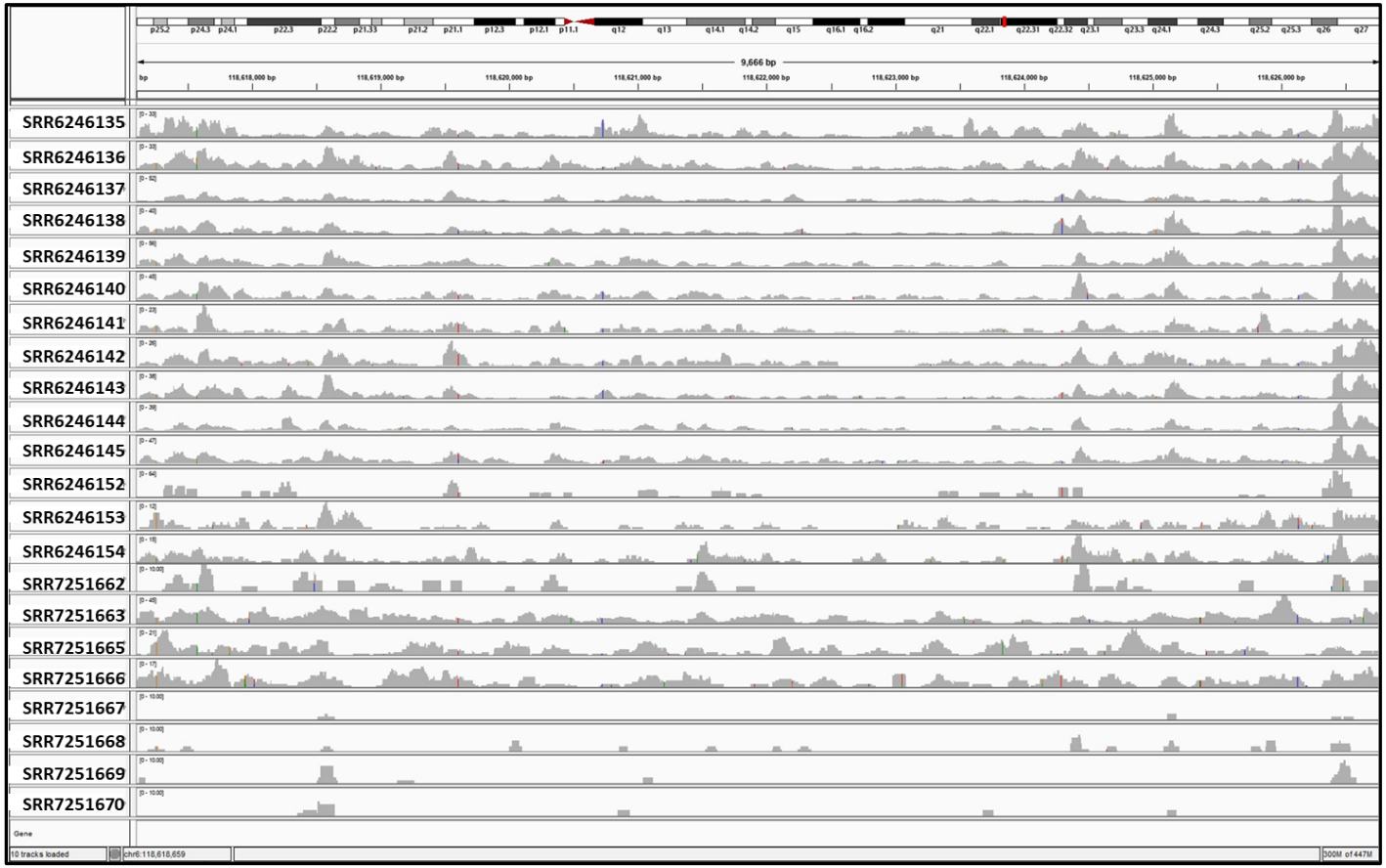


Figure S179. Read Alignment Coverage for ERVMap 2307 (HERV30)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 2307 (Chromosome 6, locus q22.31), identified as HERV30. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

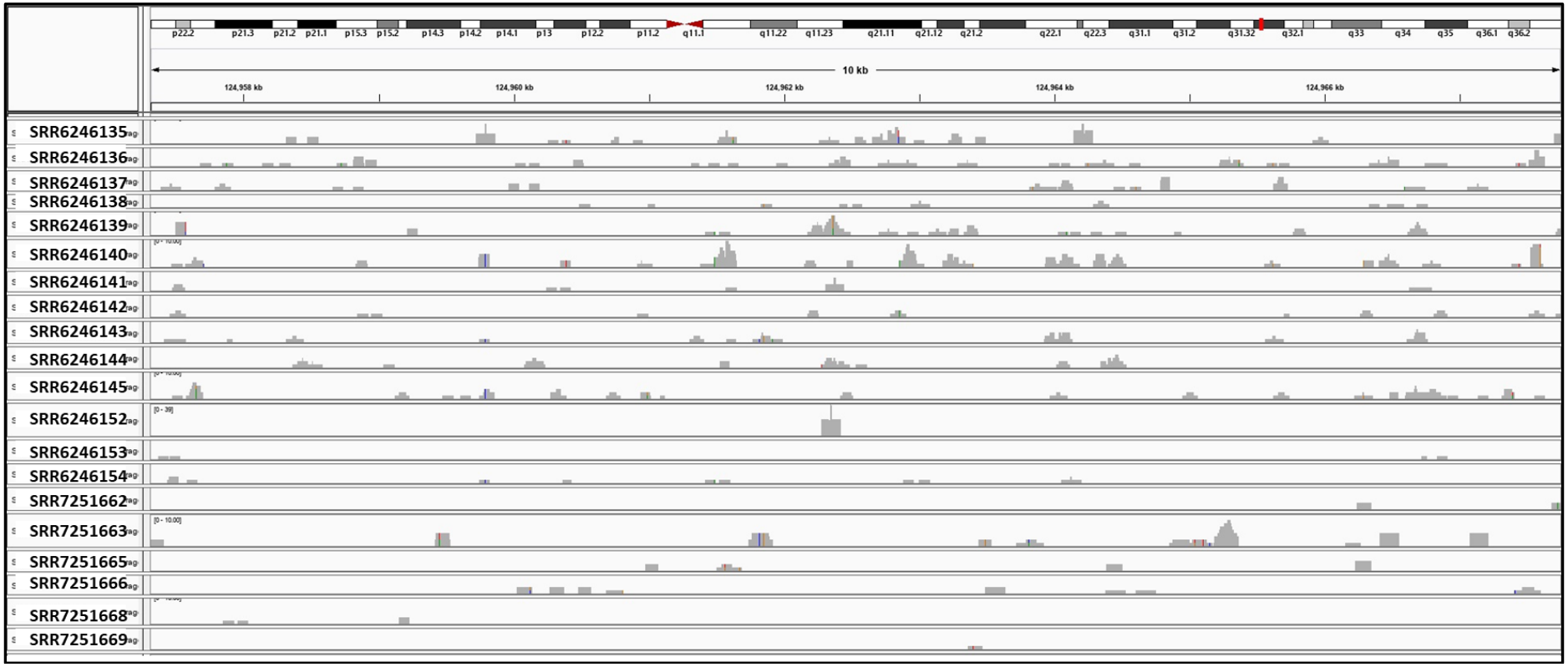


Figure S180. Read Alignment Coverage for ERVMap 2621 (HERV-H)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 2621 (Chromosome 7, locus q31.33), identified as HERV-H. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

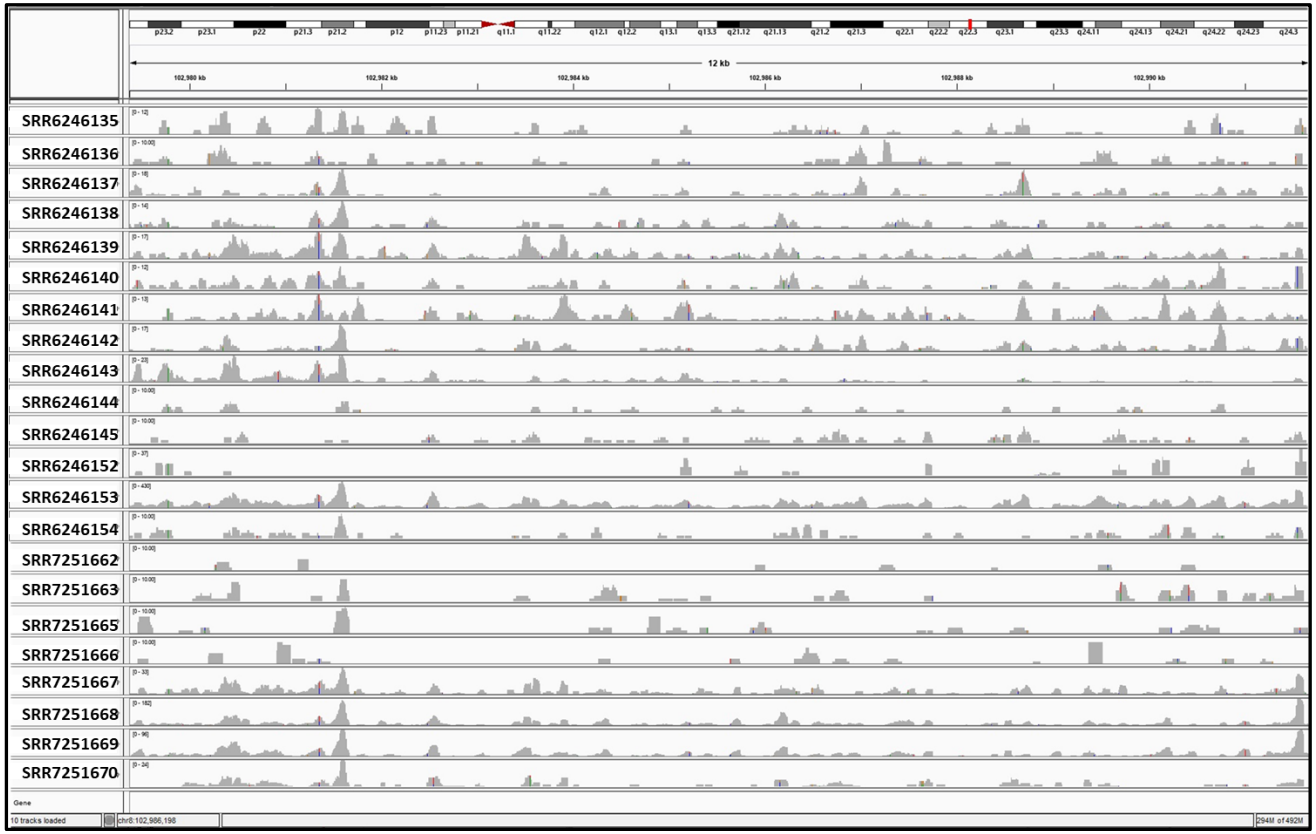


Figure S181. Read Alignment Coverage for ERVMap 2916 (MER57A)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 2916 (Chromosome 8, locus q22.3), identified as MER57A. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

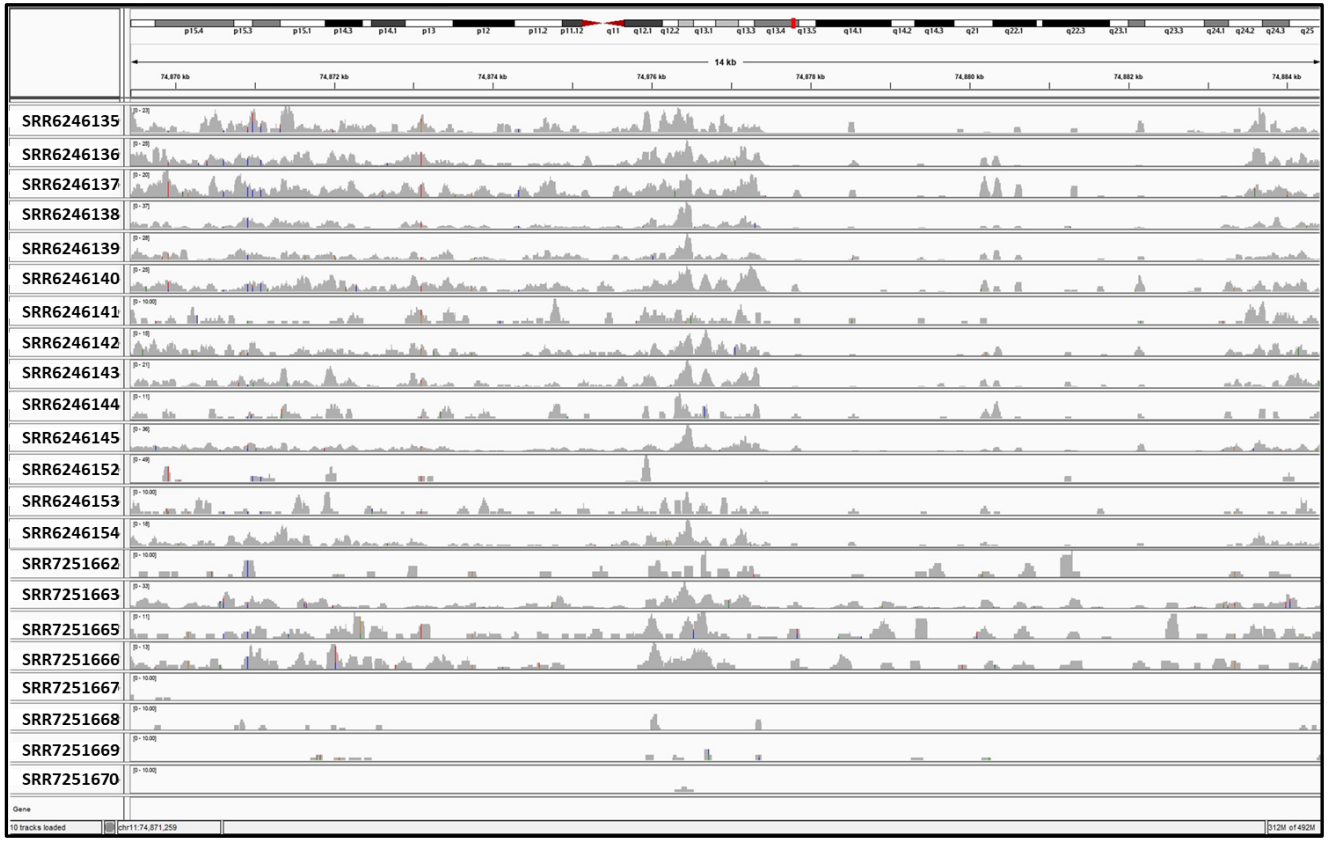


Figure S182. Read Alignment Coverage for ERVMap 3547 (HUERS-P3)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 3547 (Chromosome 11, locus q13.4), identified as HUERS-P3. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

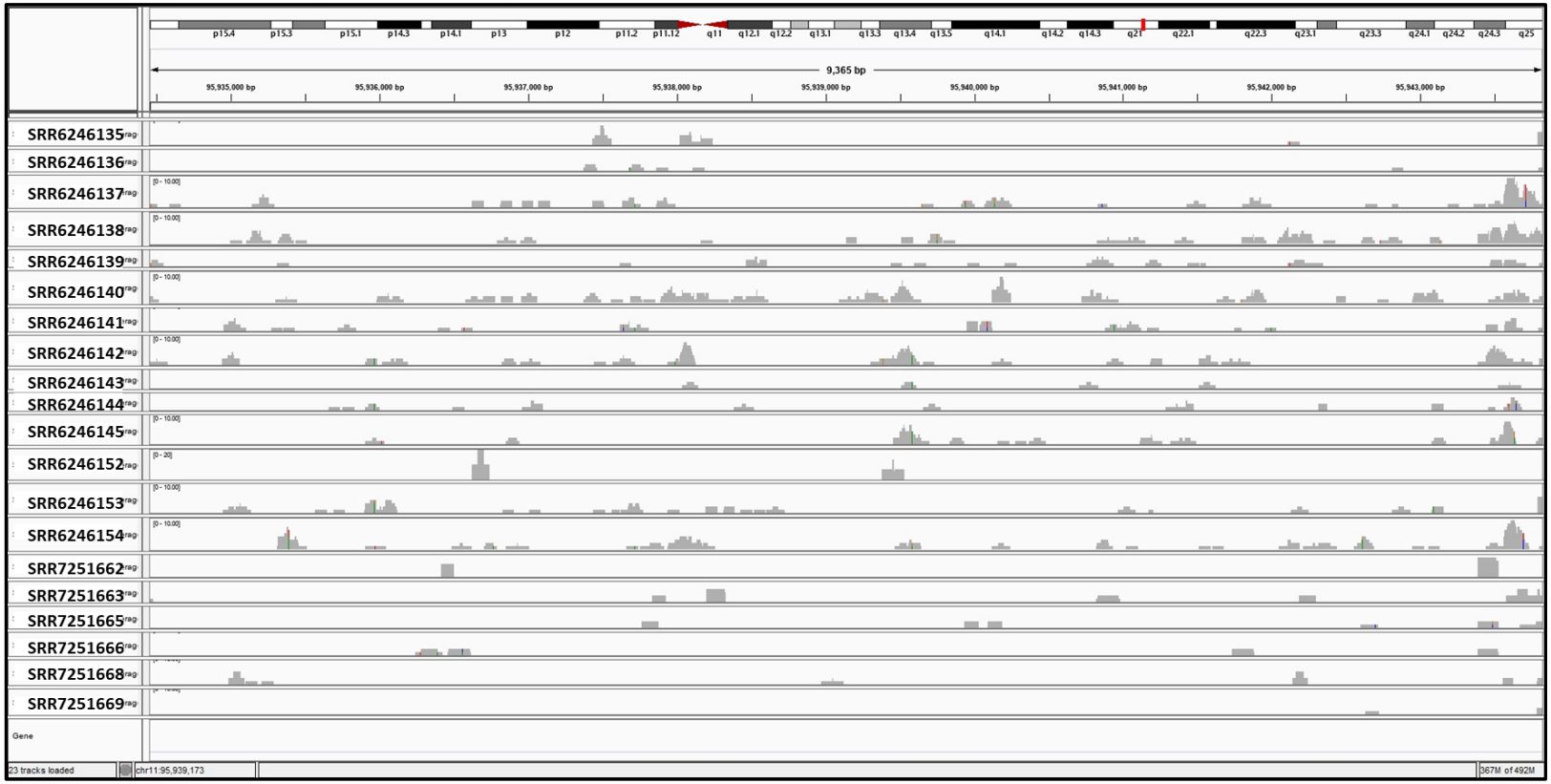


Figure S183. Read Alignment Coverage for ERVMap 3606 (HERV-K22)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 3606 (Chromosome 11, locus q21), identified as HERV-K22. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

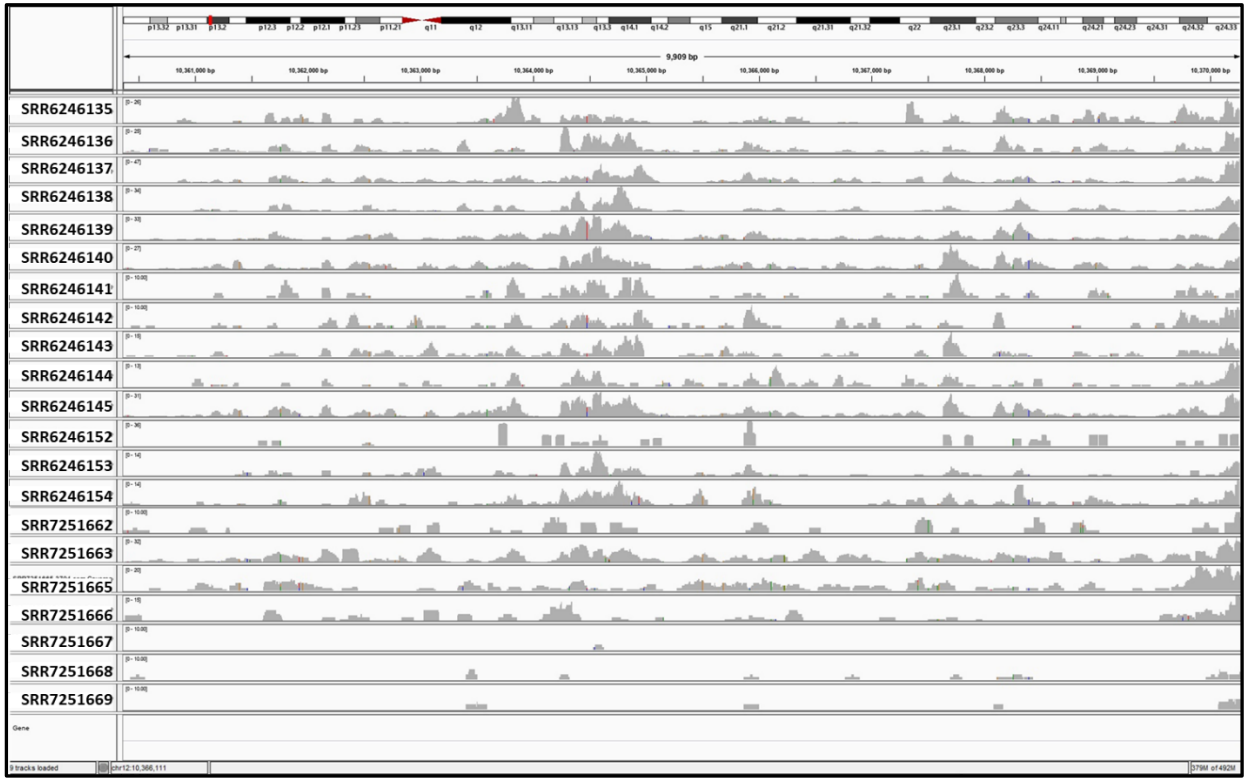


Figure S184. Read Alignment Coverage for ERVMap 3704 (HERV15)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 3704 (Chromosome 12, locus p13.2), identified as HERV15. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

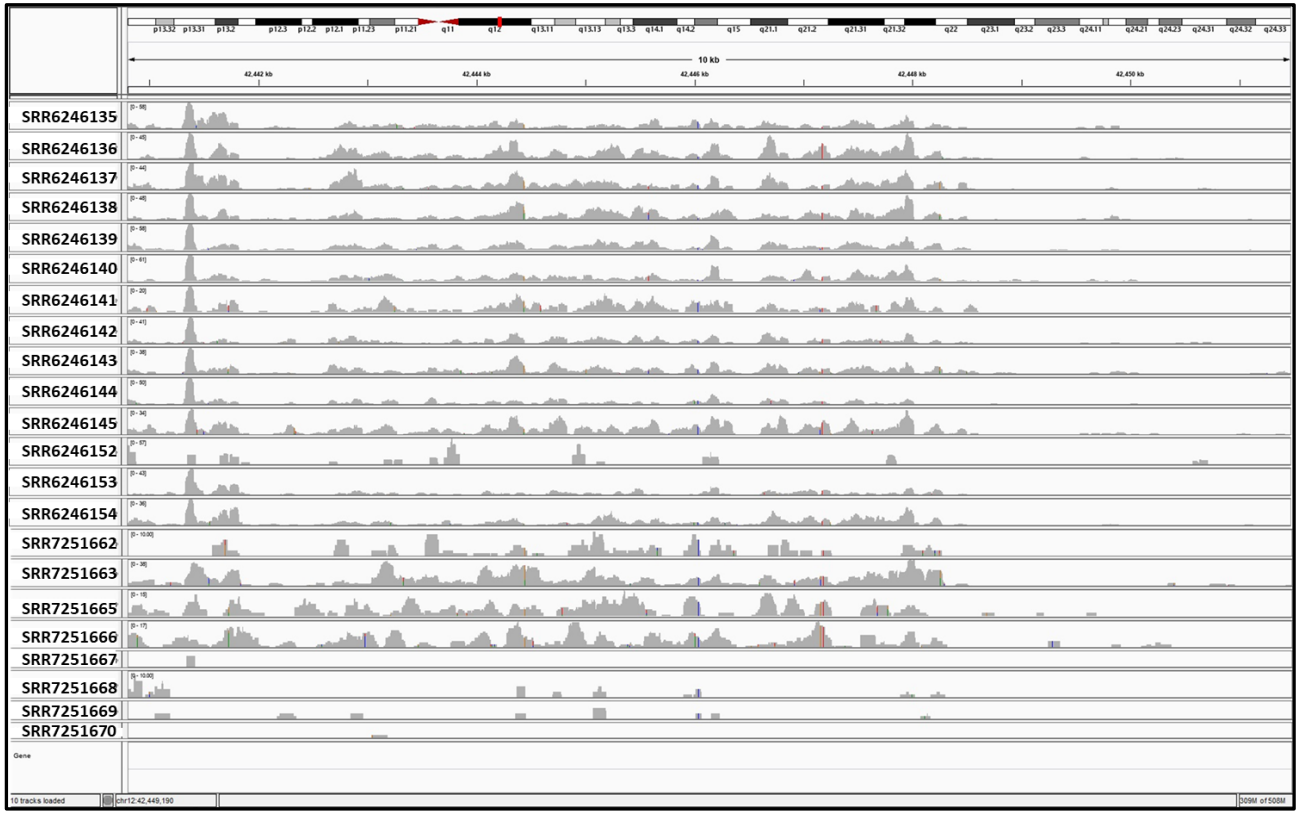


Figure S185. Read Alignment Coverage for ERVMap 3776 (HERV-K22)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 3776 (Chromosome 12, locus q12), identified as HERV-K22. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

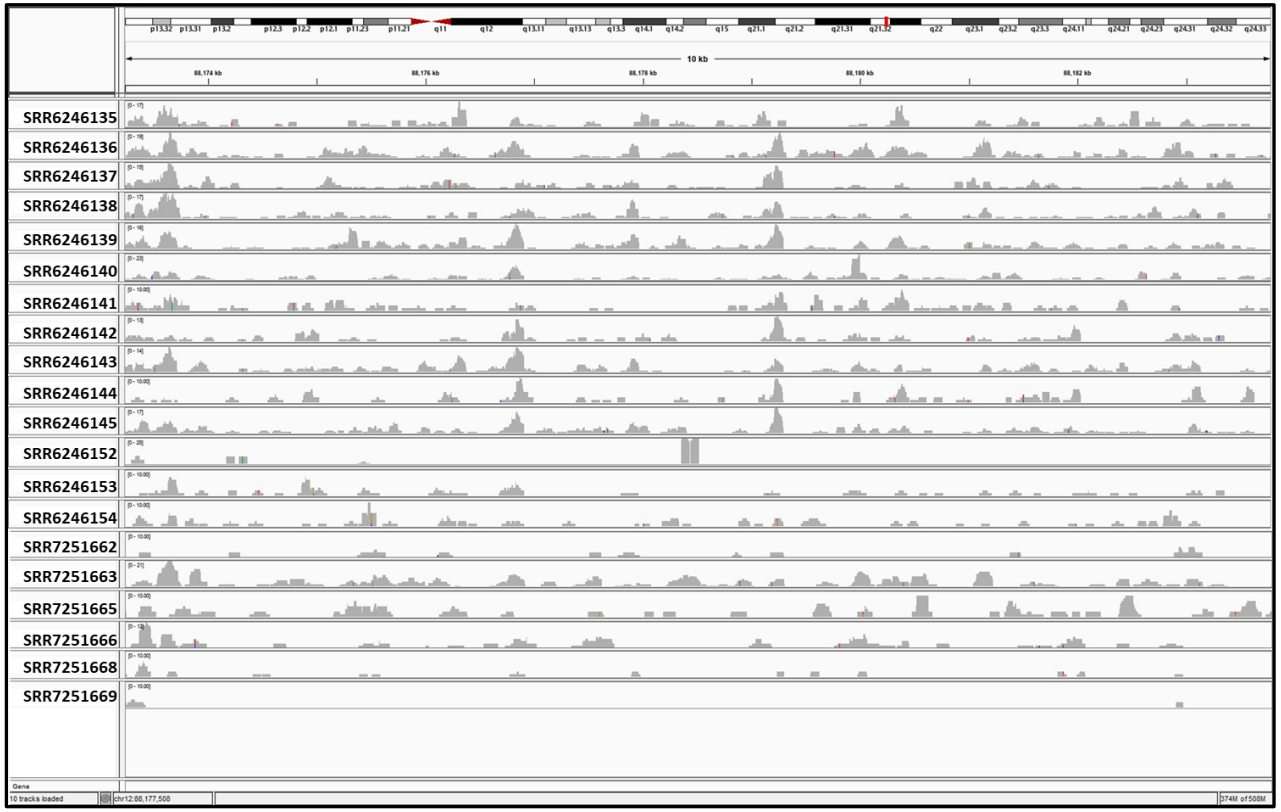


Figure S186. Read Alignment Coverage for ERVMap 3866 (HERV-K22)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 3866 (Chromosome 12, locus q21.32), identified as HERV-K22. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

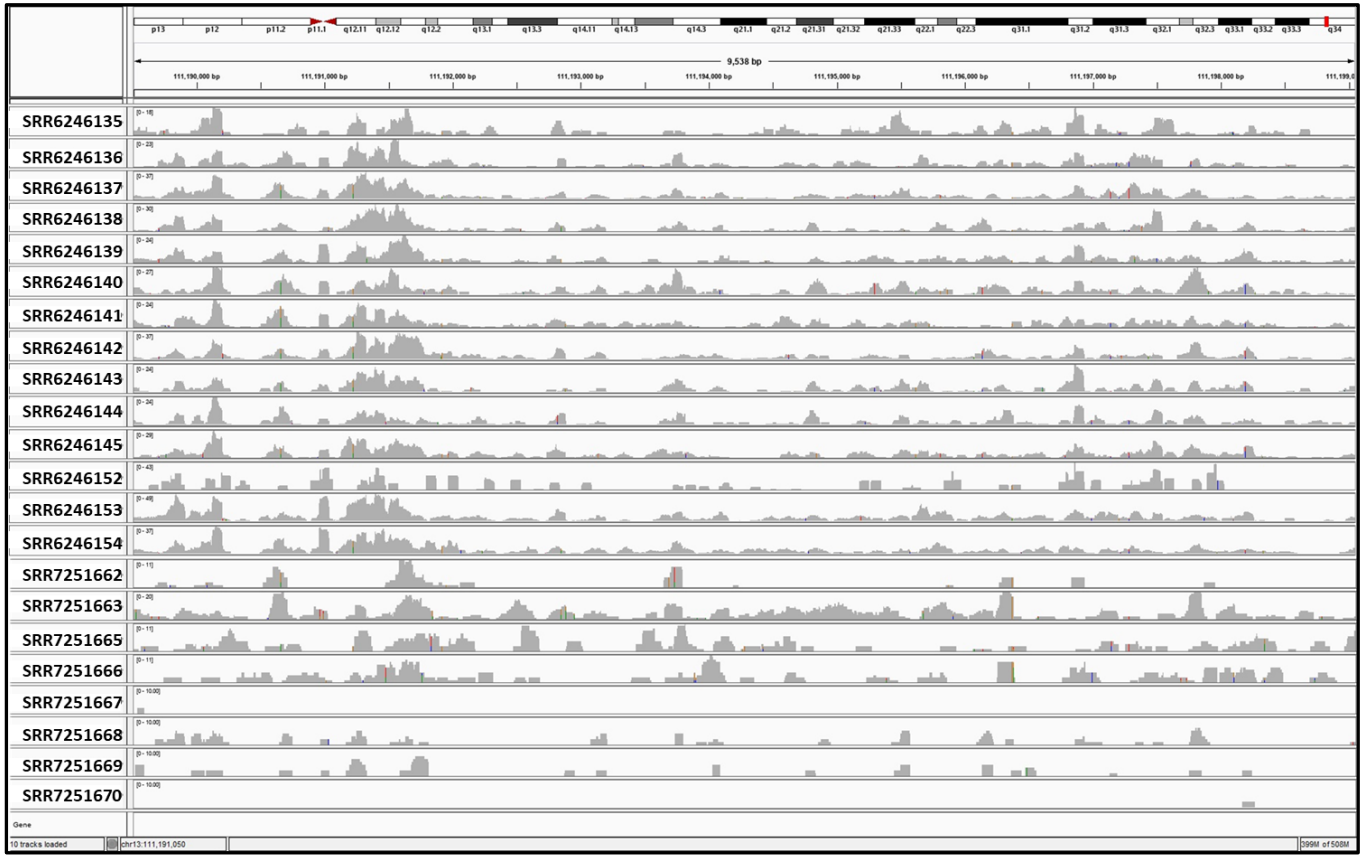


Figure S187. Read Alignment Coverage for ERVMap 4060 (HERV9)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 4060 (Chromosome 13, locus q34), identified as HERV9. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

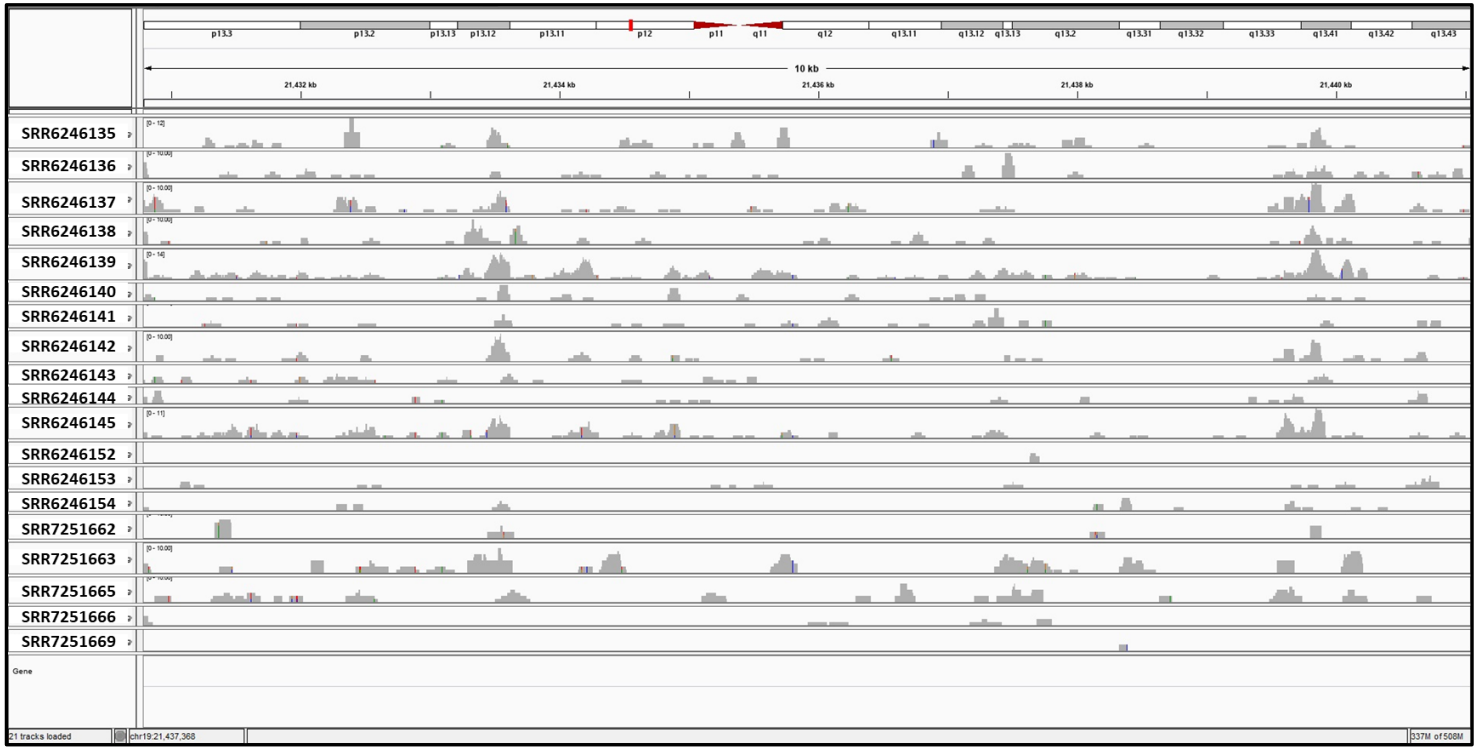


Figure S188. Read Alignment Coverage for ERVMap 4656 (HERV3)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 4656 (Chromosome 19, locus p12), identified as HERV3. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

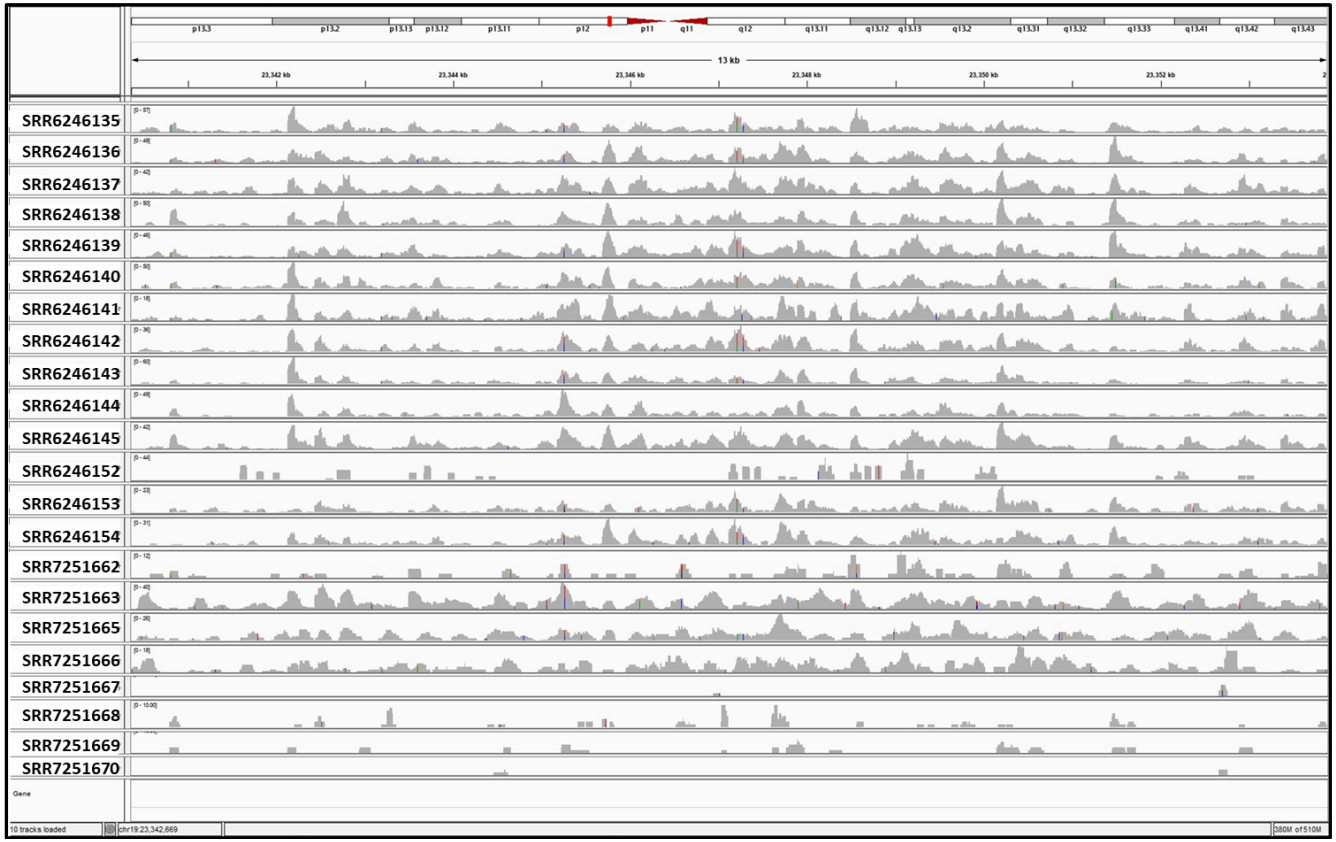


Figure S189. Read Alignment Coverage for ERVMap 4678 (HERV-H)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 4678 (Chromosome 19, locus p12), identified as HERV-H. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

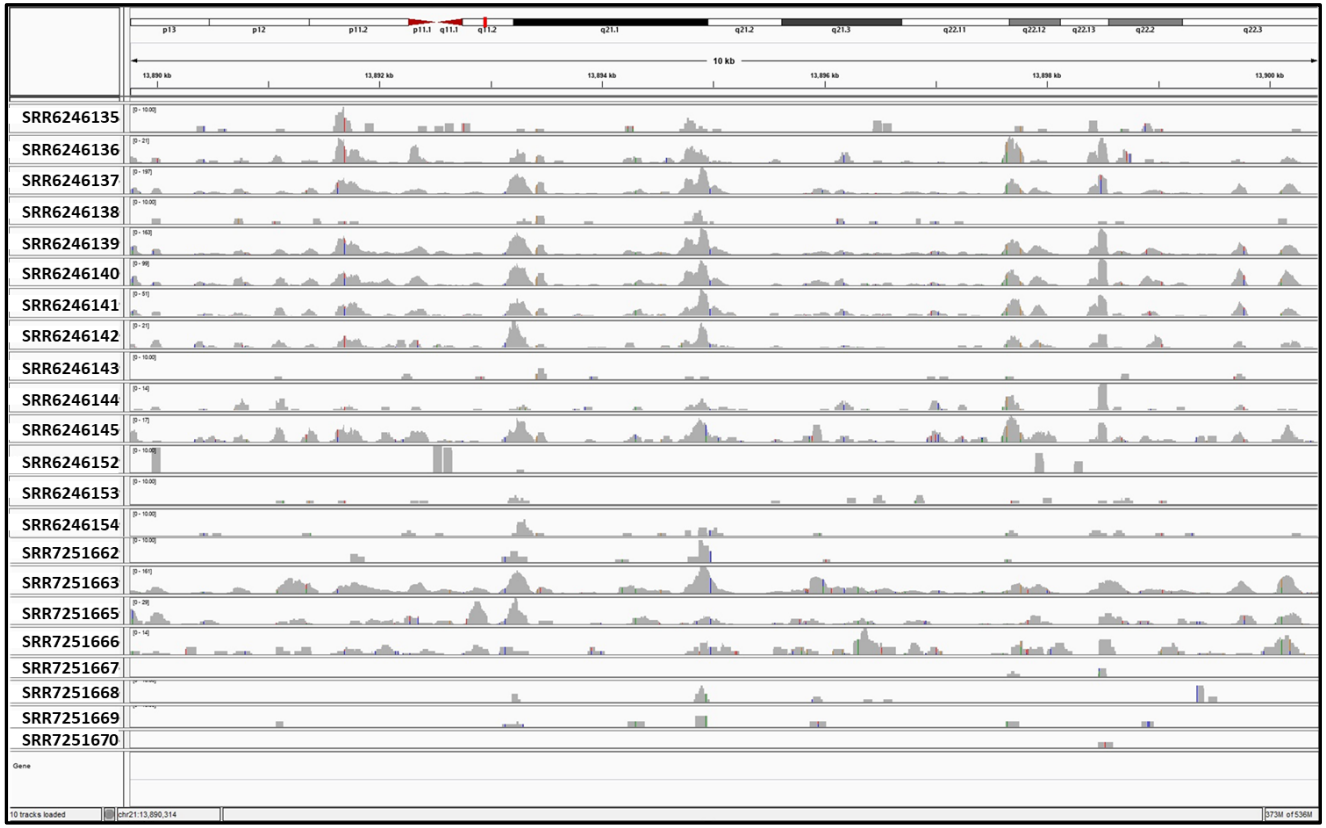


Figure S190. Read Alignment Coverage for ERVMap 4861 (HERV-L)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 4861 (Chromosome 21, locus q11.2), identified as HERV-L. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

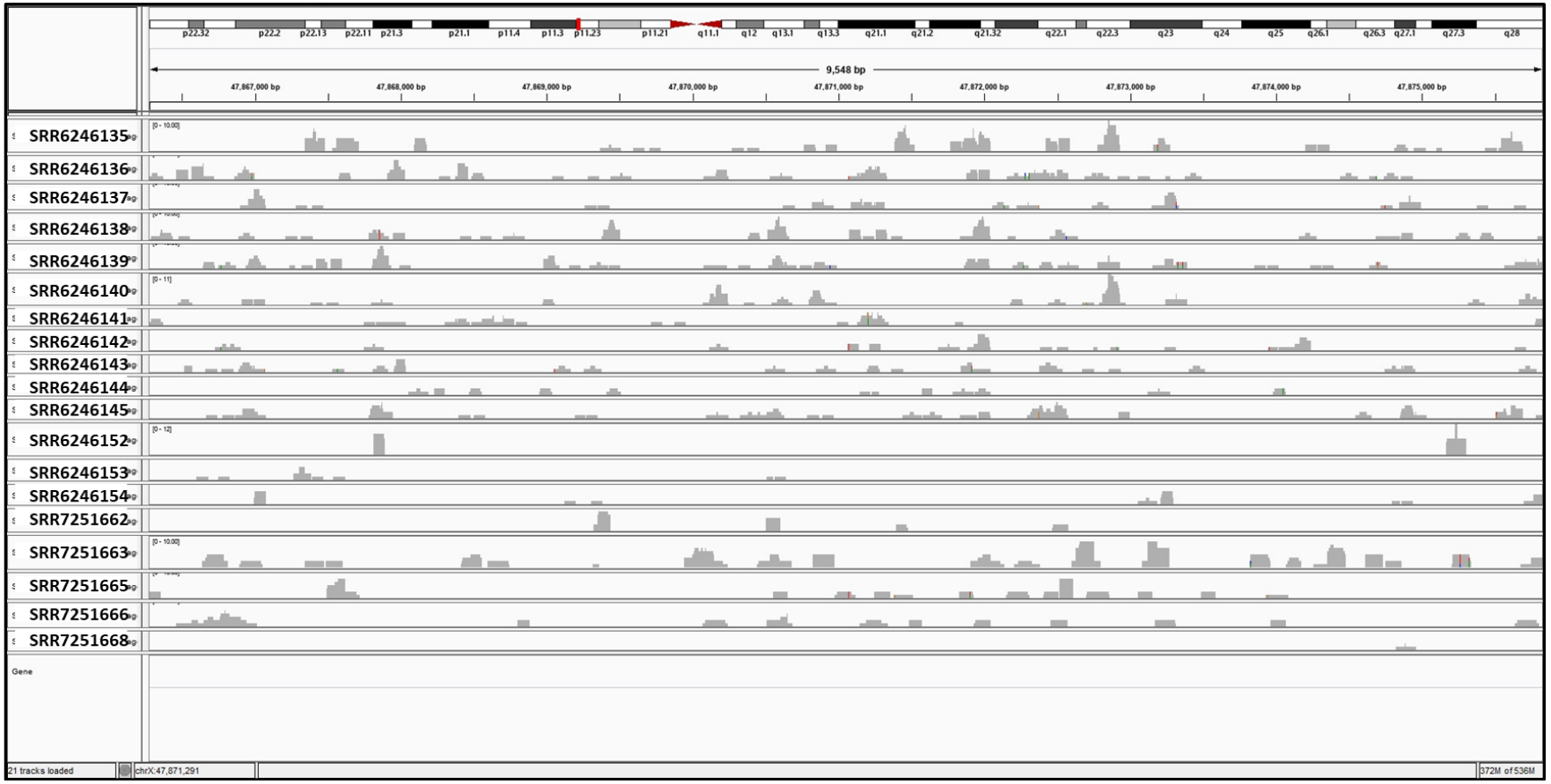


Figure S191. Read Alignment Coverage for ERVMap 5359 (MER89)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 5359 (Chromosome X, locus p11.23), identified as MER89. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

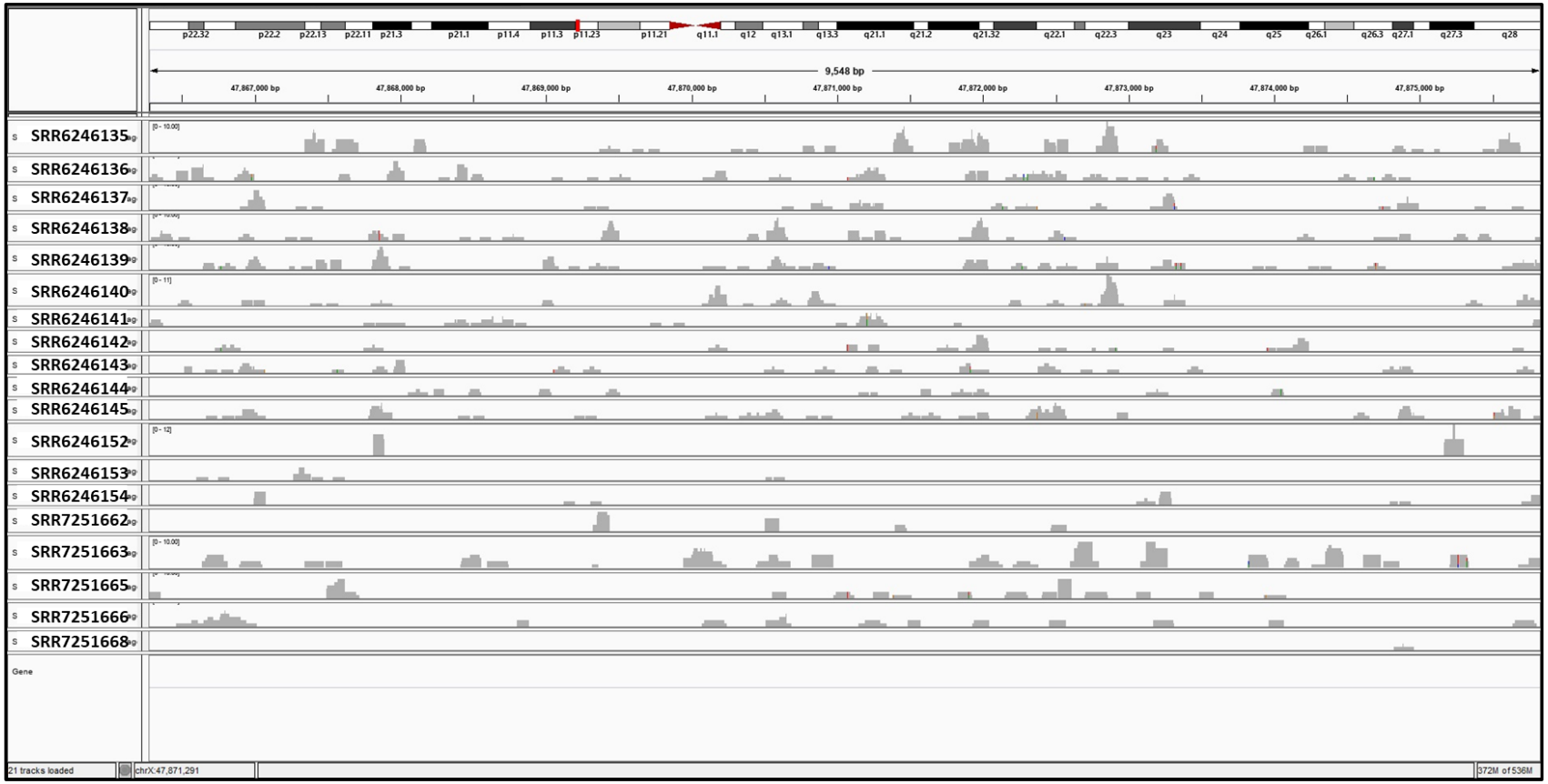


Figure S192. Read Alignment Coverage for ERVMap 5361 (HERV-E)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 5361 (Chromosome X, locus p11.23), identified as HERV-E. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

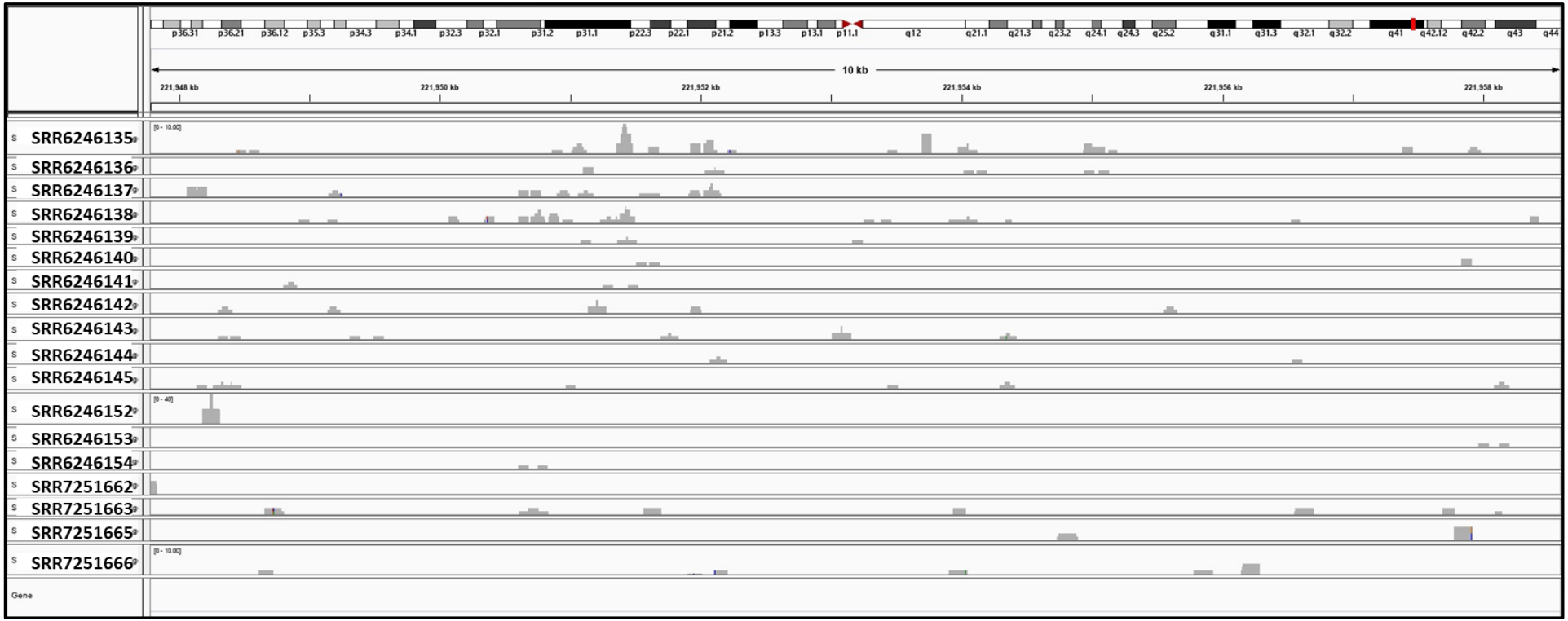


Figure S193. Read Alignment Coverage for ERVMap 6195 (HERV-H)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 6195 (Chromosome 1, locus q41), identified as HERV-H. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

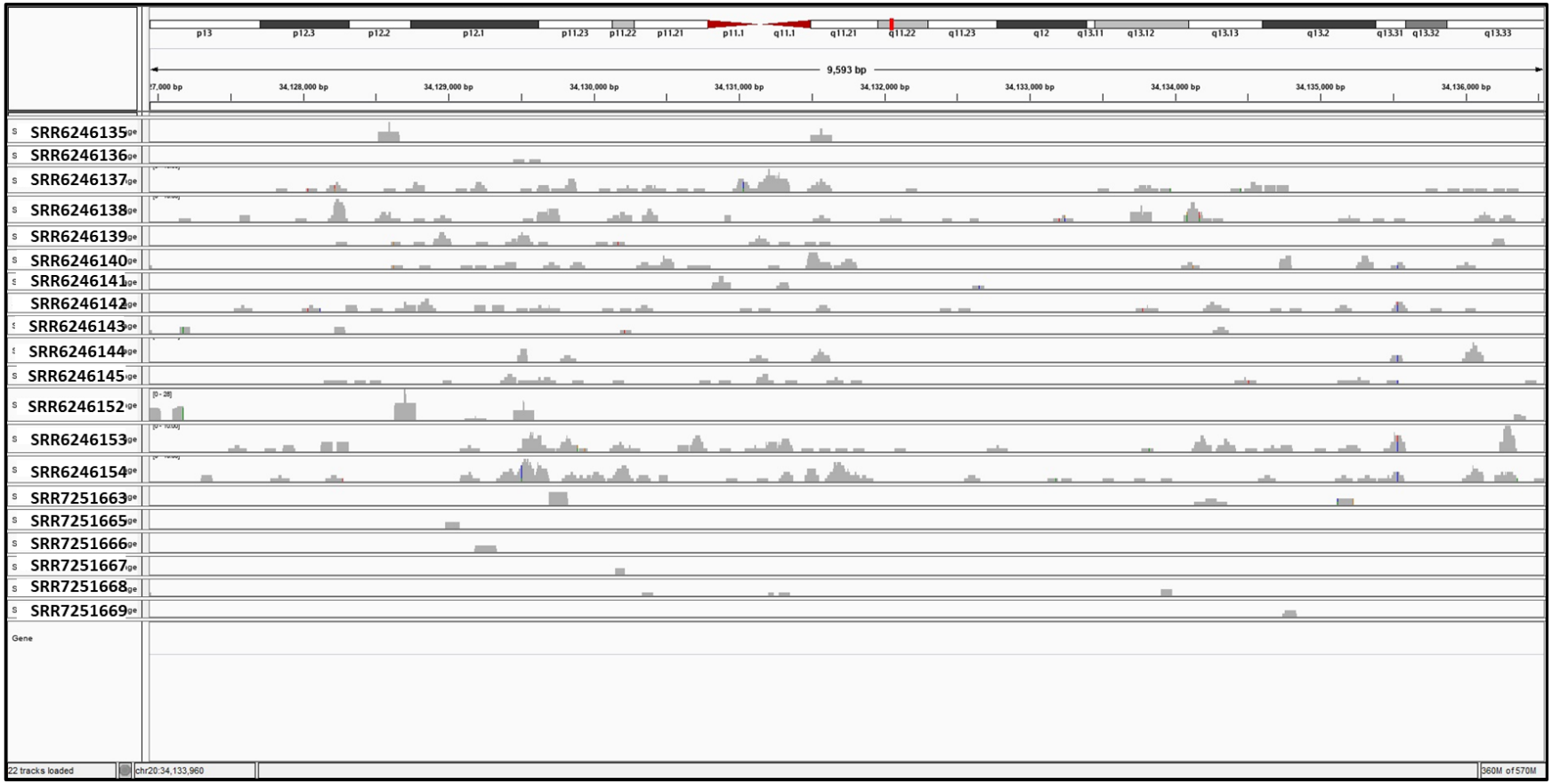


Figure S194. Read Alignment Coverage for ERVMap K-46 (HERV-K)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap K-46 (Chromosome 20, locus 11.22), identified as HERV-K. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus.

Supplementary Table S6. Clinical Information for Post-Mortem Premotor Cortex Brain Tissue Samples

Shown in the table below is relevant clinical information for post-mortem tissue samples that were used in the HERV-K and HERV-W gene expression studies using RT-qPCR. Information received about patient samples includes age at time of death, delay in retrieving post-mortem tissue after death and clinical diagnosis in the case of non-ALS control samples.

Sample ID	Clinical status	Sex	Age (years)	Post-mortem delay (hrs)	Clinical information
A375/12	ALS	M	61	20.5	Sporadic ALS
A073/12	ALS	F	68	29	Sporadic ALS
A342/13	ALS	F	50	56	Sporadic ALS
A254/15	ALS	F	68	51.5	Sporadic ALS
A355/15	ALS	M	69	52.5	Sporadic ALS
A151/10	ALS	F	75	38	Sporadic ALS
A115/08	ALS	M	83	42	Sporadic ALS
A414/12	ALS	F	72	53	Sporadic ALS
A203/11	ALS	F	57	55	Sporadic ALS
A447/13	ALS	F	90	34	Sporadic ALS
A381/11	ALS	F	86	77	Sporadic ALS
A251/09	ALS	M	78	2:30	Sporadic ALS
A205/09	ALS	M	58	35	Sporadic ALS
A162/09	ALS	F	70	27.5	Sporadic ALS
A041/07	ALS	F	87	21.5	Sporadic ALS
A308/15	ALS	M	76	51	Sporadic ALS
A401/08	ALS	F	80	36.5	Sporadic ALS
A348/08	ALS	F	69	64	Sporadic ALS
A218/09	ALS	F	65	3.7	Sporadic ALS
A292/09	Control	F	43	43	pneumonia; severe obesity
A261/12	Control	M	63	23	colon cancer
A132/14	Control	F	66	50	Alzheimer's disease BNE stage II

Supplementary Table S6. (Continued). Clinical Information for Post-Mortem Premotor Cortex Tissue Samples

Shown in the table below is relevant clinical information for postmortem brain tissue samples used in the HERV-K and HERV-W gene expression studies using RT-qPCR.. Information received about patient samples includes age at time of death, delay in retrieving post-mortem tissue after death and clinical diagnosis in the case of non-ALS control samples.

Sample ID	Clinical status	Sex	Age (years)	Postmortem delay (hrs)	Clinical information
A308/09	Control	M	66	52	cancer, renal failure
A346/10	Control	F	84	34	Sepsis; aspiration pneumonia; metastatic breast cancer
A265/08	Control	M	79	47	ruptured anterior myocardial infarction
A012/12	Control	F	51	33	lung cancer
A358/08	Control	F	55	12	COPD, respiratory failure, bronchopneumonia
A033/11	Control	M	82	47	congestive cardiac failure; aortic stenosis
A002/13	Control	M	90	45	Myocardial Infarction
A007/15	Control	F	74	66	1a. Hospital acquired pneumonia 1b. Chronic obstructive pulmonary disease 2. bilateral lung disease
A407/13	Control	F	80	22	S4 Lung cancer
A153/06	Control	F	92	17	Colon cancer
A248/11	Control	F	84	53	Myocardial infarction; diabetes
A177/14	Control	F	67	62	
A158/14	Control	F	73	27	
A273/12	Control	M	67	25	metastatic prostate cancer
A308/14	Control	F	66	78	metastatic colorectal cancer
A319/14	Control	F	90	44	metastasized cancer, no primary found.
A103/17	Control	F	55	71	1a. Multi Organ Failure 1b Severe Sepsis 1c bronchopneumonia 2 Mixed connective tissue disease

Supplementary Table S7. Clinical Information for Additional Non-ALS or ALS Associated, Control Post-Mortem Premotor Cortex Tissue Samples

Shown in the table below is relevant clinical information for postmortem tissue samples used in research work detailed in the methods section below from the Garson *et.al.* (2019) paper. Information received about patient samples includes age at time of death, delay in retrieving post-mortem tissue after death and clinical diagnosis in the case of non-ALS control samples.

Sample ID	Clinical status	Sex	Age (years)	Postmortem delay (hrs)	Clinical information
A066/00	Control	M	61	53.0	Cardiac Arrest
A167/98	Control	M	67	41.0	Myocardial Infarction
A358/99	Control	F	55	95.0	Myocardial Infarction
A136/10	Control	F	89	41	Aortic sclerosis; infection, Hypoxic-type changes and amyloid angiopathy (AD BNE modified Braak score I-II)
A310/09	Control	F	84	35	Alzheimer changes Braak II consistent with normal aging
A133/12	Control	F	88	39	Consistent with ageing; Alzheimer changes Braak 3
A145/02	Control	M	55	24.0	Dementia, Syringomyelia
A092/14	Control	F	83	74	Limbic tauopathy, BNE stage II mild small vessel disease possible control, Moderate amyloid angiopathy, capillary type

Supplementary Table S8. Clinical Information for additional Post-Mortem Premotor Cortex Brain Tissue Samples used in Garson et.al. 2019 and not used in initial RT-qPCR assay validation and HERV-K and HERV-W RT-qPCR assays.

Shown in the table below is relevant clinical information for postmortem brain tissue samples used in RT-qPCR assays to measure relative expression of HERV-H, HERV-K22 and HERV-K3 transcripts. Information received about patient samples includes age at time of death, delay in retrieving post-mortem tissue after death.

Sample ID	Clinical status	Sex	Age (years)	Postmortem delay (hrs)	RIN
A293/17	ALS	F	89	30	4.80
A148/13	ALS	M	75	20	5.00
A285/11	ALS	F	63	25	5.50
A233/14	ALS	M	68	73	5.50
A463/17	ALS	F	63	25	5.50
A484/15	ALS	M	75	98	5.60
A420/15	ALS	M	75	42	5.70
A206/16	ALS	F	49	64	5.90
A314/14	ALS	F	68	57	6.00
A098/09	ALS	M	72	26	6.00
A249/09	ALS	M	68	5.2	6.10
A217/14	ALS	M	78	79	6.20
A368/15	ALS	F	79	35	6.20
A221/12	ALS	M	77	66	6.30
A397/12	ALS	M	75	6.5	6.50
A130/13	ALS	M	74	19	6.50
A211/09	ALS	F	73	70	6.50
A343/10	ALS	M	65	64	6.50
A365/17	ALS	M	64	51	6.50
A028/13	ALS	M	55	4.5	6.70
A177/09	ALS	M	82	12.5	6.70
A112/14	ALS	F	65	68	6.70
A074/17	ALS	F	72	74	6.70
A254/12	ALS	M	54	69	6.80
A001/16	ALS	F	54	67.5	6.80
A273/09	ALS	M	66	59	6.80
A426/15	ALS	M	60	11	6.90
A401/13	ALS	M	68	14	7.20
A244/14	ALS	M	53	69	7.20
A155/14	ALS	F	62	70	7.20
A237/12	ALS	M	65	33	7.30
A211/16	ALS	F	68	64	7.40
A044/10	ALS	M	50	26	7.50
A221/16	ALS	M	66	46	8.00
A202/11	ALS	M	65	16	8.20
A049/03	Control	M	79	24	3.80
A234/17	Control	F	47	27	4.60
A345/12	Control	M	85	55	4.80
A134/00	Control	M	86	6	4.80
A048/09	Control	M	81	18	5.10

Supplementary Table S8. (Continued) Clinical Information for Post-Mortem Premotor Cortex Brain Tissue Samples from Garson et.al. 2019 not used in initial HERV-K and HERV-W RT-qPCR Assays but for HERV-H, HERV-K22 and HERV-K3 RT-qPCR assays.

Sample ID	Clinical status	Sex	Age (years)	Postmortem delay (hrs)	RIN
A185/04	Control	M	80	48.25	5.10
A200/13	Control	F	81	30	5.60
A382/12	Control	F	85	45	5.80
A006/15	Control	M	89	84	5.90
A006/16	Control	F	68	58.5	6.00
A274/07	Control	M	78	30.5	6.00
A105/14	Control	F	77	21	6.20
A209/13	Control	M	80	55	6.40
A359/08	Control	F	80	3	6.40
A042/01	Control	F	52	44	6.50
A130/09	Control	M	54	30	6.60
A309/99	Control	F	58	21	6.60
A142/16	Control	M	83	48	6.90
A388/12	Control	M	65	26	7.00
A216/15	Control	M	86	10.5	7.40

Supplementary Table S9. Clinical Information for Post-Mortem Primary Motor Cortex Brain Tissue Samples Used in RNA Sequencing Analysis

The samples listed in the table below were supplied by the MRC neurodegenerative disease brain bank. These samples were sent by our collaborators at KCL to Source Bioscience (UK) for RNA extraction, quantification and sequencing using their Illumina Hi-Seq Next Generation Sequencing (NGS) platform for RNA seq analysis using ERVMap RNA Seq pipeline and were selected to match patient samples used in our initial RT-qPCR validation and HERV-K and HERV-W RT-qPCR experiments.

Sample ID	Clinical Status	Sex	Age (Years)	PMD (Hours)	Source Bioscience RIN
A002_13	Control	M	90	45	7.40
A007_15	Control	F	74	66	6.30
A033_11	Control	M	82	47	6.80
A041_07	ALS	F	87	21.5	6.60
A066_00	Control	M	61	53.0	5.80
A073_12	ALS	F	68	29	8.00
A103_17	Control	F	55	71	5.00
A132_14	Control	F	66	50	6.90
A133_12	Control	F	88	39	6.30
A136_10	Control	F	89	41	6.40
A151_10	ALS	F	75	38	7.20
A158_14	Control	F	70	27.5	4.80
A162_09	ALS	F	70	27.5	6.80
A167_98	Control	M	67	41.0	4.60
A205_09	ALS	M	58	35	7.90
A218_09	ALS	F	65	3.7	5.70
A248_11	Control	F	84	53	7.40
A251_09	ALS	M	78	2.5	6.00
A292_09	Control	F	43	43	7.80
A348_08	ALS	F	69	64	4.80
A358_08	Control	F	55	12	6.50
A358_99	Control	F	55	95.0	6.30
A375_12	ALS	M	61	20.5	6.70
A401_08	ALS	F	80	36.5	7.60
A447_13	ALS	F	90	34	5.20

Supplementary Table S10. Clinical Information for the Publicly Sourced RNA-Seq Peripheral Blood Mononuclear Cell (PBMC) Dataset.

The following table shows clinical information for RNA-Seq samples obtained from the publicly available PBMC dataset published in Zucca *et al.*, 2019 from an Italian cohort. All RNA RIN values were >8.0.

Assay ID	Disease Status	Sex	Age at Time of Sampling
SRR6246135	Sporadic ALS	M	66
SRR6246136	Sporadic ALS	M	63
SRR6246137	Sporadic ALS	F	61
SRR6246138	Sporadic ALS	F	70
SRR6246139	Sporadic ALS	M	56
SRR6246140	Sporadic ALS	F	55
SRR6246141	Sporadic ALS	F	56
SRR6246142	Sporadic ALS	F	83
SRR6246143	Sporadic ALS	M	66
SRR6246144	Sporadic ALS	F	58
SRR6246145	Sporadic ALS	M	68
SRR6246152	Non-ALS Control	M	48
SRR6246153	Non-ALS Control	M	60
SRR6246154	Non-ALS Control	M	68
SRR7251662	Sporadic ALS	M	86
SRR7251663	Sporadic ALS	F	68
SRR7251665	Sporadic ALS	M	71
SRR7251666	Sporadic ALS	F	69
SRR7251667	Non-ALS Control	M	38
SRR7251668	Non-ALS Control	F	36
SRR7251669	Non-ALS Control	F	40
SRR7251670	Non-ALS Control	F	40

Supplementary Table S11. Clinical Information for the Publicly Sourced RNA-Seq cerebellum and frontal cortex sample dataset (Prudencio *et.al.* 2017).

Prudencio et.al. (2017) Sample №	SRA Accession	Tissue	Status	Sex	Age (Years)	PMD (hours)
11	SRR1927035	Cerebellum	ALS	Female	61.4	10
	SRR1927036	Frontal Cortex				
12	SRR1927037	Cerebellum	ALS	Male	50.1	9
	SRR1927038	Frontal Cortex				
13	SRR1927039	Cerebellum	ALS	Male	53.8	6
	SRR1927040	Frontal Cortex				
14	SRR1927041	Cerebellum	ALS	Male	66.6	13
	SRR1927042	Frontal Cortex				
70	SRR1927043	Cerebellum	ALS	Female	53.2	16
	SRR1927044	Frontal Cortex				
71	SRR1927045	Cerebellum	ALS	Male	47.5	5
	SRR1927046	Frontal Cortex				
72	SRR1927047	Cerebellum	ALS	Female	60.9	16
	SRR1927048	Frontal Cortex				
73	SRR1927049	Cerebellum	ALS	Female	65.2	7
	SRR1927050	Frontal Cortex				
74	SRR1927051	Cerebellum	ALS	Female	69.2	3
	SRR1927052	Frontal Cortex				
75	SRR1927053	Cerebellum	ALS	Female	70.4	12
	SRR1927054	Frontal Cortex				
24	SRR1927055	Cerebellum	Control	Female	64.9	30
	SRR1927056	Frontal Cortex				
86	SRR1927057	Cerebellum	Control	Male	76.6	18
	SRR1927058	Frontal Cortex				
90	SRR1927059	Cerebellum	Control	Female	82.1	2
	SRR1927060	Frontal Cortex				
91	SRR1927061	Cerebellum	Control	Female	72.9	18
	SRR1927062	Frontal Cortex				
93	SRR1927063	Cerebellum	Control	Male	80.8	15
	SRR1927064	Frontal Cortex				
94	SRR1927065	Cerebellum	Control	Male	75.1	17
	SRR1927066	Frontal Cortex				
95	SRR1927067	Cerebellum	Control	Male	78.2	13
	SRR1927068	Frontal Cortex				
97	SRR1927069	Cerebellum	Control	Male	78.3	8
	SRR1927070	Frontal Cortex				
100	SRR1927071	Frontal Cortex	Control	Male	55	13

Supplementary Table S11. (Continued) Sample Metadata for Cerebellum and Frontal Cortex Samples

Prudencio et.al. (2017) Sample Nº	SRA Accession	Tissue	Status	Sex	Age (Years)	PMD (Hours)
1	SRR1927020	Cerebellum	ALS	Female	64.3	15
	SRR1927019	Frontal Cortex				
2	SRR1927022	Cerebellum	ALS	Male	58.3	6
	SRR1927021	Frontal Cortex				
6	SRR1927028	Cerebellum	ALS	Male	58.7	18
	SRR1927027	Frontal Cortex				
7	SRR1927034	Cerebellum	ALS	Male	53.6	7
	SRR1927033	Frontal Cortex				
34	SRR1927024	Cerebellum	ALS	Female	50.2	8
	SRR1927023	Frontal Cortex				
57	SRR1927026	Cerebellum	ALS	Female	51	18
	SRR1927025	Frontal Cortex				
62	SRR1927030	Cerebellum	ALS	Female	42.6	3
	SRR1927029	Frontal Cortex				
63	SRR1927030	Cerebellum	ALS	Female	49.5	6
	SRR1927031	Frontal Cortex				

Supplementary Table S12. Clinical Information for the Publicly Sourced RNA-Seq medial motor cortex tissue sample dataset obtained from New York Genomic Centre in Partnership with Target ALS

Raw RNA-Seq files and metadata were requested from Target ALS and downloaded from NYGC web portal.

Sample ID	Status	Sex	Age	PMD	RIN
STAR-00257	ALS	Male	59	18.8	5.4
STAR-00461	ALS	Female	61	7	7.9
STAR-00467	Control	Female	57	32	3.8
STAR-00477	Control	Male	59	18	4.3
STAR-00482	ALS	Male	51	10	7
STAR-00488	ALS	Male	59	13	7.3
STAR-00627	ALS	Male	72	9.5	4.7
STAR-00646	Control	Female	71	5	3.4
STAR-00677	ALS	Male	69	19	4.2
STAR-00732	ALS	Male	69	19	6.1
STAR-01180	Control	Male	51	23	6
STAR-01279	ALS	Male	60	17	7.5
STAR-01312	Control	Female	74	3	6
STAR-01320	ALS	Female	51	7	6.3
STAR-01380	Control	Male	68	10	3.8
STAR-01386	ALS	Female	54	8	6.1
STAR-01394	ALS	Male	48	7	6.2
STAR-01398	ALS	Male	72	4.5	5.6
STAR-01414	ALS	Male	64	9	5.9
STAR-01424	ALS	Female	66	26	4.2
STAR-01452	ALS	Female	69	6	6
STAR-01487	ALS	Female	32	5.16	3.8
STAR-01493	ALS	Male	54	6.25	5.6
STAR-01499	ALS	Female	65	14.4	6.5
STAR-01505	ALS	Male	64	8	5.2
STAR-01511	ALS	Male	48	8.25	6.1
STAR-01517	ALS	Male	53	28	6.8
STAR-01527	ALS	Female	45	5.25	7.5
STAR-01531	ALS	Male	67	3	7.4
STAR-01535	ALS	Male	70	5.25	6
STAR-01539	ALS	Male	59	13.85	6.3
STAR-01562	ALS	Female	58	6.8	7
STAR-01568	ALS	Male	67	6	6.7
STAR-01574	ALS	Male	60	13	5.5
STAR-01580	ALS	Female	55	21.73	6.5
STAR-01608	ALS	Male	68	15	6.3
STAR-01629	ALS	Male	67	5.25	7.2
STAR-01636	ALS	Female	69	5.25	6.8
STAR-01983	ALS	Male	55	15	5
STAR-02157	ALS	Male	40	6	6.8

Supplementary Table S13. Clinical Information for the Publicly Sourced RNA-Seq lateral motor cortex brain tissue samples obtained from New York Genomic Centre

Sample ID	Clinical Status	Sex	Age (Years)	PMD (Hours)	RIN
STAR-00462	ALS	Female	61	7	7
STAR-00478	Control	Male	59	18	3.9
STAR-00483	ALS	Male	51	10	6.4
STAR-00489	ALS	Male	59	13	6.6
STAR-00628	ALS	Male	72	9.5	5.7
STAR-00647	Control	Female	71	5	3.6
STAR-00678	ALS	Male	69	19	4.2
STAR-00733	ALS	Male	69	19	5.9
STAR-01181	Control	Male	51	23	6.7
STAR-01280	ALS	Male	60	17	6.9
STAR-01313	Control	Female	74	3	6.3
STAR-01321	ALS	Female	51	7	7.8
STAR-01381	Control	Male	68	10	5.8
STAR-01384	ALS	Female	59	10	6.1
STAR-01387	ALS	Female	54	8	7.2
STAR-01399	ALS	Male	72	4.5	5.6
STAR-01403	ALS	Female	66	12	5.8
STAR-01407	ALS	Female	56	4	4.4
STAR-01415	ALS	Male	64	9	5.9
STAR-01425	ALS	Female	66	26	5.6
STAR-01445	ALS	Male	44	8.5	7.6
STAR-01453	ALS	Female	69	6	6.3
STAR-01461	ALS	Female	61	5	5
STAR-01488	ALS	Female	32	5.16	3.9
STAR-01494	ALS	Male	54	6.25	6.7
STAR-01500	ALS	Female	65	14.4	5.6
STAR-01506	ALS	Male	64	8	4.4
STAR-01512	ALS	Male	48	8.25	5.5
STAR-01518	ALS	Male	53	28	6.5
STAR-01524	ALS	Male	67	5.25	7.1
STAR-01528	ALS	Female	45	5.25	6.5
STAR-01532	ALS	Male	67	3	7.4
STAR-01536	ALS	Male	70	5.25	5.5
STAR-01540	ALS	Male	59	13.85	6.1
STAR-01563	ALS	Female	58	6.8	6.7
STAR-01569	ALS	Male	67	6	6.8
STAR-01575	ALS	Male	60	13	6.2
STAR-01581	ALS	Female	55	21.73	5
STAR-01587	ALS	Male	78	5.9	6.2
STAR-01609	ALS	Male	68	15	4.7
STAR-01628	ALS	Male	67	5.25	6.6

Supplementary Table S13 (Continued) Sample Metadata for Lateral Motor Cortex Tissue Samples Obtained from New York Genomic Centre

Sample ID	Clinical Status	Sex	Age (Years)	PMD (Hours)	RIN
STAR-01635	ALS	Female	69	5.25	7.4
STAR-01646	ALS	Female	68	8	7
STAR-01978	Control	Female	63	27	6.8
STAR-01984	ALS	Male	55	15	7

Supplementary Table S14. Quantification of Total RNA Extracted from n=20 ALS and n=20 Non-ALS Premotor Cortex brain tissue Samples obtained at post-mortem.

In the table below information is given on RIN values obtained from Agilent Bioanalyser 2100 along with RNA yield as measured using the Qubit BR (Broad Range) assay. Qubit means were derived from duplicate/triplicate values that were within 40ng/µl of each other, and only those values used for the mean quantification of RNA yield are given in the table.

Sample ID	Clinical Status	RIN Value	QuBit Mean RNA concentration ng/µl
A375/12	ALS	6.6	560
A073/12	ALS	7.2	530
A342/13	ALS	7.7	421
A254/15	ALS	6.2	450
A355/15	ALS	5.4	405
A151/10	ALS	7.8	616
A115/08	ALS	5.9	824
A414/12	ALS	6.1	637
A203/11	ALS	5.8	420
A447/13	ALS	5	608
A381/11	ALS	6.6	328
A251/09	ALS	6.2	547
A205/09	ALS	7.2	432
A162/09	ALS	6.8	607
A041/07	ALS	6.5	771
A308/15	ALS	7	765
A401/08	ALS	7	529
A348/08	ALS	4.9	631
A218/09	ALS	6.7	364
A331/09	ALS	N/A	Too Low to Measure
A292/09	Control	6.9	593
A261/12	Control	5.8	514
A132/14	Control	5.8	568
A308/09	Control	4.9	131
A346/10	Control	5.9	841
A265/08	Control	6.6	162
A012/12	Control	4.9	576
A358/08	Control	5.4	536
A033/11	Control	6	609
A002/13	Control	6.6	571
A007/15	Control	5.8	407
A407/13	Control	6.6	375
A153/06	Control	6.6	307
A248/11	Control	6.8	446
A177/14	Control	6.7	492
A158/14	Control	4.1	646
A273/12	Control	5.6	603
A308/14	Control	4.4	290
A319/14	Control	4.1	389
A103/17	Control	4.3	450

Supplementary Table S15. Quantification of Total RNA Extracted from n=10 ALS and n=10 Non-ALS Post-Mortem Primary Motor Cortex Brain Tissue Samples.

In the table below information is given on RIN values obtained from Agilent Bioanalyser 2100 along with RNA yield as measured using the Qubit BR (Broad Range) assay. Qubit means were derived from duplicate/triplicate values that were within 40ng/µl of each other, and only those values used for the mean quantification of RNA yield are given in the table.

Sample ID	Status	Sex	RIN Value	QuBit Mean RNA concentration ng/µl
A098/09	ALS	Male	6.2	499
A358/08	Control	Female	5.4	399
A046/00	ALS	Male	6.5	415
A185/04	Control	Male	3.4	490
A130/09	Control	Male	6.7	429
A213/12	Control	Male	6.4	373
A257/17	ALS	Female	5.2	530
A140/94	ALS	Male	7.0	654
A042/01	Control	Female	6.2	542
A103/17	Control	Female	5.0	464
A388/12	Control	Male	7.4	207
A141/99	ALS	Male	4.9	552
A217/93	ALS	Male	7.4	199.5
A309/99	Control	Female	6.8	393
A261/12	Control	Male	4.9	445
A358/99	Control	Female	6.7	673
A081/91	ALS	Female	8.1	536
A134/02	ALS	Male	7.2	611
A251/09	ALS	Male	6.7	451
A083/01	ALS	Male	7.0	518

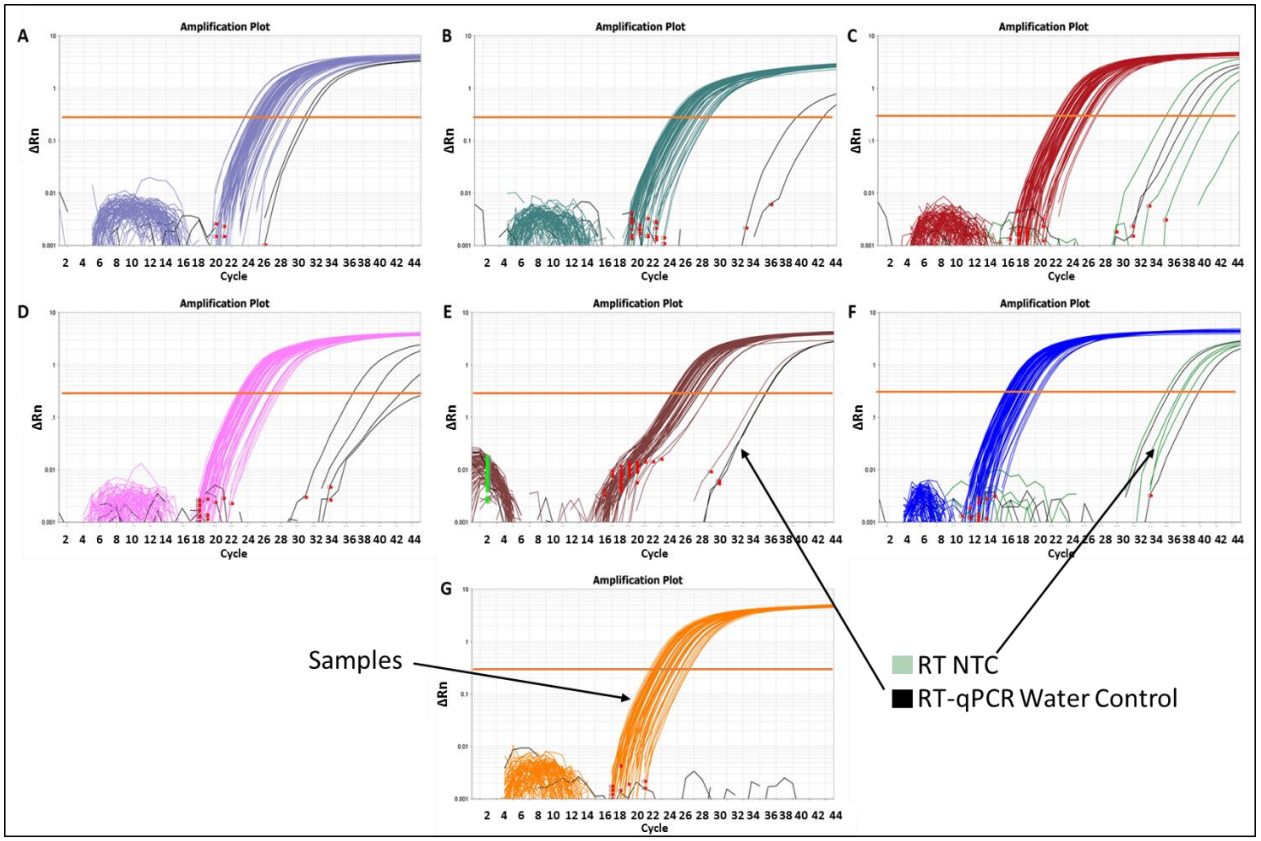


Figure S195. Amplification Plots generated by RT-qPCR following cDNA Amplification of HERV-K and HERV-W transcripts present in n=19 ALS and n=20 non-ALS brain tissue samples.

The figure above shows amplification plots for A) HERV-K *gag*, B) HERV-K *pol*, C) HERV-K *env*, D) HERV-K *RT*, E) HERV-W *env*, F) GAPDH and G) XPNPEP1 gene targets. The black lines in each image represent water control reactions for the RT-qPCR assay with the green lines in C & F showing the amplification curve for no Reverse Transcription RT-qPCR products. The Orange line on each image represents the baseline for measuring the Ct value for the samples in each gene target (ΔRn = 0.21).

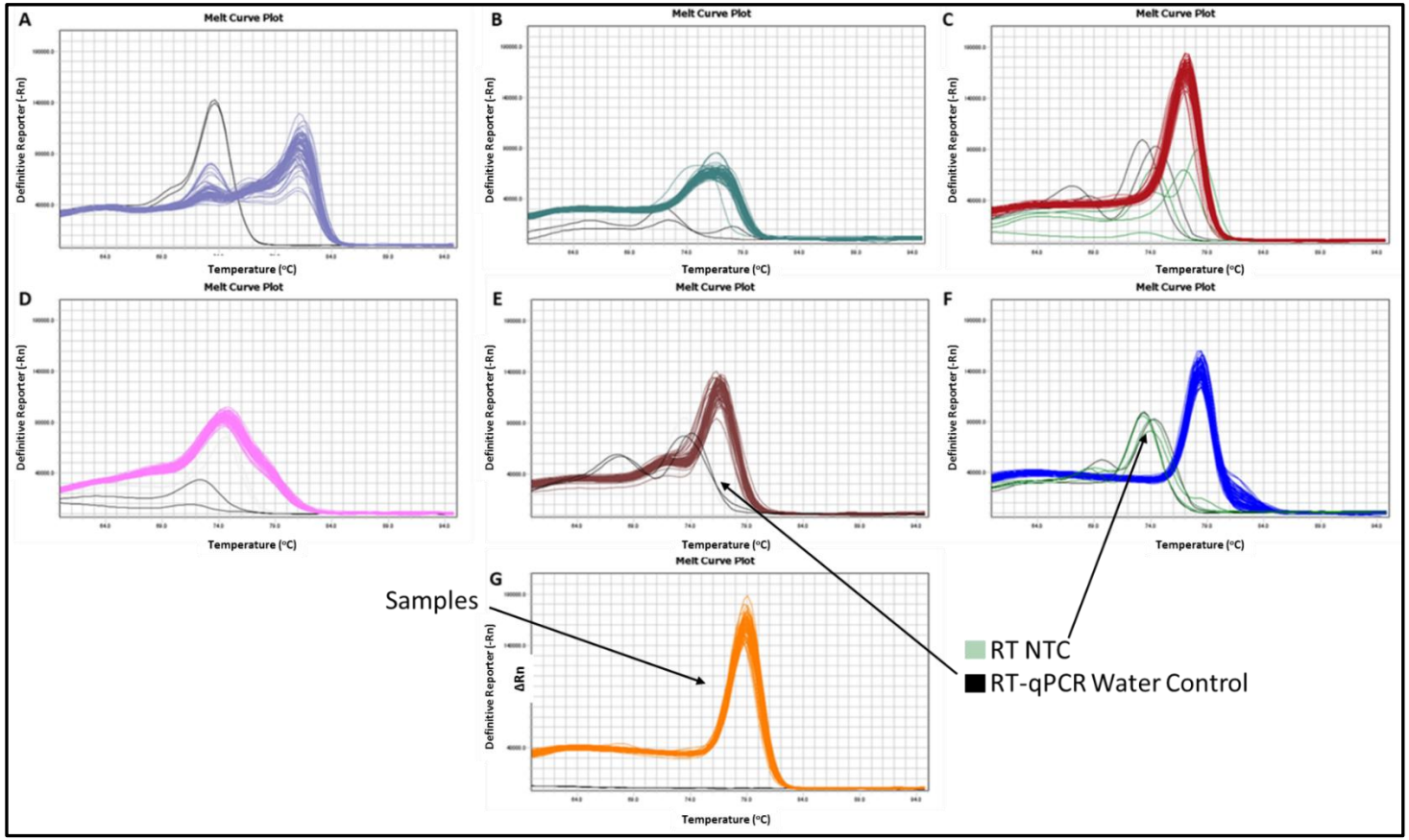


Figure S196. Melt Curve Plots generated by RT-qPCR following cDNA Amplification of HERV-K and HERV-W transcripts present in n=19 ALS and n=20 non-ALS brain tissue samples

The figure above shows melt curve plots for A) HERV-K *gag*, B) HERV-K *pol*, C) HERV-K *env*, D) HERV-K *RT*, E) HERV-W *env*, F) GAPDH and G) XPNPEP1 gene targets. The black lines in each image represent water control reactions for the RT-qPCR plates with the green lines in C & F showing the amplification curves for No Reverse Transcription controls.

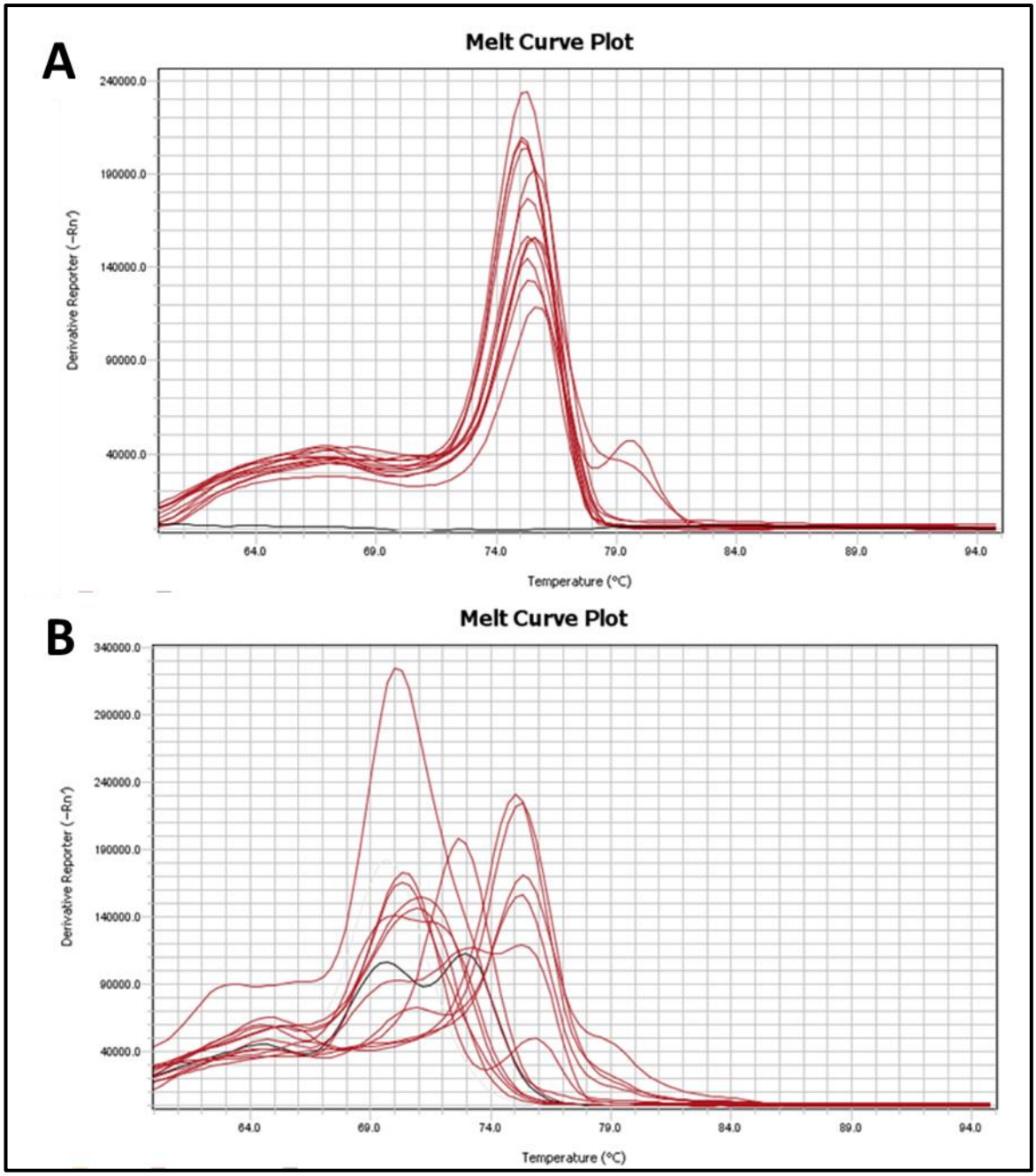


Figure S197. cDNA Melt Curve Plots for HERV-K3 *env* Primer Targets.
The figure above shows Melt Curve plots for HERV-K3 *env* primer targets. These have been divided into primers utilising the ALS samples (A) and non-ALS Control (B). Black lines in the reference gene selection plot indicate NTC reactions for this gene target

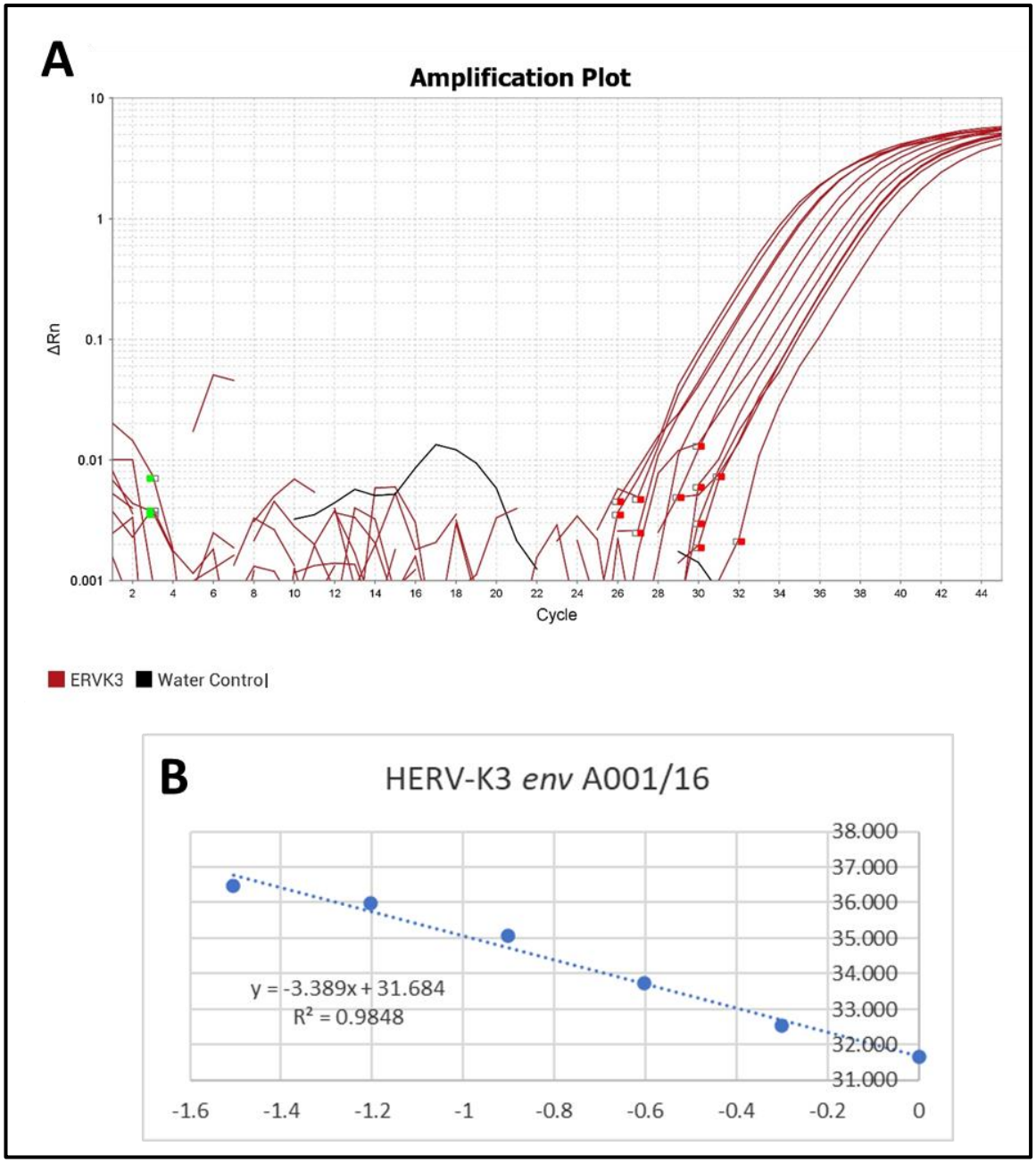


Figure S198. Amplification Plot and primer efficiency graph for HERV-K3 *env* Primer Target.

The figure above displays the amplification plot and standard curve graph for ALS Patient Sample A001/16 using the HERV-K3 *env* primer set. The axes for the standard curve graph display Ct values on the y-axis plotted against log transformed dilution factors performed on the cDNA on the x-axis.

Supplementary Table S16 Summary of Amplification Efficiency Data for HERV-K3 *env* Tested on ALS Patient Sample A001/16.

The table below shows primer efficiency data obtained from Standard curves generated from cDNA amplification efficiency graphs shown in Figure 5.2. Efficiency Percentages were generated from the equation E = 10(-1/slope)*100.

Primer Target (Sample ID)	Slope	R^2	Efficiency
ERVK3 ALS (A001/16)	-3.389	0.9848	97.27%

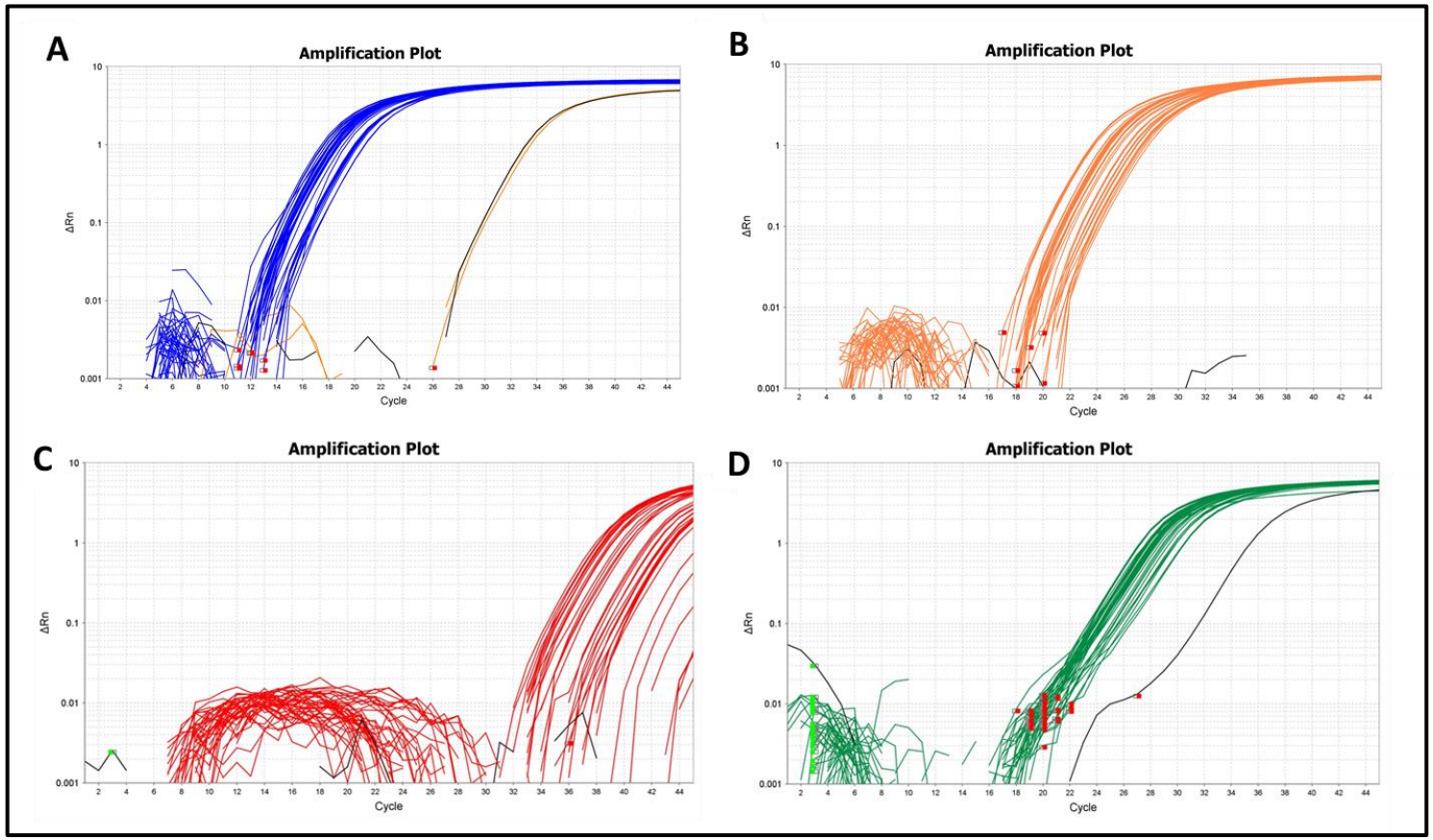


Figure S199. Amplification Plots generated by RT-qPCR following cDNA Amplification of GAPDH, XPNPEP1, HERV-K3 *env* and HERV-W *env* transcripts present in n=10 ALS and n=10 non-ALS Primary Motor Cortex Tissue Samples.

The figure above shows amplification plots for A) GAPDH, B) XPNPEP1, C) HERV-K3 *env* and D)HERV-W *env* gene targets. The black lines in each image represent water control reactions for the RT-qPCR assay with the additional lines clustered around the water control for GAPDH representing the no template controls from the Reverse transcriptase reaction.

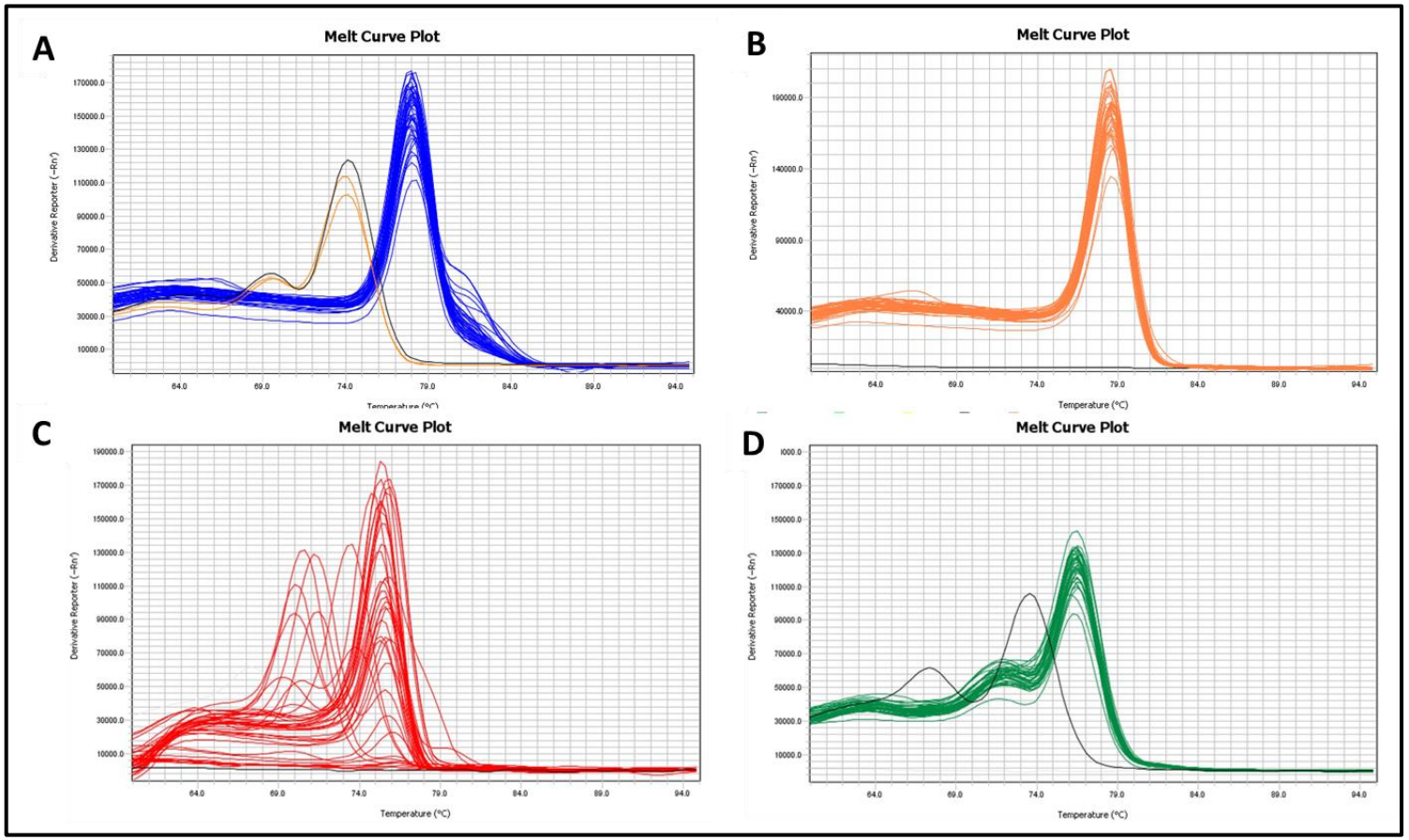


Figure S200. Melt Curve Plots generated by RT-qPCR following cDNA Amplification of GAPDH, XPNPEP1, HERV-K3 *env* and HERV-W *env* transcripts present in n=10 ALS and n=10 non-ALS Primary Motor Cortex Tissue Samples.

The figure above shows melt curve plots for A) GAPDH, B) XPNPEP1, C) HERV-K3 *env* and D)HERV-W *env* gene targets. The black lines in each image represent water control reactions for the RT-qPCR assay with the additional lines clustered around the water control for GAPDH representing the non template controls from the Reverse transcriptase reaction.

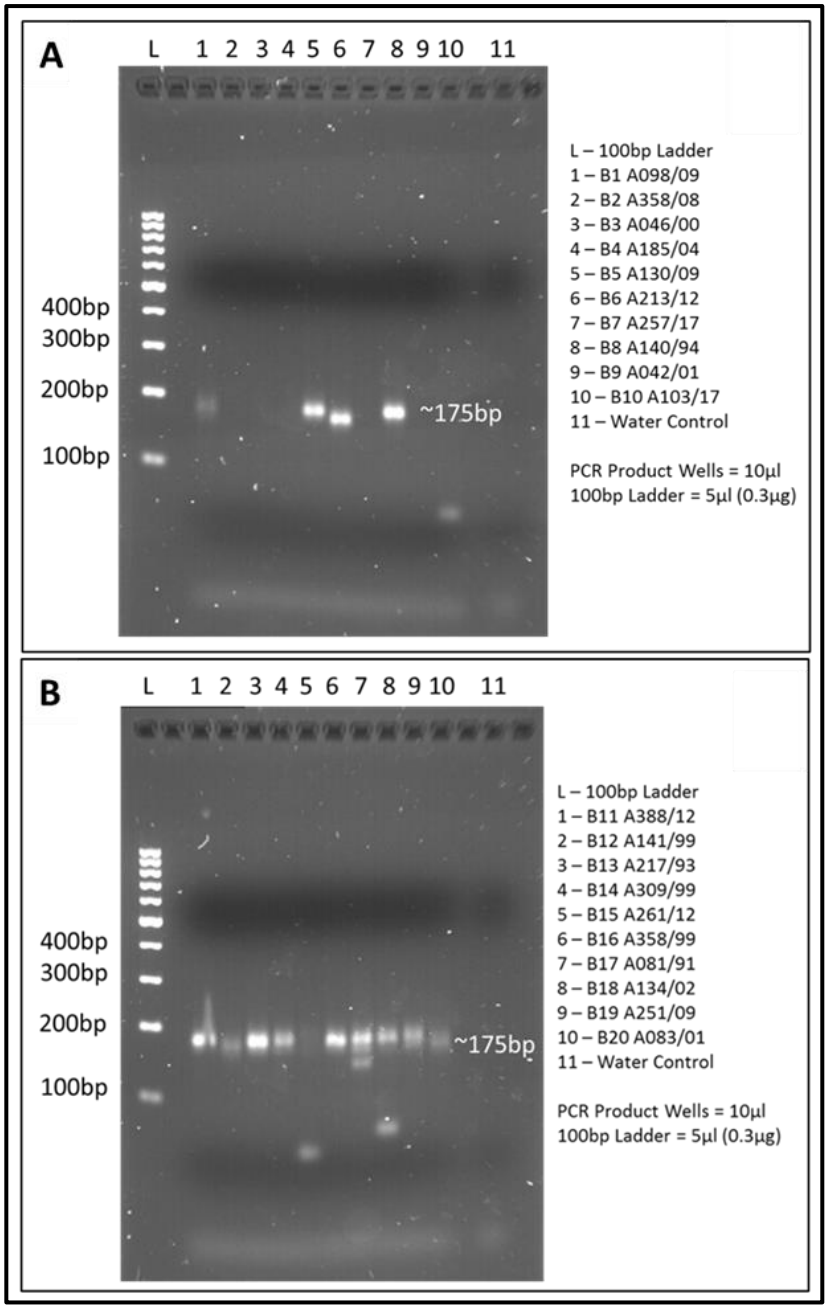


Figure S201. Agarose Gel Electrophoresis Analysis of HERV-K3 *env* Amplicons Produced by RT-qPCR

The figure above shows the gel electrophoresis results for HERV-K3 *env* cDNA amplification by RT-qPCR. The gel in the image above was made to a 2% concentration in TBE buffer with the 100bp ladder Generuler (ThermoFisher Scientific, SM0241).

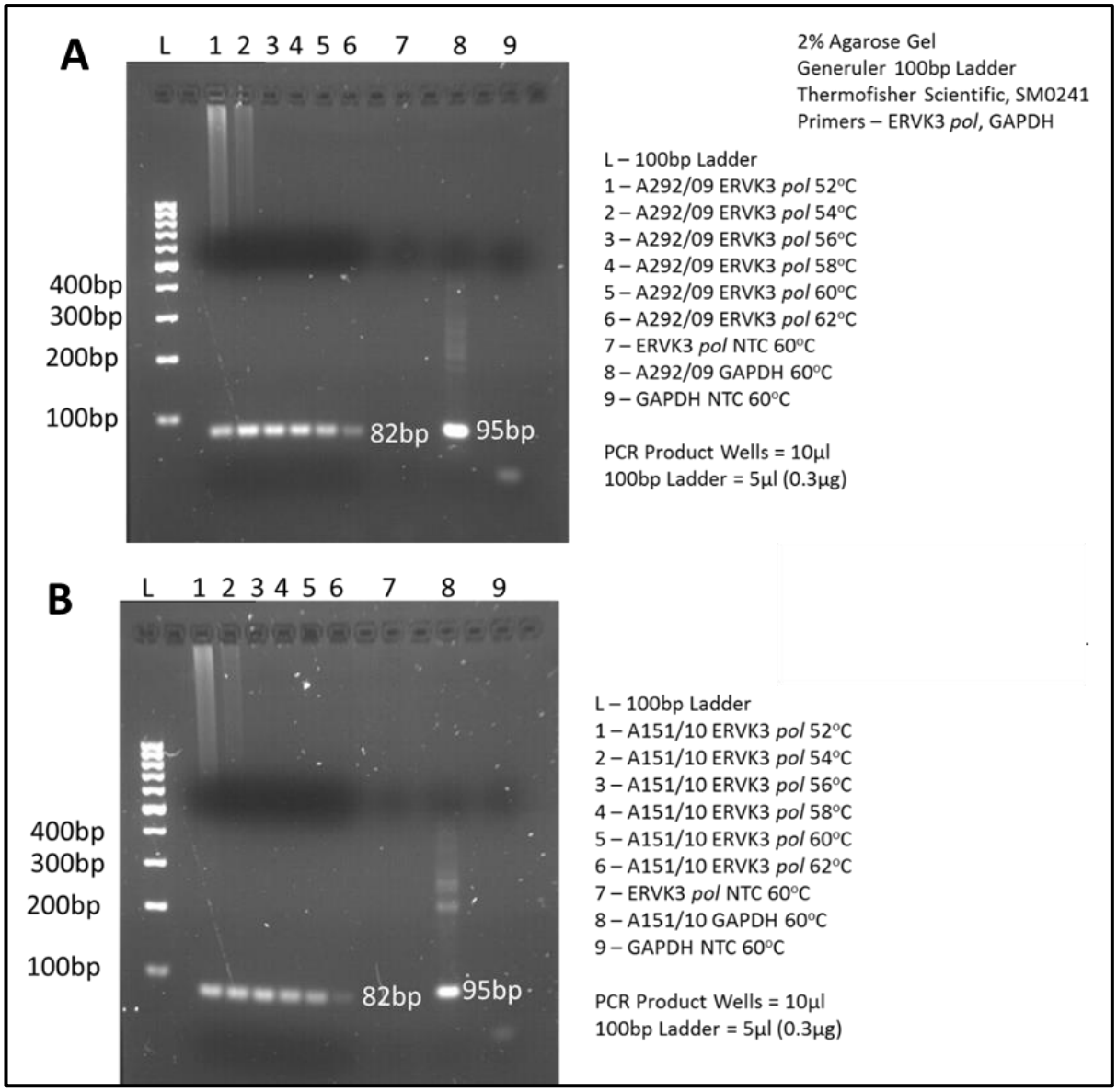


Figure S202. 2% Gel Electrophoresis Images for Gradient PCR of HERV-K3 *pol* Amplicons Using GAPDH as a Control.

The 2% gel electrophoresis images shown in the figure above show amplicon bands for each temperature zone in the thermal cycler utilising the HERV-K3 *pol* primers. GAPDH has been included as a positive control to establish that the cDNA template and PCR conditions are optimal.. Gel Image A) shows the PCR products for A292/09 non-ALS control sample and Gel Image B) shows the PCR products for A151/10 ALS sample.

Supplementary Table S17. Summary of ERVK3 Primer Efficiency Results

The table below details the slope, R^2 and calculated primer efficiency for the primer efficiency experiments utilising Water and Poly-A carrier RNA as a dilution medium utilising primary motor cortex sample A081/91.

Primer Target (Sample ID)	Slope	R^2	Efficiency
ERVK3 *pol* (A081/91) Water	-4.412	0.9965	68.52%
ERVK3 *pol* (A081/91) Carrier RNA	-3.3625	0.9963	98.33%

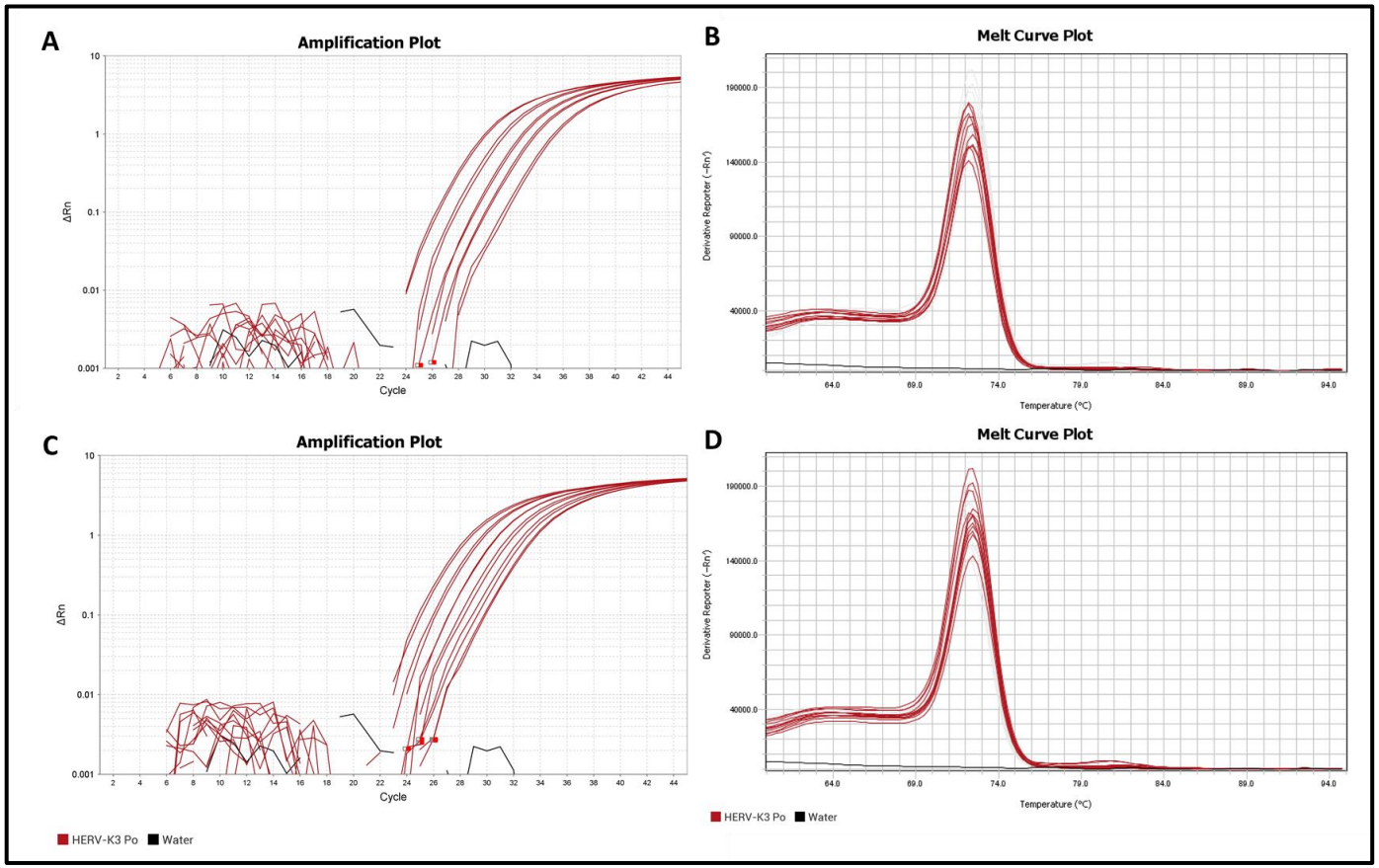


Figure S203. Primer Efficiency Amplification and Melt Curve Plots for HERV-K3 *pol* Amplicons When Using Standard Dilution Series in nuclease free water and Dilution Series Performed Utilising Poly-A Carrier RNA.

The Figure above shows the Amplification Plots (A&C) and Melt Curve Plots (B&D) for Primer Efficiency assays performed using dilution series performed using standard methods with nuclease free water (A&B) and the dilution series using a diluent solution containing 50µg/ml of Poly-A Carrier RNA (C&D).

Supplementary Table S18. Sequencing Data for HERV-K3 *pol* Amplicon Utilising Known ALS (A151/10) and Non-ALS Control (A292/09) Samples

The table below shows Sanger sequencing data obtained from Eurofins GATC Sequencing service and the closest match for the sequence from the NCBI BLASTn search tool.

Primer Target	Sequence Obtained from GATC Sequencing	NCBI Reference Sequence and Accession Number of Closest Match	Sequence Coverage
HERV-K3 *pol* A292	GTATATCCCTCGAGCCAC ACCTATAGAAGGATGTAA TCCACGAGGTAAGACG	AC098613.2 Homo sapiens chromosome 3 clone RP11-24F11, complete sequence	46/46(100%)
HERV-K3 *pol* A151	TCGAGCCACACCTATAGA AGGATGTAATCCACGAGG TAAGA	AC098613.2 Homo sapiens chromosome 3 clone RP11-24F11, complete sequence	37/37(100%)

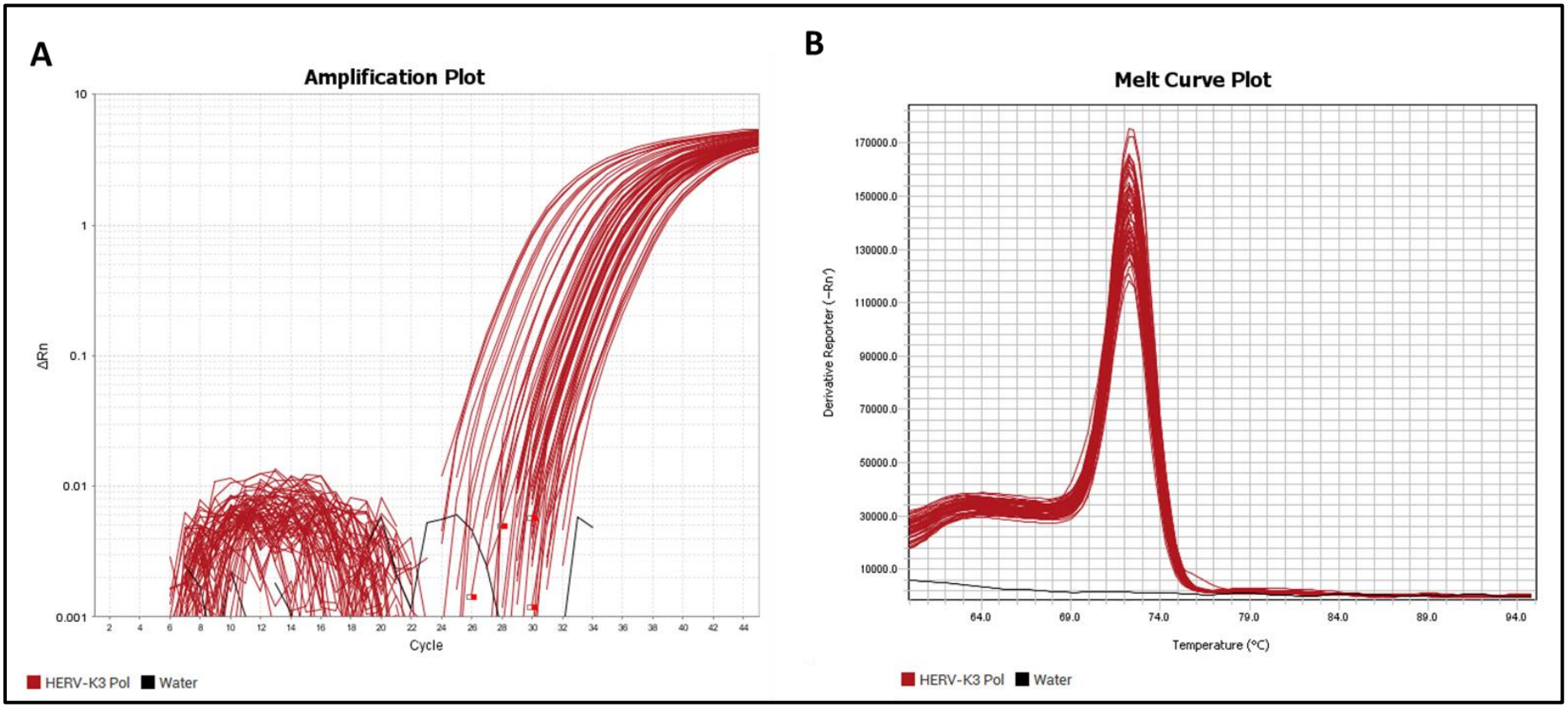


Figure S204. Amplification and Melt Curve Outputs for HERV-K3 *pol* RT-qPCR Assay Utilising n=10 ALS and n=10 Non-ALS Controls from Postmortem Primary Motor Cortex Brain Tissue Samples.

The figure above shows amplification (A) and melt curve (B) plots for HERV-K3 *pol env*. The black lines in each image represent water control reactions for the RT-qPCR assay. The baseline for measuring the Ct value for the samples in each gene target was ΔRn = 0.21.

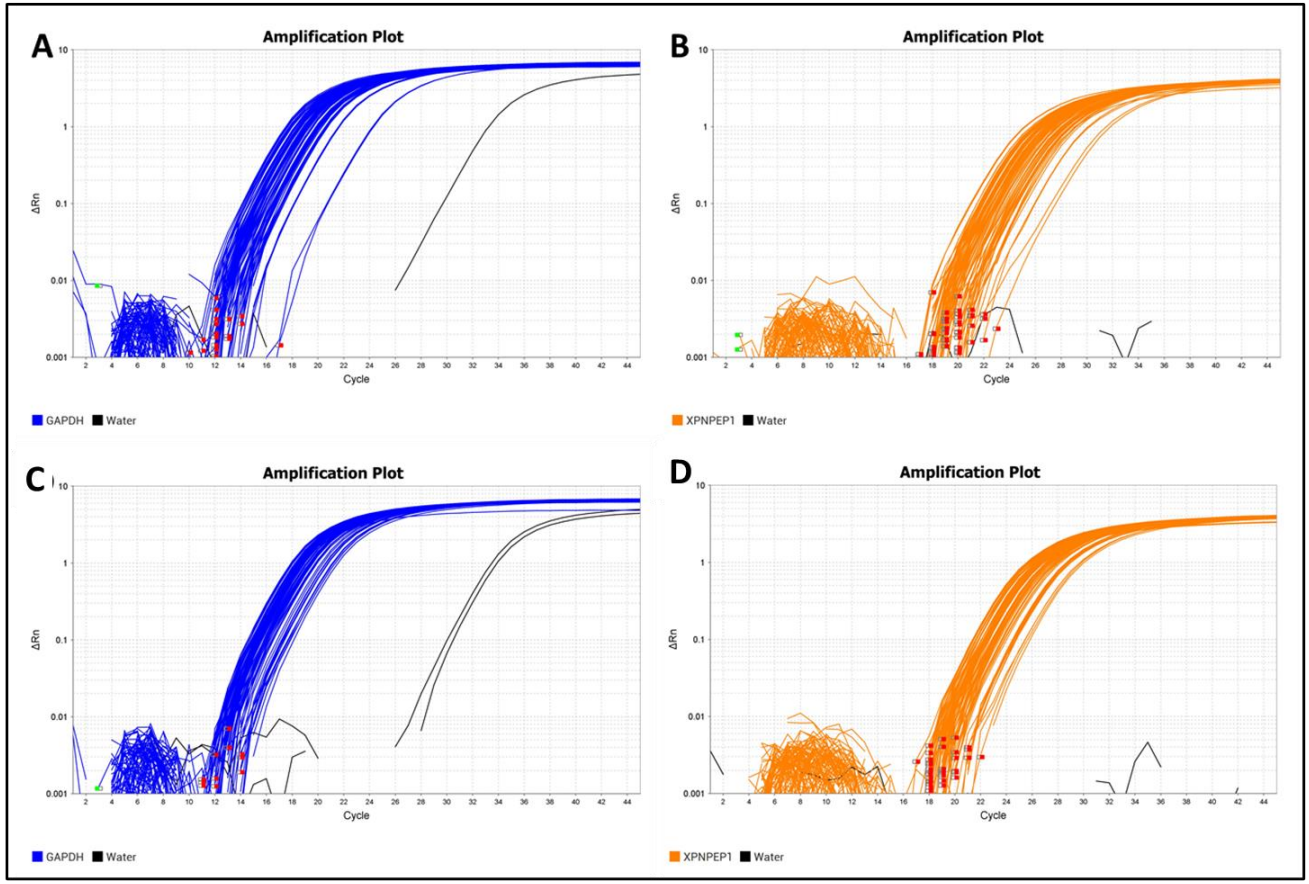


Figure S205. Amplification Plots generated by RT-qPCR following cDNA Amplification of GAPDH and XPNEP1Transcripts Present in n=54 ALS and n=37 non-ALS brain tissue samples.

The figure above shows amplification plots for A&C) GAPDH and B&D) XPNPEP1. The black lines in each image represent water control reactions for the RT-qPCR assay. The top row of amplification plots are from Run 1 of the gene targets and the bottom row from Run 2. The baseline for measuring the Ct value for the samples in each gene target was the same as in previous differential expression assays (ΔRn = 0.21).

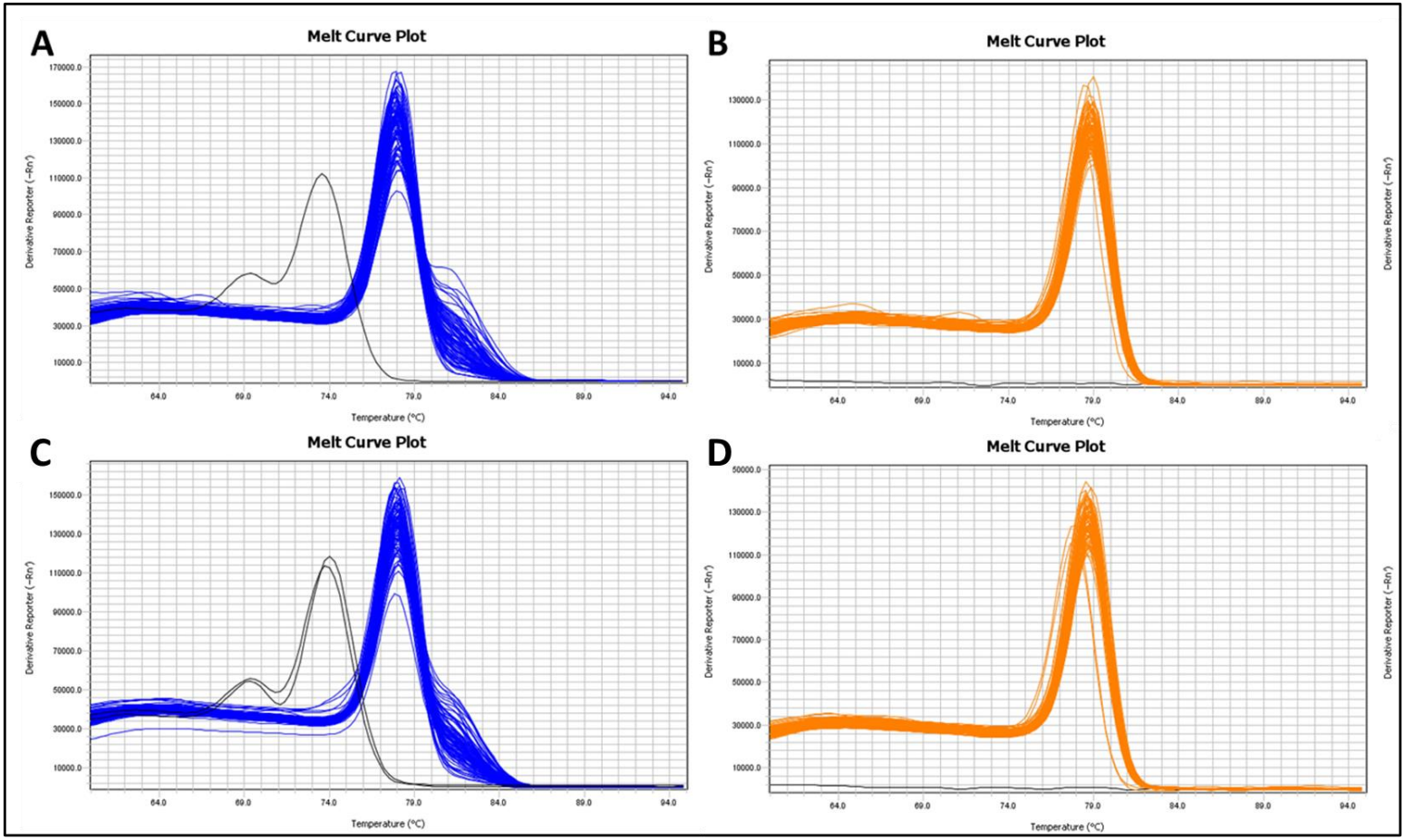


Figure S206. Melt Curve Plots generated by RT-qPCR following cDNA Amplification of GAPDH and XPNEP1 transcripts present in n=54 ALS and n=37 non-ALS brain tissue samples

The figure above shows melt curve plots for A&C) GAPDH, B&D) XPNPEP1. The black lines in each image represent water control reactions for the RT-qPCR assays. The top row of melt curve plots are from Run 1 of the gene targets and the bottom row from Run 2.

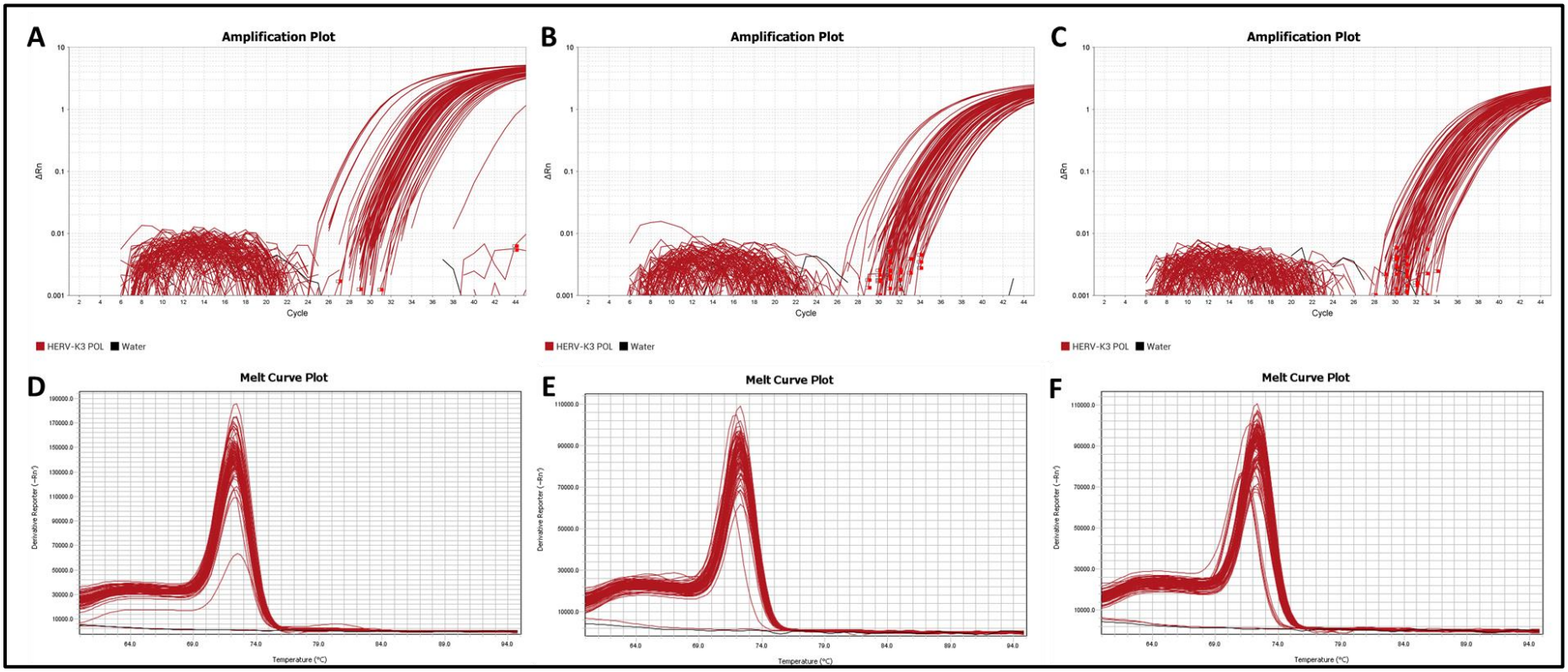


Figure S207. Amplification and Melt Curve Plots generated by RT-qPCR following cDNA Amplification of HERV-K3 *pol* transcripts present in n=54 ALS and n=37 non-ALS brain tissue samples

The figure above shows amplification and melt curve plots for HERV-K3 *pol* gene targets. The black lines in each image represent water control reactions for the RT-qPCR assays. The amplification and melt curve plots are differentiated in the figure above by A&D) Run 1, B&E) Run 2 and C&F) Run 3. The baseline for measuring the Ct value for the samples was the same as in previous differential expression assays (ΔRn = 0.21).

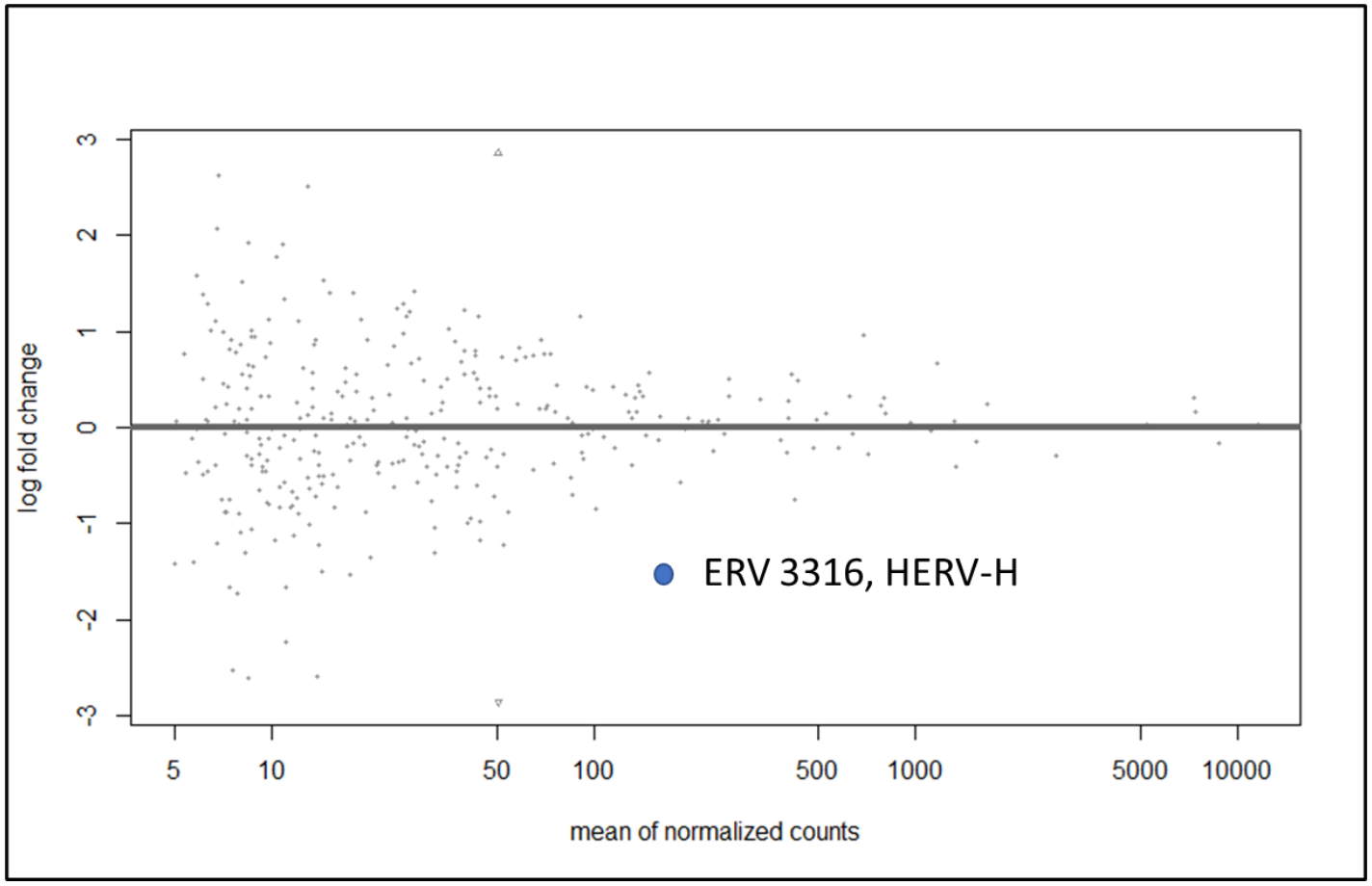


Figure S208 MA Plot of Log2 Fold Changes in Expression between in Postmortem Frontal Cortex ALS and Non-ALS Controls for ERVs.
The MA plot in the figure above shows the distribution of Log2 fold changes against the mean of normalised counts for individual ERVs identified in ALS and non-ALS controls. Statistically significant ERVs (p-value less than 0.01) would be identified by blue points if present. The significant ERV in this graph has been highlighted by increasing the size of the blue dot.

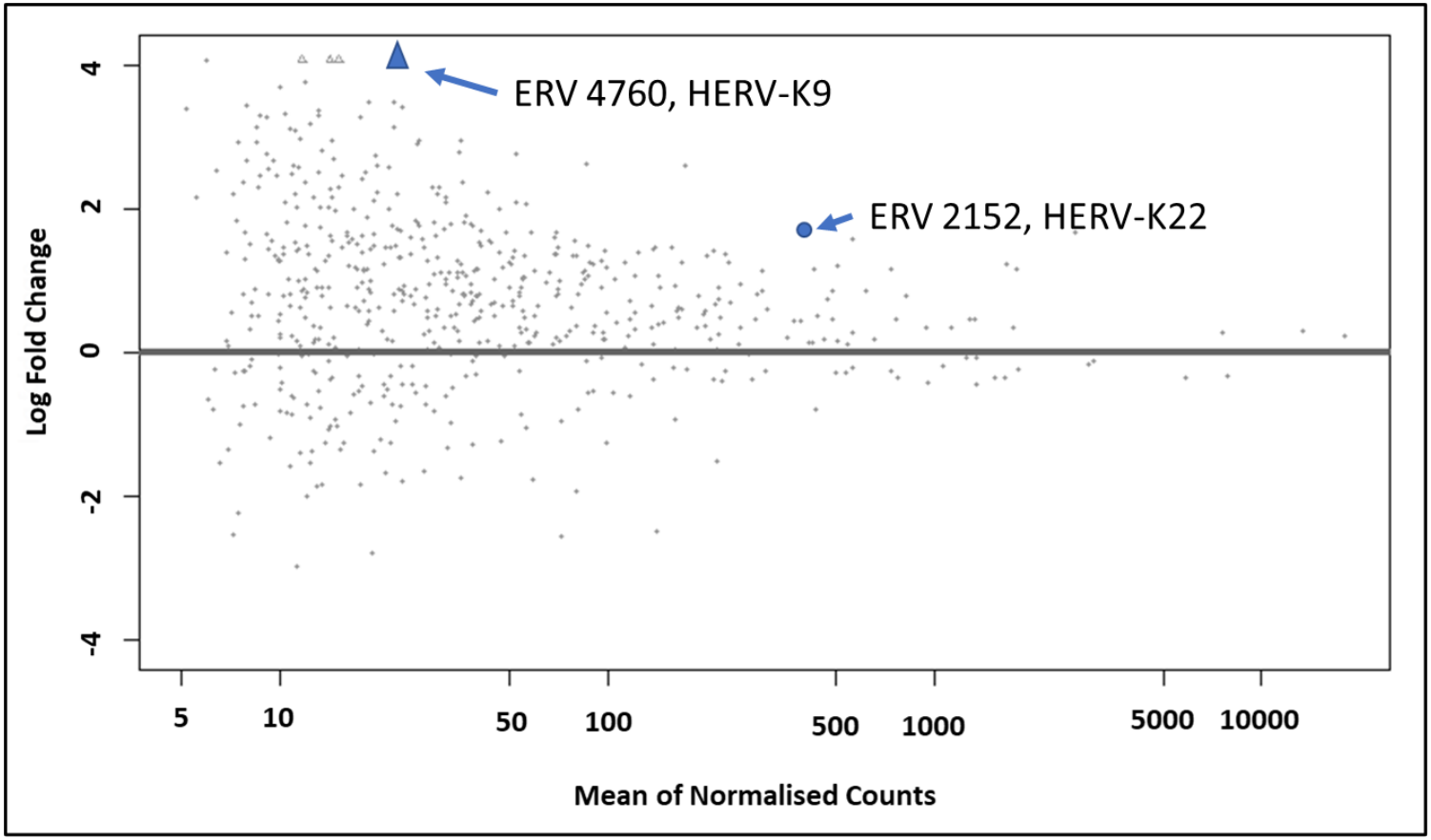


Figure S209. MA Plot of Log2 Fold Changes in Expression between in Postmortem Cerebellum ALS and Non-ALS Controls for ERVs.
The MA plot in the figure above shows the distribution of Log2 fold changes against the mean of normalised counts for individual ERVs identified in ALS and non-ALS controls. Statistically significant ERVs (p-value less than 0.01) would be identified by blue points if present. The significant ERV has been highlighted by a increasing the size of the blue dot and adding an arrow for ERV 4760 where it appears out of the plot's range.

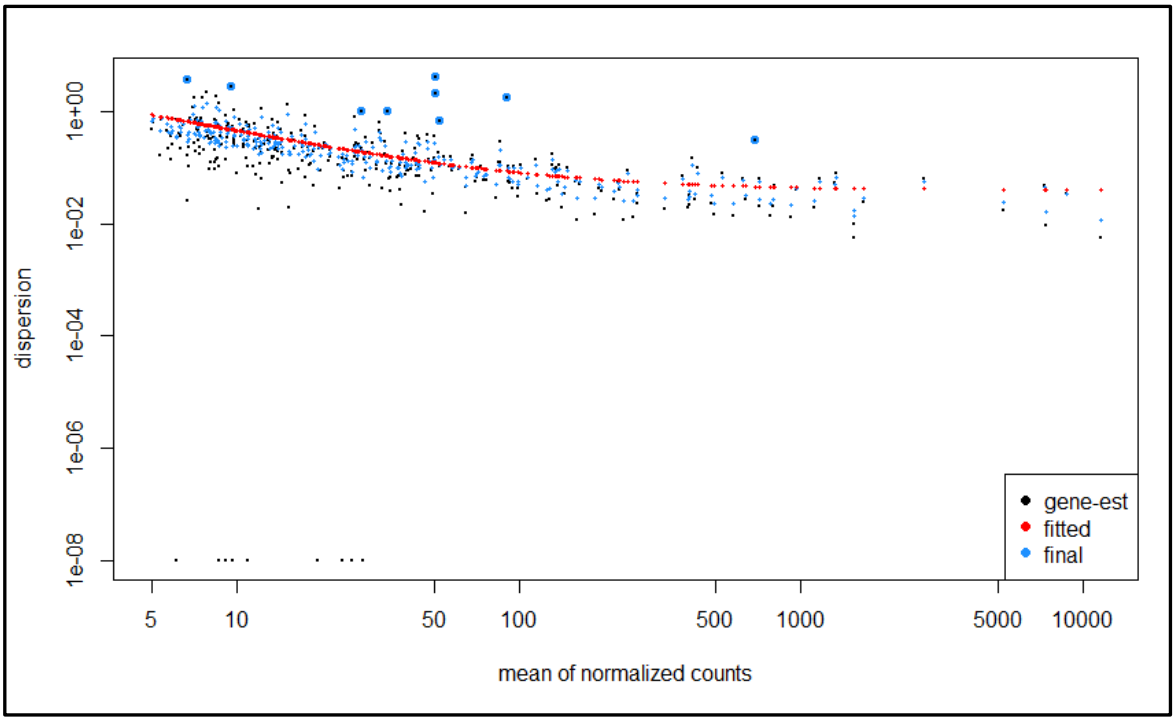


Figure S210. Dispersion Estimate Plot for Differential Expression Data Generated by DESeq2 from Frontal Cortex ALS and Non-ALS Controls.

The Figure above plots dispersion estimates for endogenous retrovirus differential expression data between ALS and Non-ALS control tissue from postmortem primary motor cortex samples over the mean count of those ERVs (Love, Huber and Anders, 2014). This is shown as shrinkage towards the red line (consensus) for non-outlier data points. The black points circled in blue above the main data set are defined as dispersion outliers and not shrunk towards the consensus line.

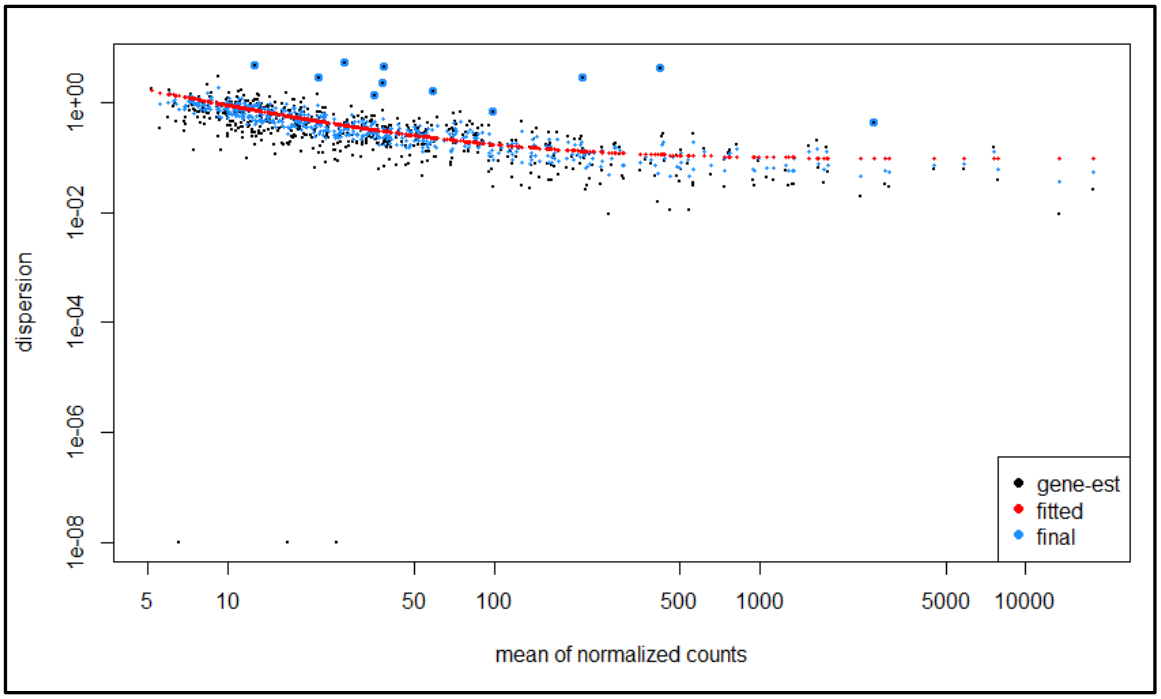


Figure S211. Dispersion Estimate Plot for Differential Expression Data Generated by DESeq2 from Cerebellum ALS and Non-ALS Controls.

The Figure above plots dispersion estimates for endogenous retrovirus differential expression data between ALS and Non-ALS control tissue from postmortem primary motor cortex samples over the mean count of those ERVs (Love, Huber and Anders, 2014). This is shown as shrinkage towards the red line (consensus) for non-outlier data points. The black points circled in blue above the main data set are defined as dispersion outliers and not shrunk towards the consensus line.

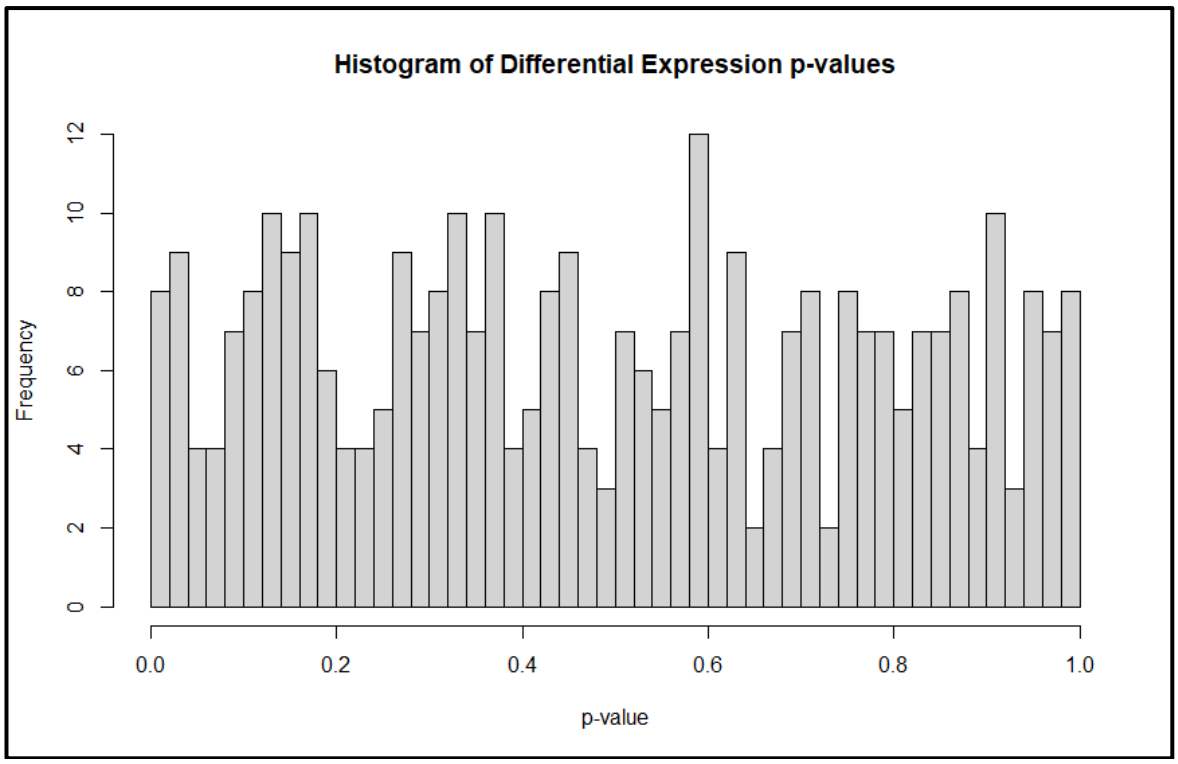


Figure S212. Histogram of p-value Frequency within DESeq2 Differential Expression Analysis of Frontal Cortex Tissue Samples

The figure above displays the frequency at which a given p-value occurs within the ALS vs Non-ALS control differential expression sample set. The p-values, while varied at different values broadly shows a uniform distribution in the sample set. This indicates that the data has no statistically significant differential expression values for ERVs in ALS when compared with controls.

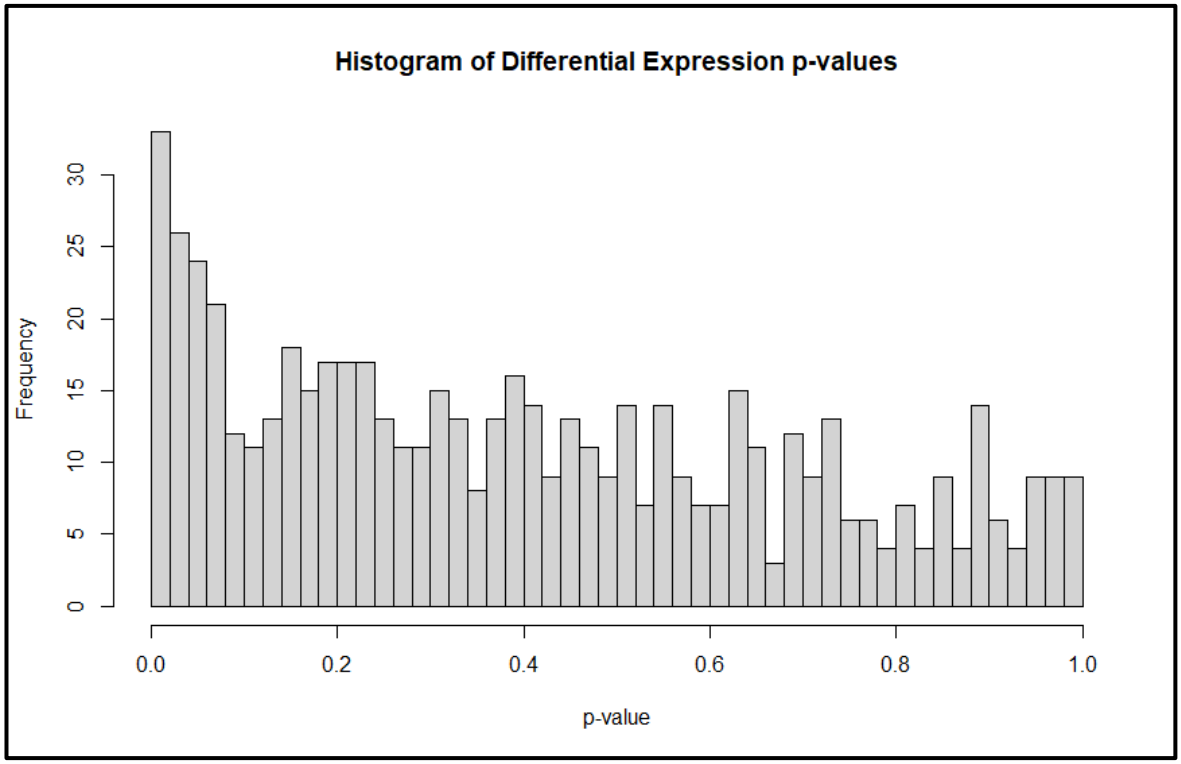


Figure S213. Histogram of p-value Frequency within DESeq2 Differential Expression Analysis of Cerebellum Tissue Samples

The figure above displays the frequency at which a given p-value occurs within the ALS vs Non-ALS control differential expression sample set. The p-values show a conservative distribution in the sample set, showing several samples which would be considered as significant by unadjusted p-value.

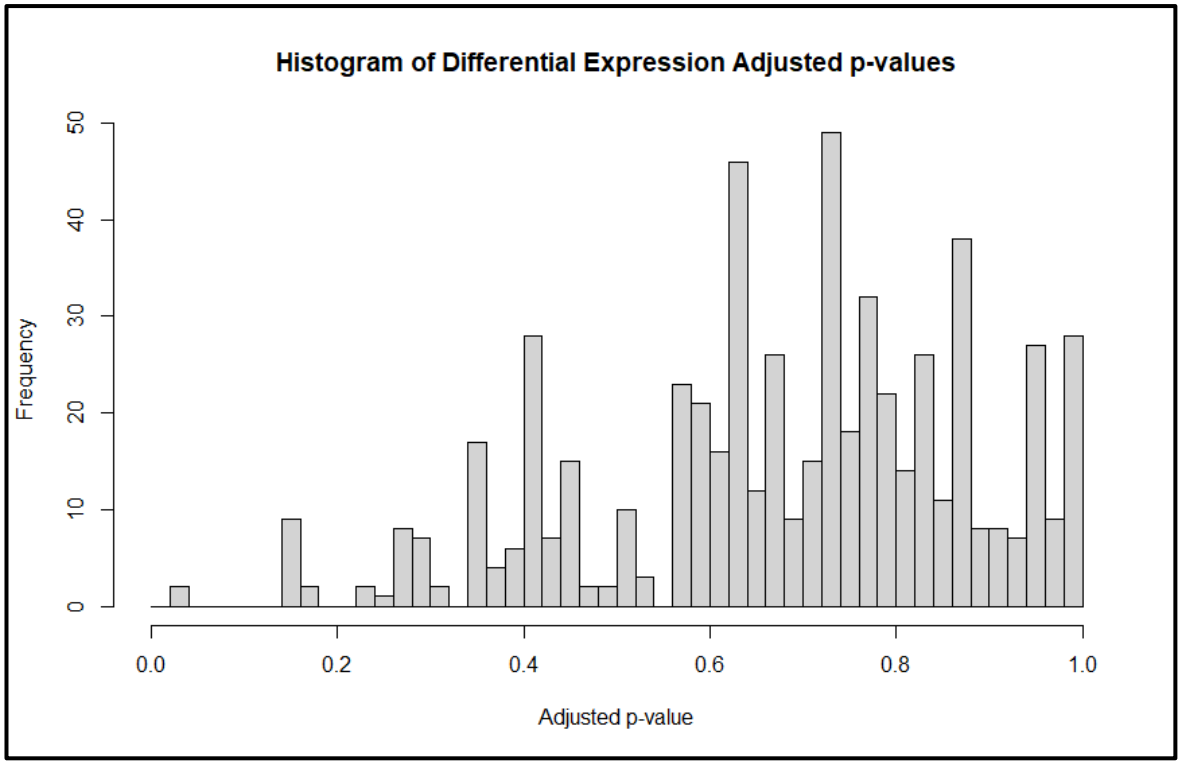


Figure S214. Histogram of Adjusted p-value Frequency within DESeq2 Differential Expression Analysis of Cerebellum Tissue Samples

The figure above displays the frequency at which a given adjusted p-value occurs within the ALS vs Non-ALS control differential expression sample set. The adjusted p-values, while varied at different values broadly shows a conservative distribution in the sample set.

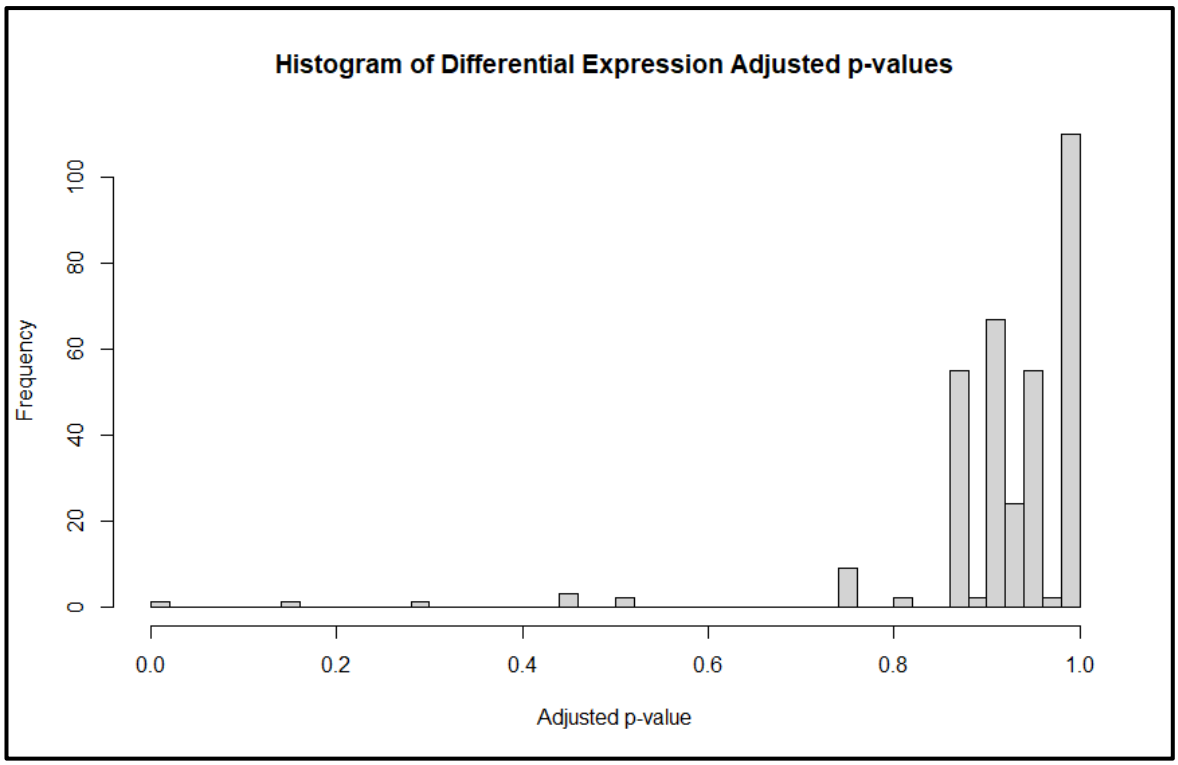


Figure S215. Histogram of Adjusted p-value Frequency within DESeq2 Differential Expression Analysis of Frontal Cortex Tissue Samples

The figure above displays the frequency at which a given adjusted p-value occurs within the ALS vs Non-ALS control differential expression sample set. The adjusted p-values, while varied at different values strongly shows a conservative distribution in the sample set.

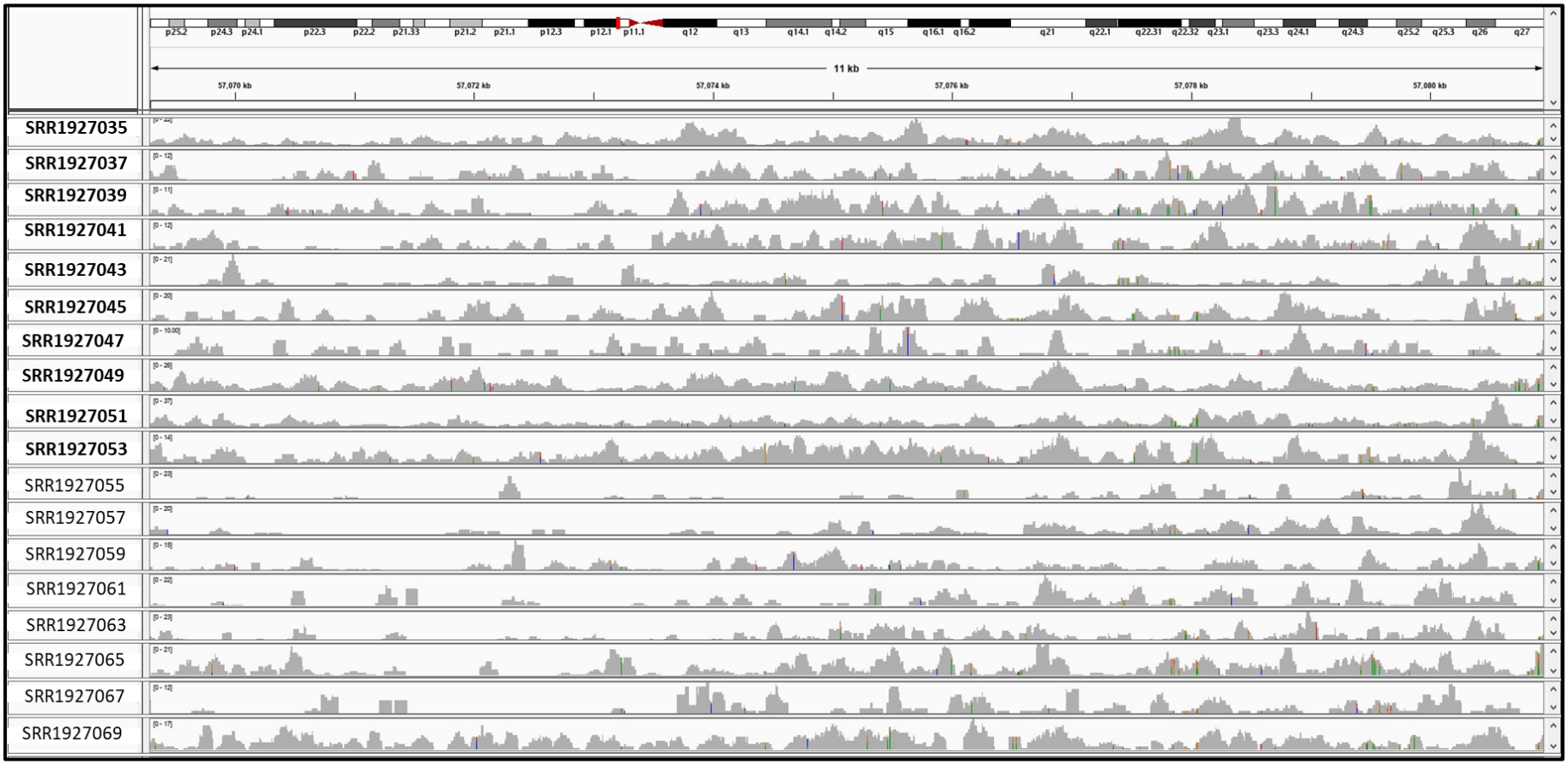


Figure S216. Read Alignment Coverage for ERVMap 2152 (HERV-K22)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 2152 (Chromosome 6, locus p12.1), identified as HERV-K22. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 6 with the red bar indicating the ERV location within the locus. The sample IDs featured in bold text in the figure above show ALS samples while the non-bold text are the non-ALS controls.

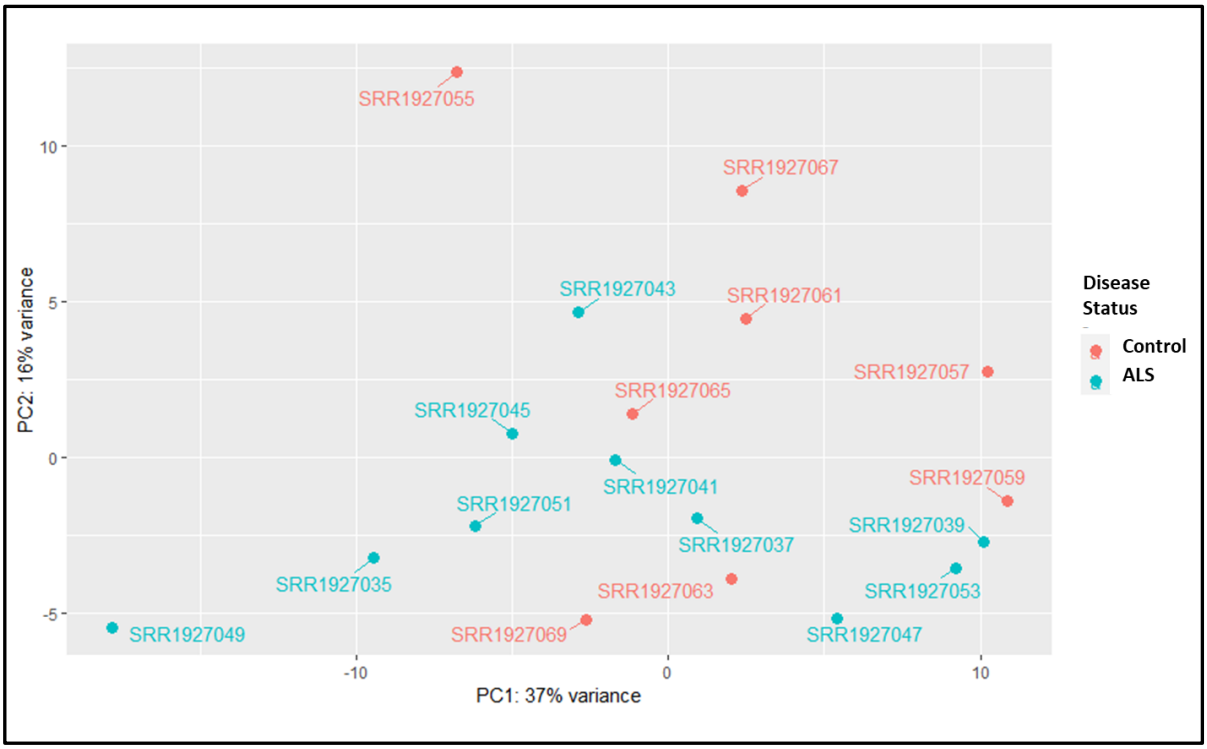


Figure S217. Principal Component Analysis (PCA) plots for DESeq2 Generated Normalised Counts from Differential Expression Analysis of Cerebellum Tissue from ALS and Non-ALS Controls.

The principal component analysis (PCA) shown in the figure above plots variation between all ERVs in each sample using normalised counts for each gene per patient sample. These principal component values are then plotted against each other to group samples together based on similar expression profiles of ERVs.

Figure S218. Principal Component Analysis (PCA) plots for DESeq2 Generated Normalised Counts from Differential Expression Analysis of Frontal Cortex Tissue from ALS and Non-ALS Controls.

The principal component analysis (PCA) shown in the figure above plots variation between all ERVs in each sample using normalised counts for each gene per patient sample. These principal component values are then plotted against each other to group samples together based on similar expression profiles of ERVs.

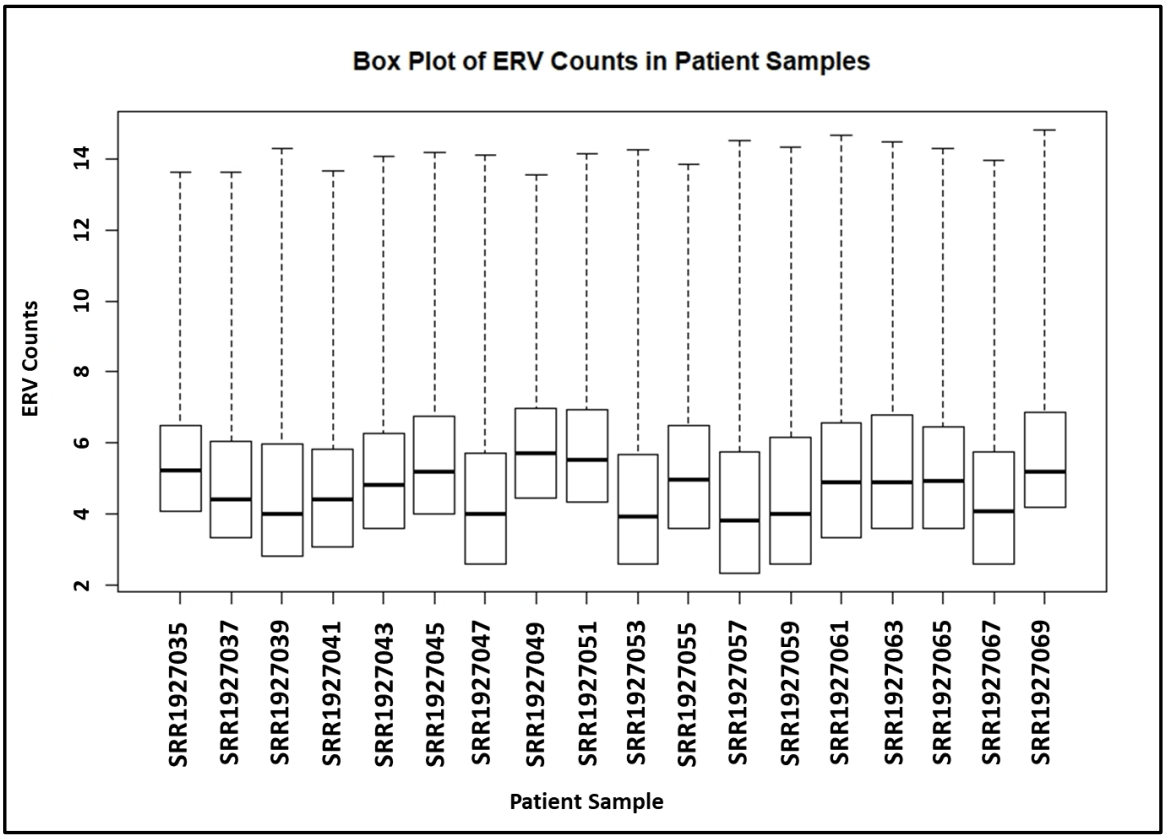


Figure S219. Box Plot of Endogenous Retrovirus Normalised Counts in Cerebellum Tissue between n=10 ALS and n=8 Non-ALS controls.

The figure above displays statistical information on the counts data for ERVs within the n=10 ALS and n=8 non-ALS control cerebellum sample set. The thick line inside each of the plots shows the median value for the data while the upper and lower end of the bars represent the interquartile ranges. The “whisker” above the plot, represented by the dotted line, shows the highest count recorded for an individual ERV within the sample.

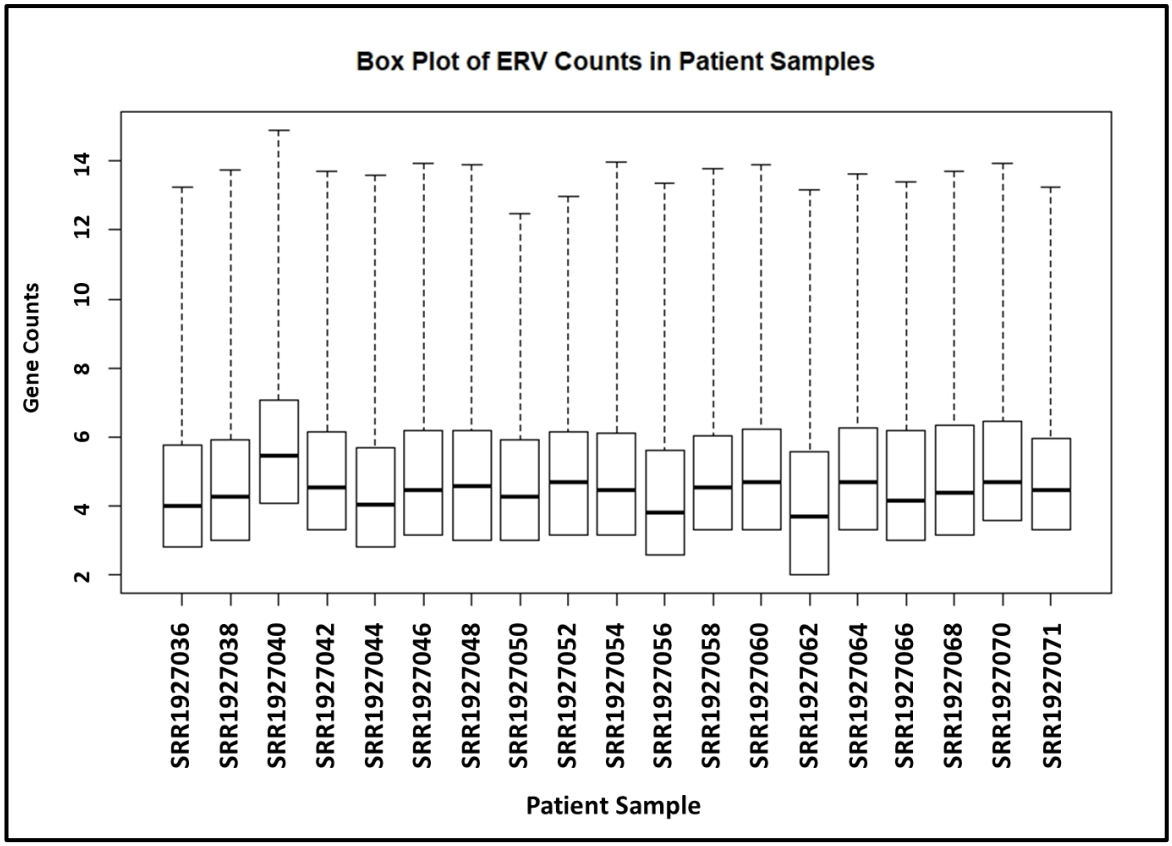


Figure S220. Box Plot of Endogenous Retrovirus Normalised Counts in Frontal Cortex Tissue between n=10 ALS and n=8 Non-ALS controls.

The figure above displays statistical information on the counts data for ERVs within the n=10 ALS and n=8 non-ALS control frontal cortex sample set. The thick line inside each of the plots shows the median value for the data while the upper and lower end of the bars represent the interquartile ranges. The “whisker” above the plot, represented by the dotted line, shows the highest count recorded for an individual ERV within the sample.

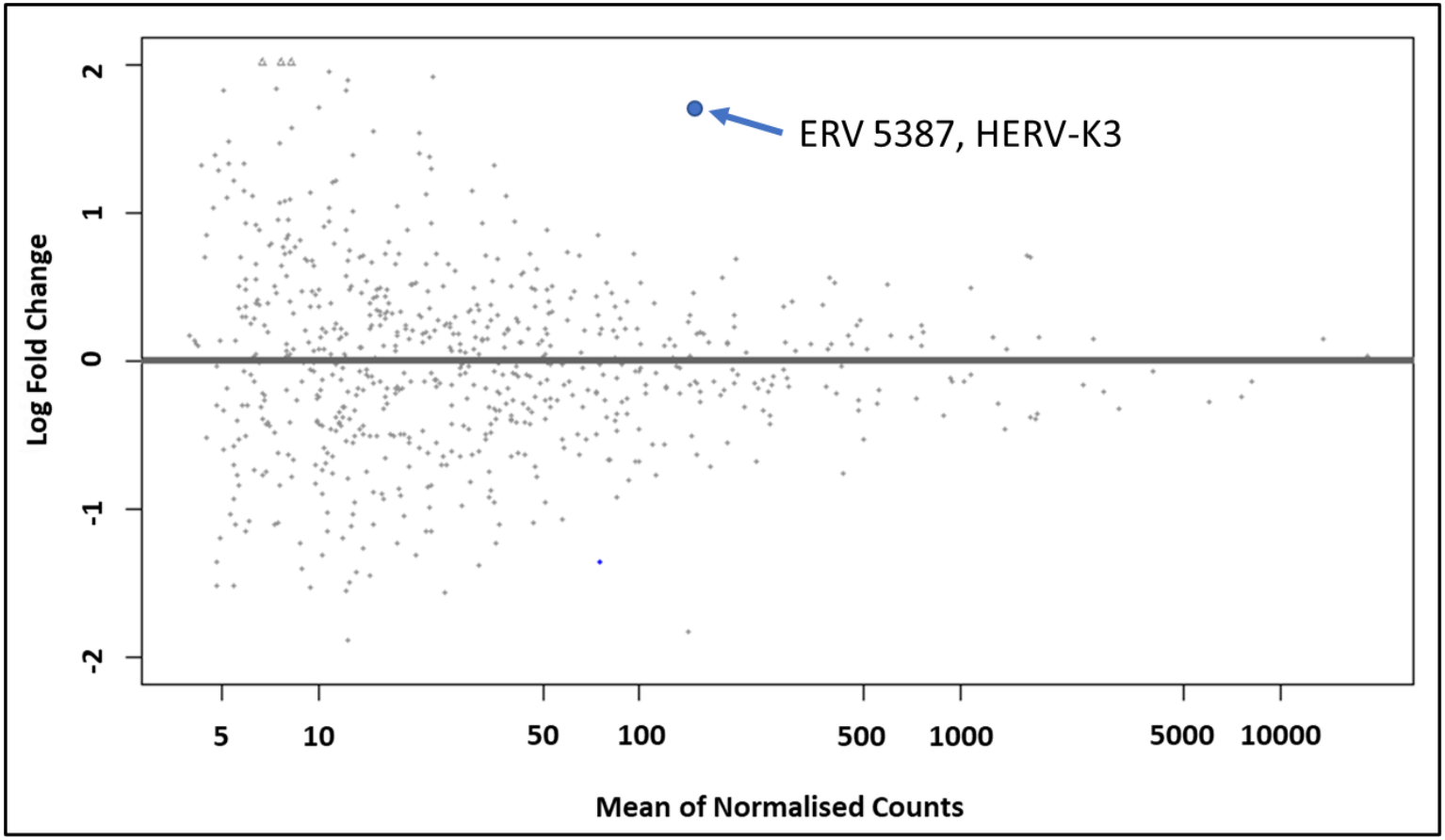


Figure S221. MA Plot of Log2 Fold Changes in Expression between Postmortem Cerebellum ALS and Non-ALS Controls, Inclusive of C9orf72 Patient Samples.

The MA plot in the figure above shows the distribution of Log2 fold changes against the mean of normalised counts for individual ERVs identified in ALS and non-ALS controls. Statistically significant ERVs (p-value less than 0.01) would be identified by blue points if present. The significant ERV in this graph has been highlighted by increasing the size of the blue dot.

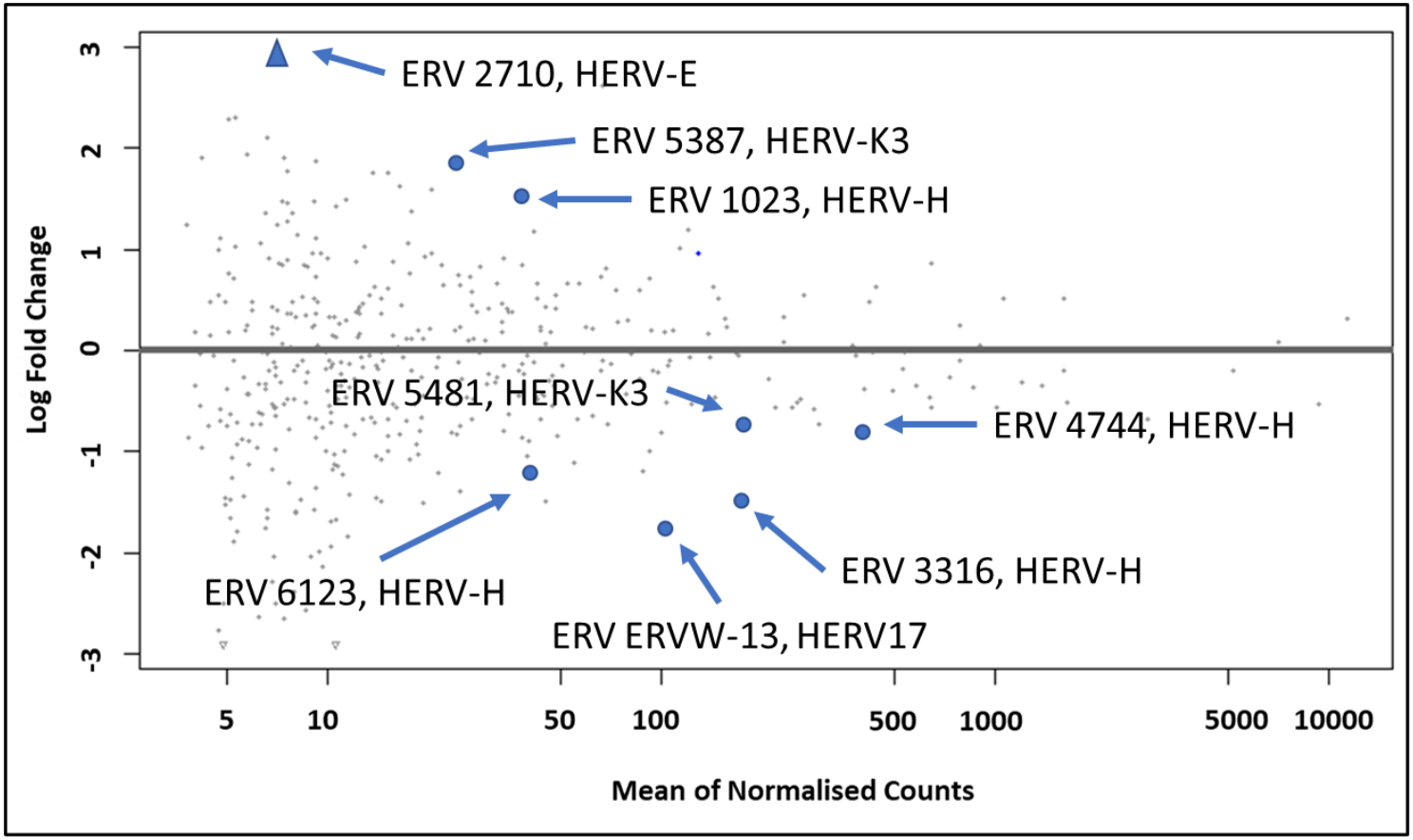


Figure S222. MA Plot of Log2 Fold Changes in Expression between Postmortem Frontal Cortex ALS and Non-ALS Controls, Inclusive of C9orf72 Patient Samples.

The MA plot in the figure above shows the distribution of Log2 fold changes against the mean of normalised counts for individual ERVs identified in ALS and non-ALS controls. Statistically significant ERVs (p-value less than 0.01) would be identified by blue points if present. The significant ERV in this graph has been highlighted by increasing the size of the blue dot. The blue arrow indicates that ERV 2710 is outside of the plot range.

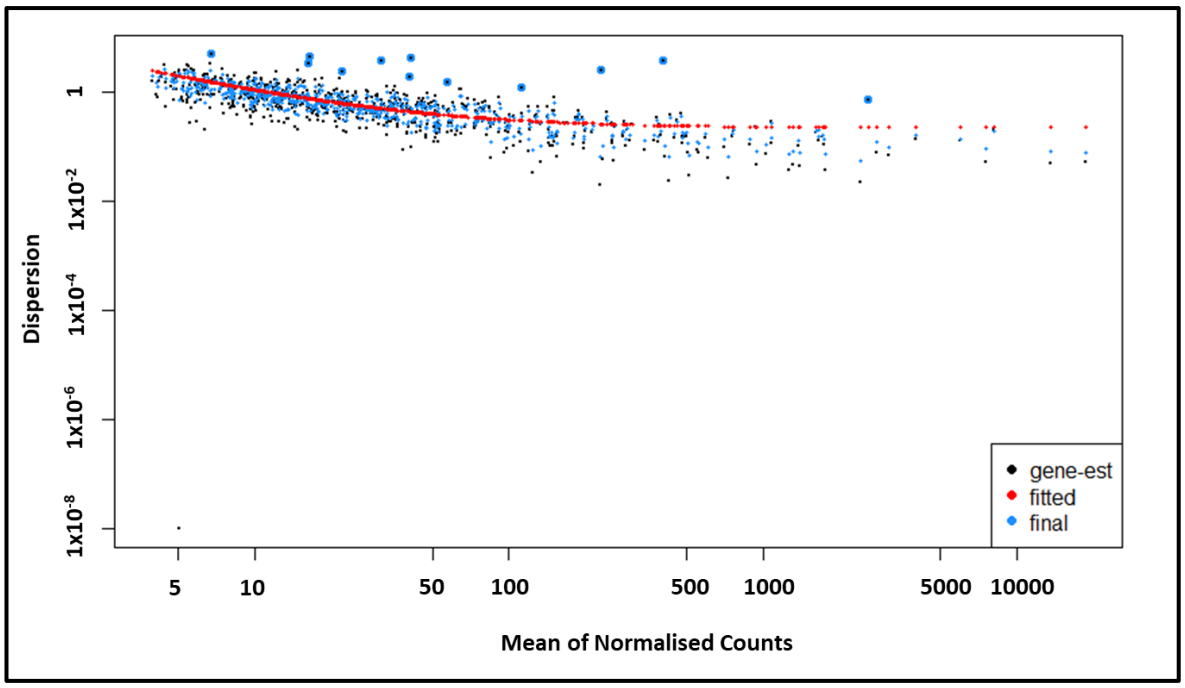


Figure S223. Dispersion Estimate Plot for Differential Expression Data Generated by DESeq2 from Cerebellum ALS and Non-ALS Controls, Inclusive of C9orf72 Samples.

The Figure above plots dispersion estimates for endogenous retrovirus differential expression data between ALS and Non-ALS control tissue from postmortem cerebellum samples, inclusive of C9orf72 samples, over the mean count of those ERVs (Love, Huber and Anders, 2014). This is shown as shrinkage towards the red line (consensus) for non-outlier data points. The black points circled in blue above the main data set are defined as dispersion outliers and not shrunk towards the consensus line.

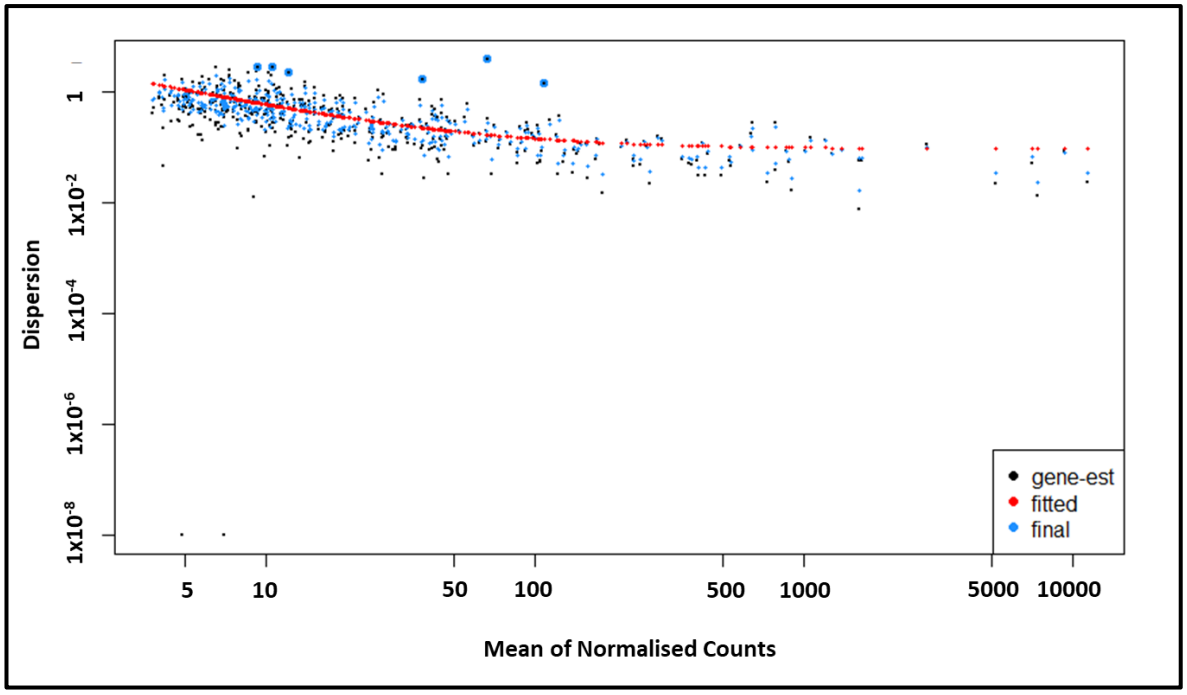


Figure S224. Dispersion Estimate Plot for Differential Expression Data Generated by DESeq2 from Frontal Cortex ALS and Non-ALS Controls, Inclusive of C9orf72 Samples.

The Figure above plots dispersion estimates for endogenous retrovirus differential expression data between ALS and Non-ALS control tissue from postmortem frontal cortex samples, inclusive of C9orf72 samples, over the mean count of those ERVs (Love, Huber and Anders, 2014). This is shown as shrinkage towards the red line (consensus) for non-outlier data points. The black points circled in blue above the main data set are defined as dispersion outliers and not shrunk towards the consensus line.

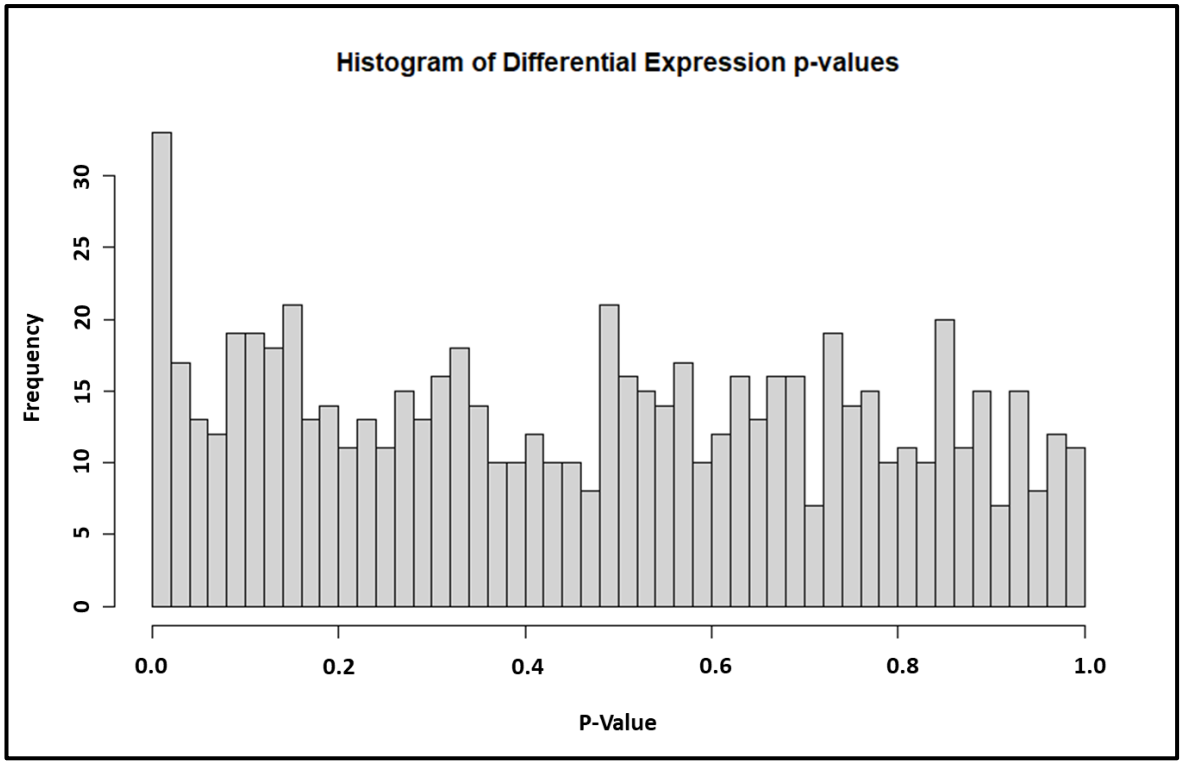


Figure S225. Histogram of p-value Frequency within DESeq2 Differential Expression Analysis of Cerebellum Tissue Samples, Inclusive of C9orf72 samples.

The figure above displays the frequency at which a given p-value occurs within the ALS vs Non-ALS control differential expression sample set. The p-values, while varied at different values broadly shows a uniform distribution in the sample set. This indicates that the data has no statistically significant differential expression values for ERVs in ALS when compared with controls.

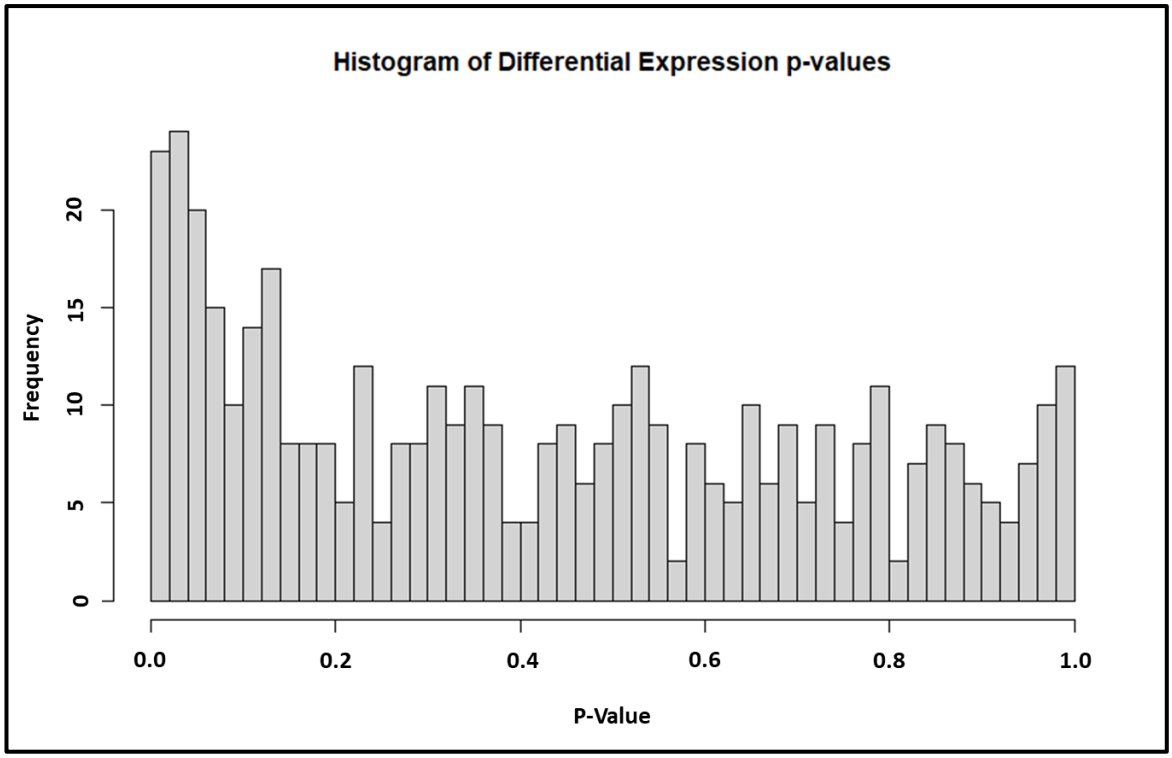


Figure S226. Histogram of p-value Frequency within DESeq2 Differential Expression Analysis of Frontal Cortex Tissue Samples, Inclusive of C9orf72 samples.

The figure above displays the frequency at which a given p-value occurs within the ALS vs Non-ALS control differential expression sample set. The p-values, while varied at different values broadly shows a conservative distribution in the sample set, skewed towards the 0.0 side of the figure.

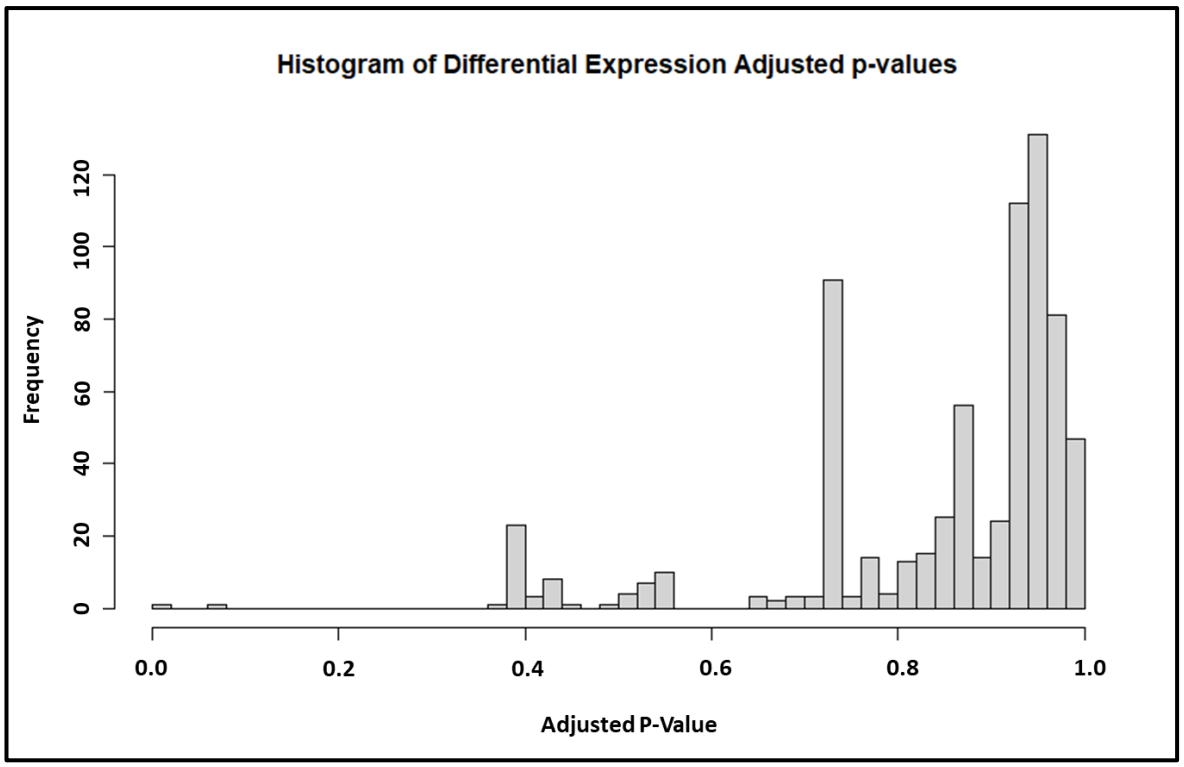


Figure S227. Histogram of Adjusted p-value Frequency within DESeq2 Differential Expression Analysis of Cerebellum Tissue Samples

The figure above displays the frequency at which a given adjusted p-value occurs within the ALS vs Non-ALS control differential expression sample set. The adjusted p-values, while varied at different values broadly shows a conservative distribution in the sample set.

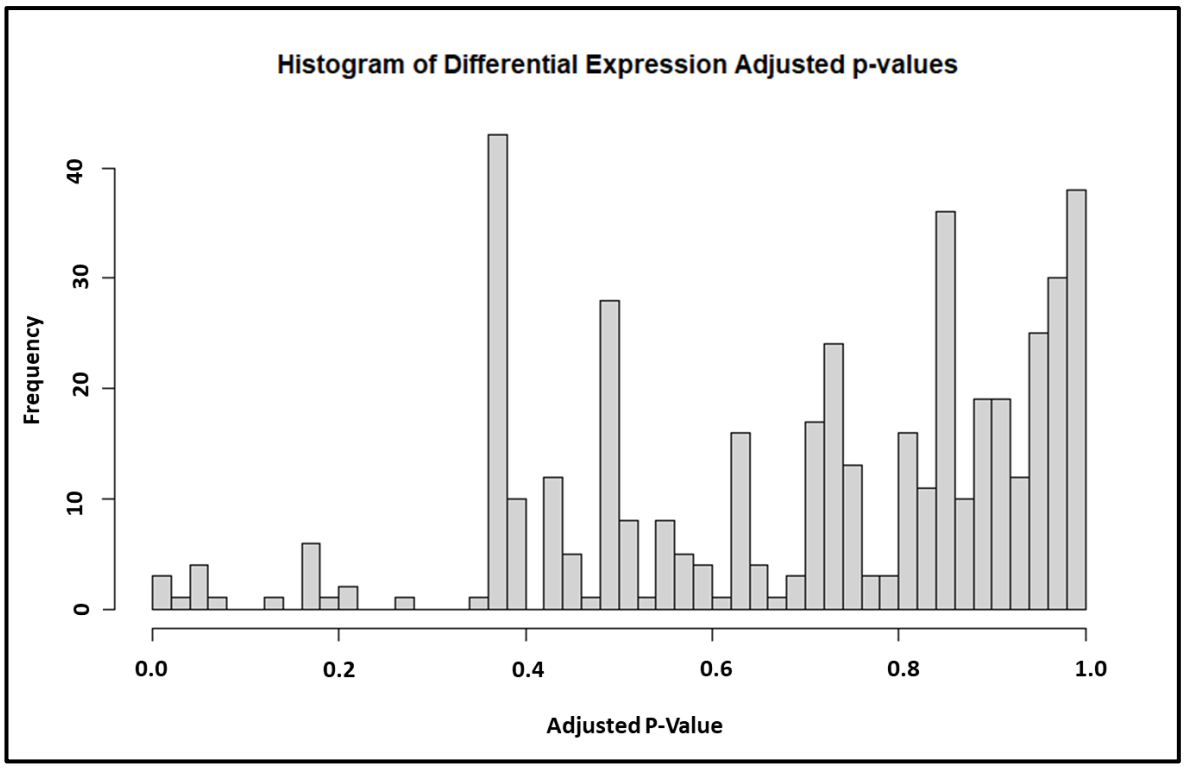


Figure S228. Histogram of Adjusted p-value Frequency within DESeq2 Differential Expression Analysis of Frontal Cortex Tissue Samples

The figure above displays the frequency at which a given adjusted p-value occurs within the ALS vs Non-ALS control differential expression sample set. The adjusted p-values, while varied at different values broadly shows a conservative distribution in the sample set.

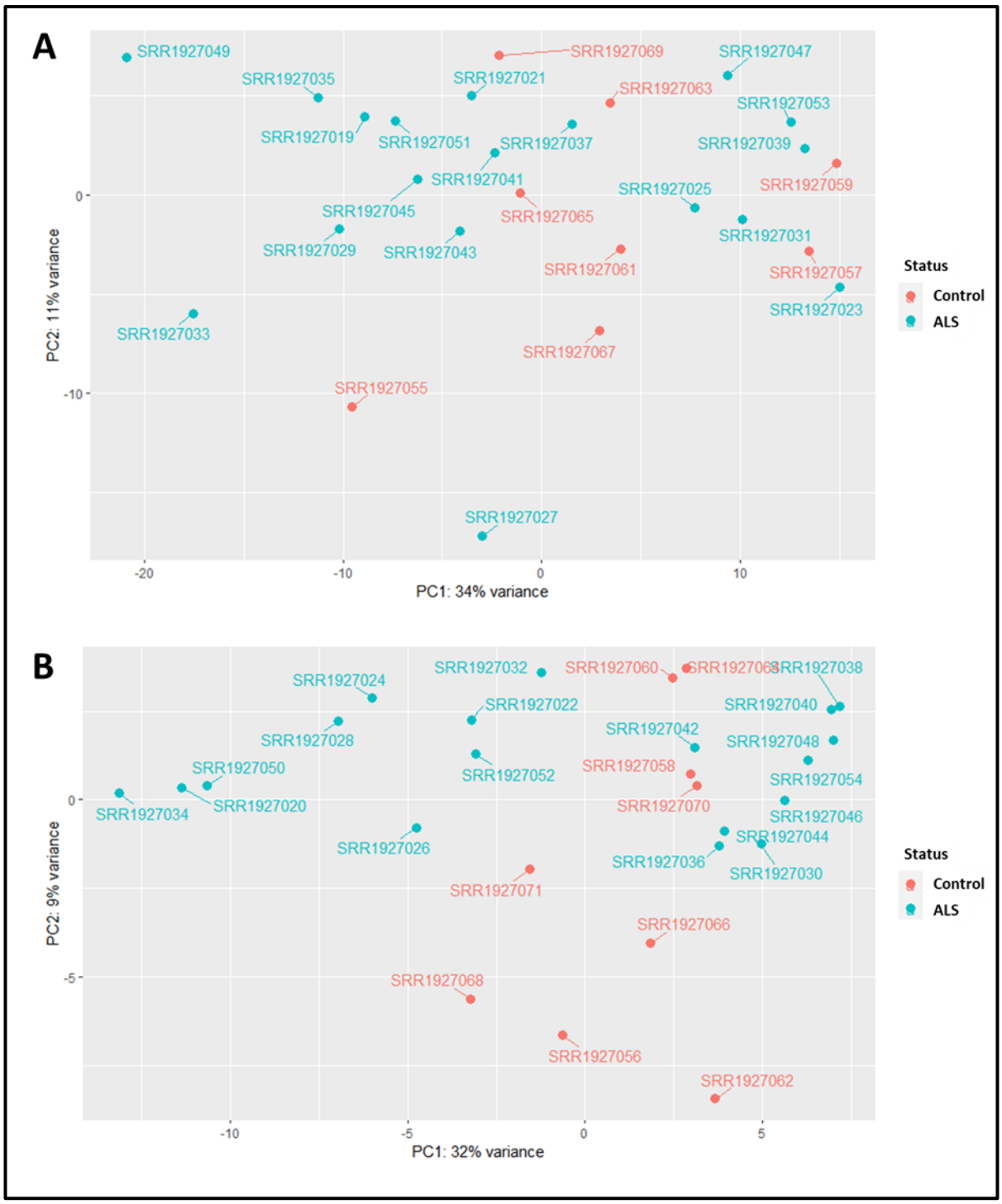


Figure S229. Principal Component Analysis (PCA) plots for DESeq2 Generated Normalised Counts from Differential Expression Analysis of Cerebellum and Frontal Cortex Tissue from C9orf72 ALS and Non-ALS Control Samples

The principal component analysis (PCA) shown in the figure above plots variation between all ERVs in each sample using normalised counts for each gene per patient sample. These principal component values are then plotted against each other to group samples together based on similar expression profiles of ERVs. In the Figure above A) shows the Control vs ALS comparison for Cerebellum Tissue and B) shows the comparison for Frontal Cortex Tissue.

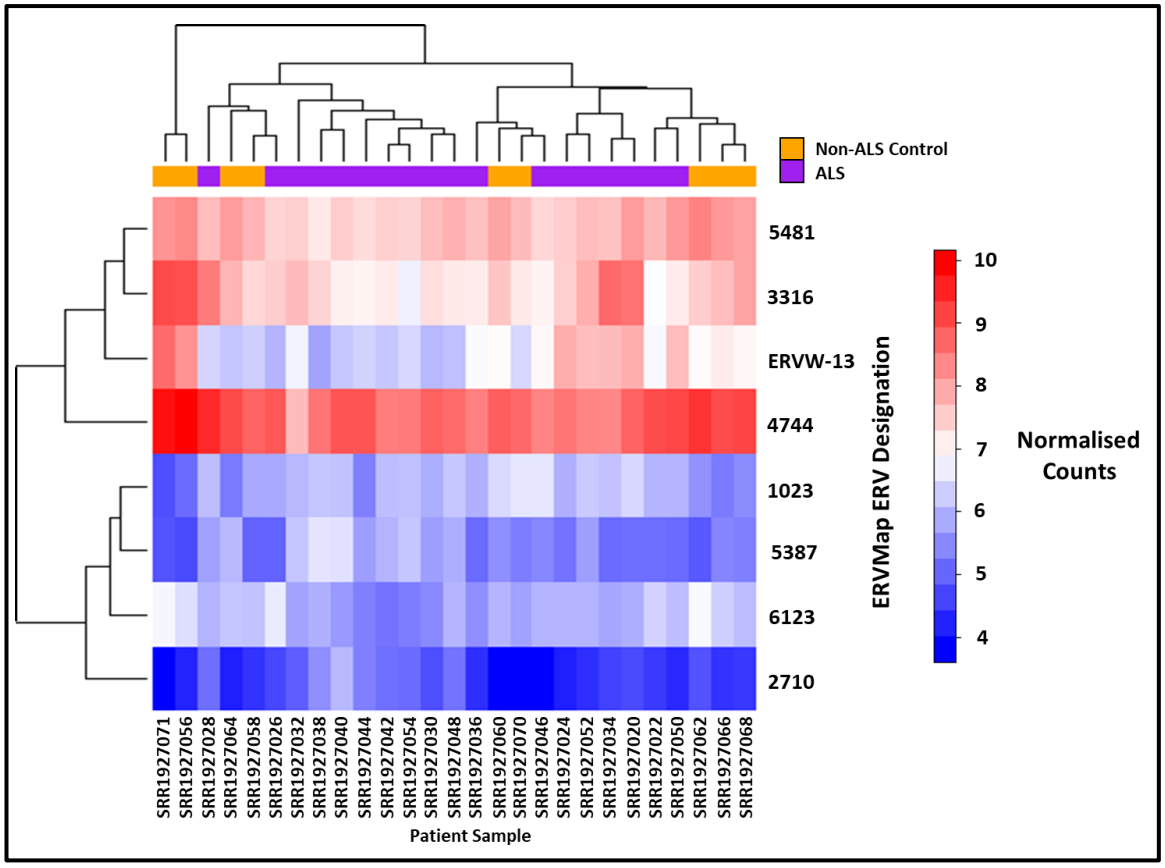


Figure S230. Heatmap of Differentially Expressed ERVs Identified in Frontal Cortex C9orf72 ALS vs Non-ALS Controls.

The heatmap displayed in the figure above shows the normalised counts data for ERVs identified by the ERVmap.bed file with low expressed ERV members filtered out. The rows and columns are hierarchically clustered to group together samples and ERVs with similar expression profiles based on normalised counts data generated from DESeq2. Also included in the cells above the counts matrix identifies those samples which are from the ALS (purple) and non-ALS control (orange) sample sets.

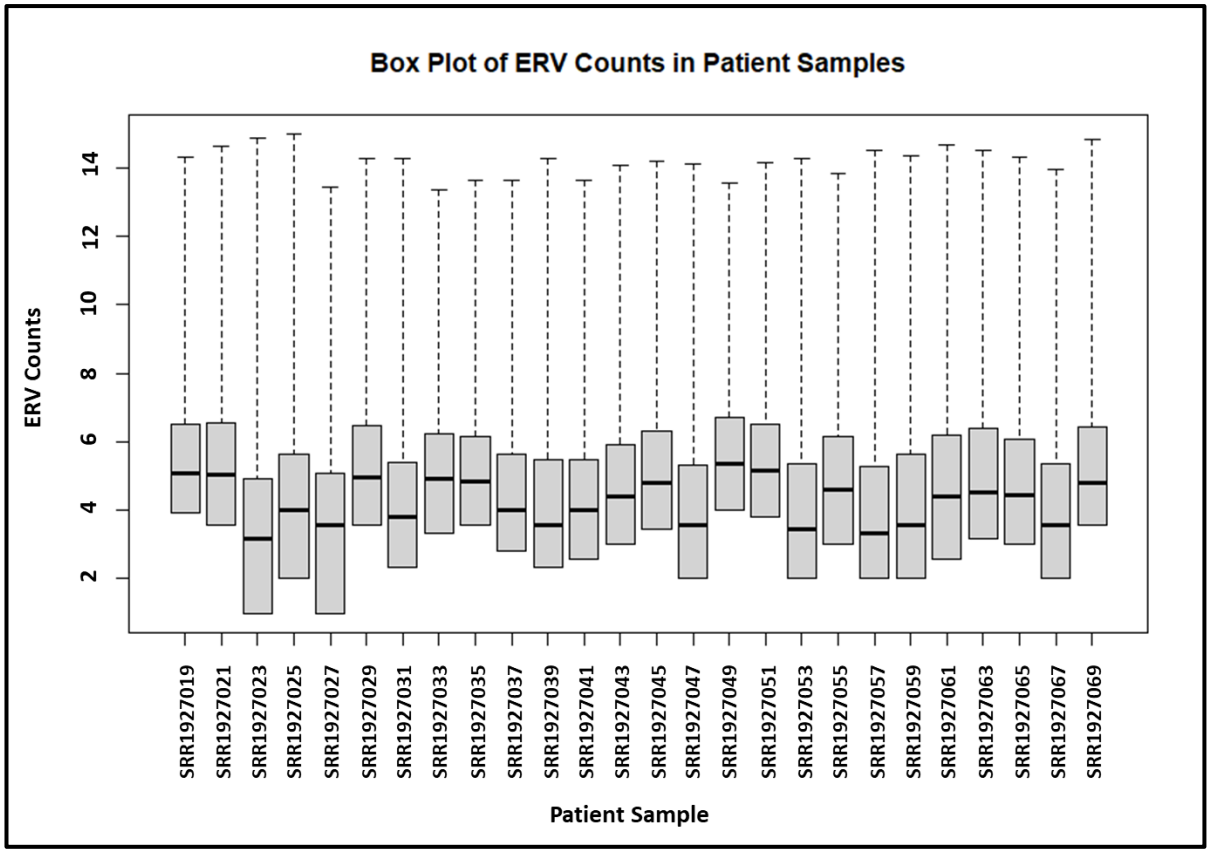


Figure S231. Box Plot of Endogenous Retrovirus Normalised Counts in Cerebellum Tissue between n=18 ALS and n=8 Non-ALS controls.

The figure above displays statistical information on the counts data for ERVs within the n=18 ALS and n=8 non-ALS control cerebellum sample set, inclusive of C9orf72 samples. The thick line inside each of the plots shows the median value for the data while the upper and lower end of the bars represent the interquartile ranges. The “whisker” above the plot, represented by the dotted line, shows the highest count recorded for an individual ERV within the sample.

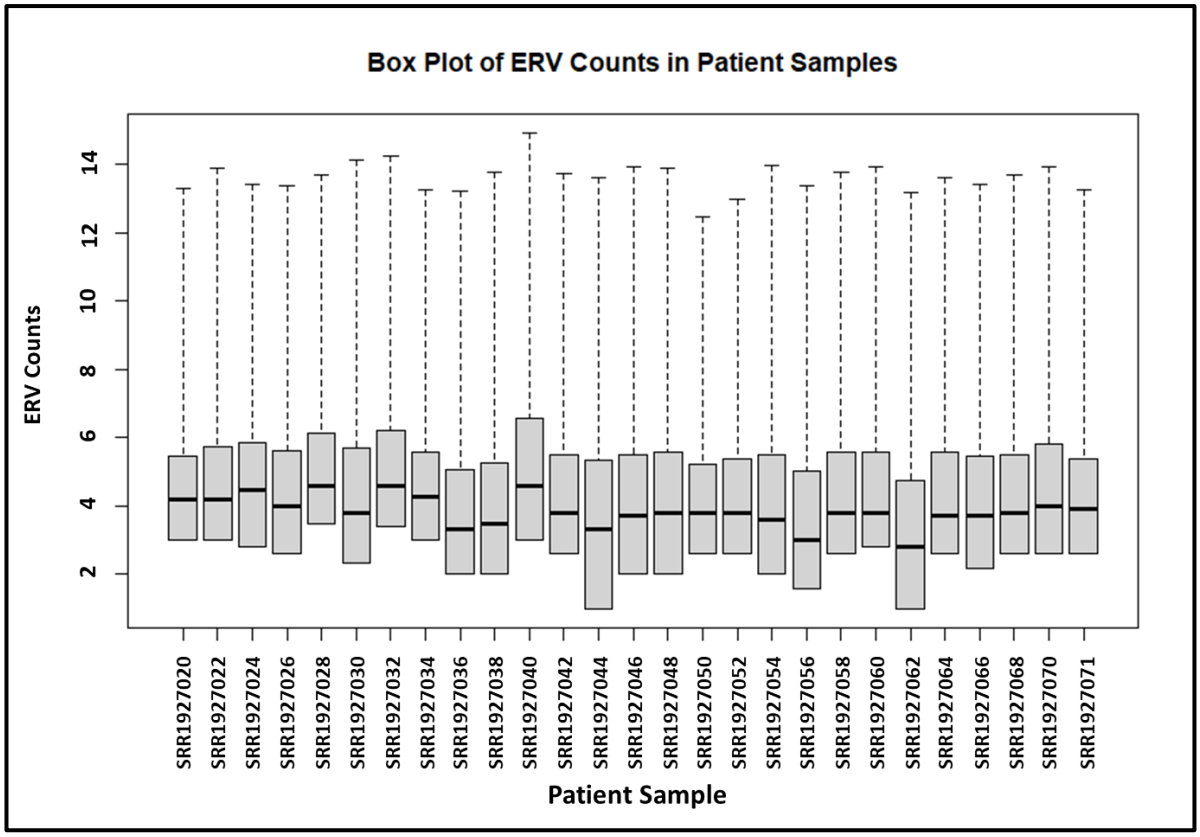


Figure S232. Box Plot of Endogenous Retrovirus Normalised Counts in Frontal Cortex Tissue between n=18 ALS and n=8 Non-ALS controls.

The figure above displays statistical information on the counts data for ERVs within the n=18 ALS and n=8 non-ALS control frontal cortex sample set, inclusive of C9orf72 samples. The thick line inside each of the plots shows the median value for the data while the upper and lower end of the bars represent the interquartile ranges. The “whisker” above the plot, represented by the dotted line, shows the highest count recorded for an individual ERV within the sample.

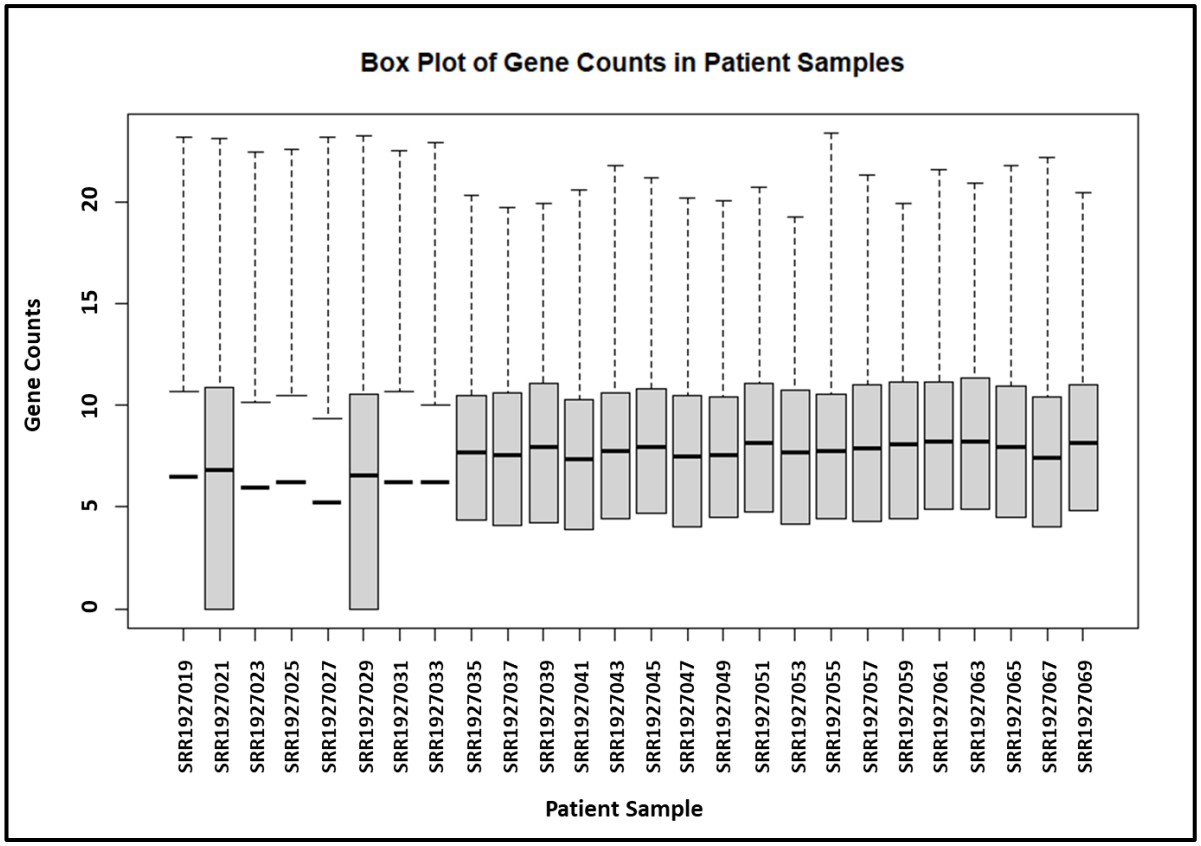


Figure S233. Box Plot of Normalised Gene Counts in Cerebellum Tissue between n=18 ALS and n=8 Non-ALS controls.

The figure above displays statistical information on the counts data for genes within the n=18 ALS and n=8 non-ALS control cerebellum sample set, inclusive of C9orf72 samples. The thick line inside each of the plots shows the median value for the data while the upper and lower end of the bars represent the interquartile ranges. The “whisker” above the plot, represented by the dotted line, shows the highest count recorded for an individual ERV within the sample.

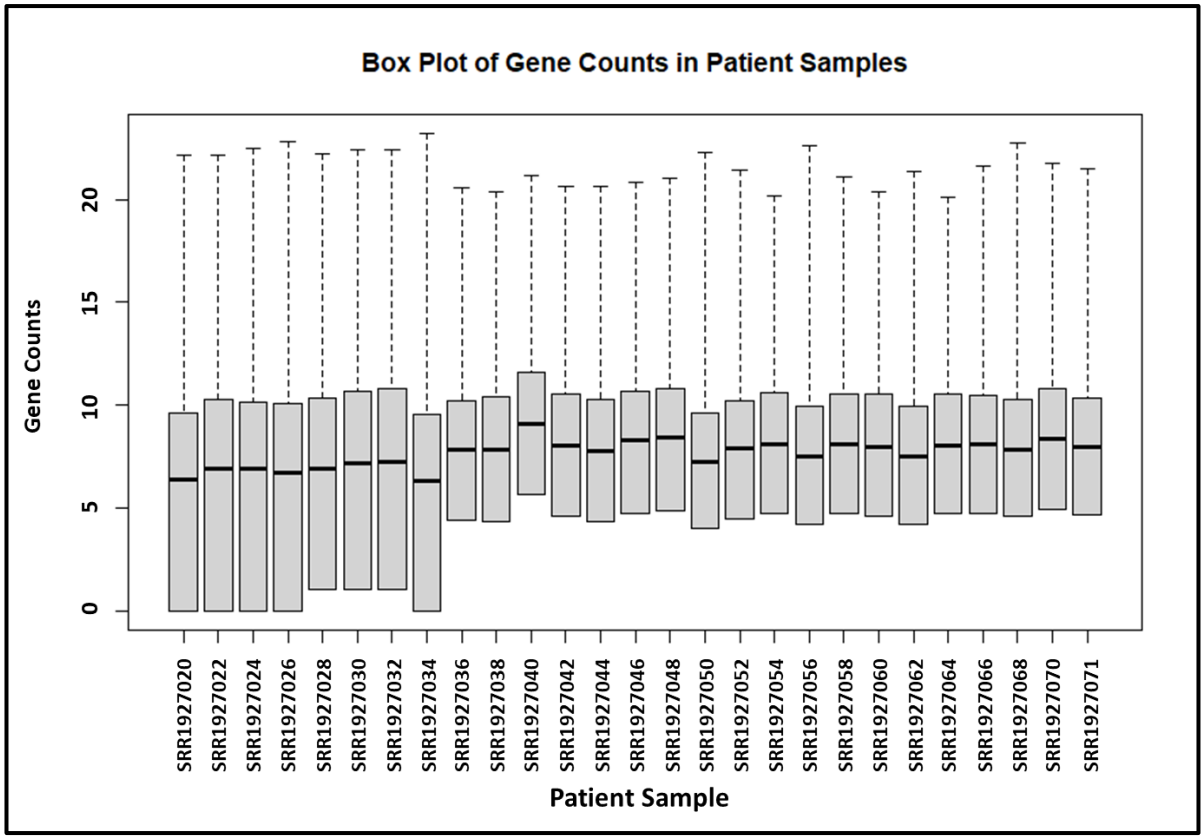


Figure S234. Box Plot of Normalised Gene Counts in Frontal Cortex Tissue between n=18 ALS and n=8 Non-ALS controls.

The figure above displays statistical information on the counts data for genes within the n=18 ALS and n=8 non-ALS control frontal cortex sample set, inclusive of C9orf72 samples. The thick line inside each of the plots shows the median value for the data while the upper and lower end of the bars represent the interquartile ranges. The “whisker” above the plot, represented by the dotted line, shows the highest count recorded for an individual ERV within the sample.

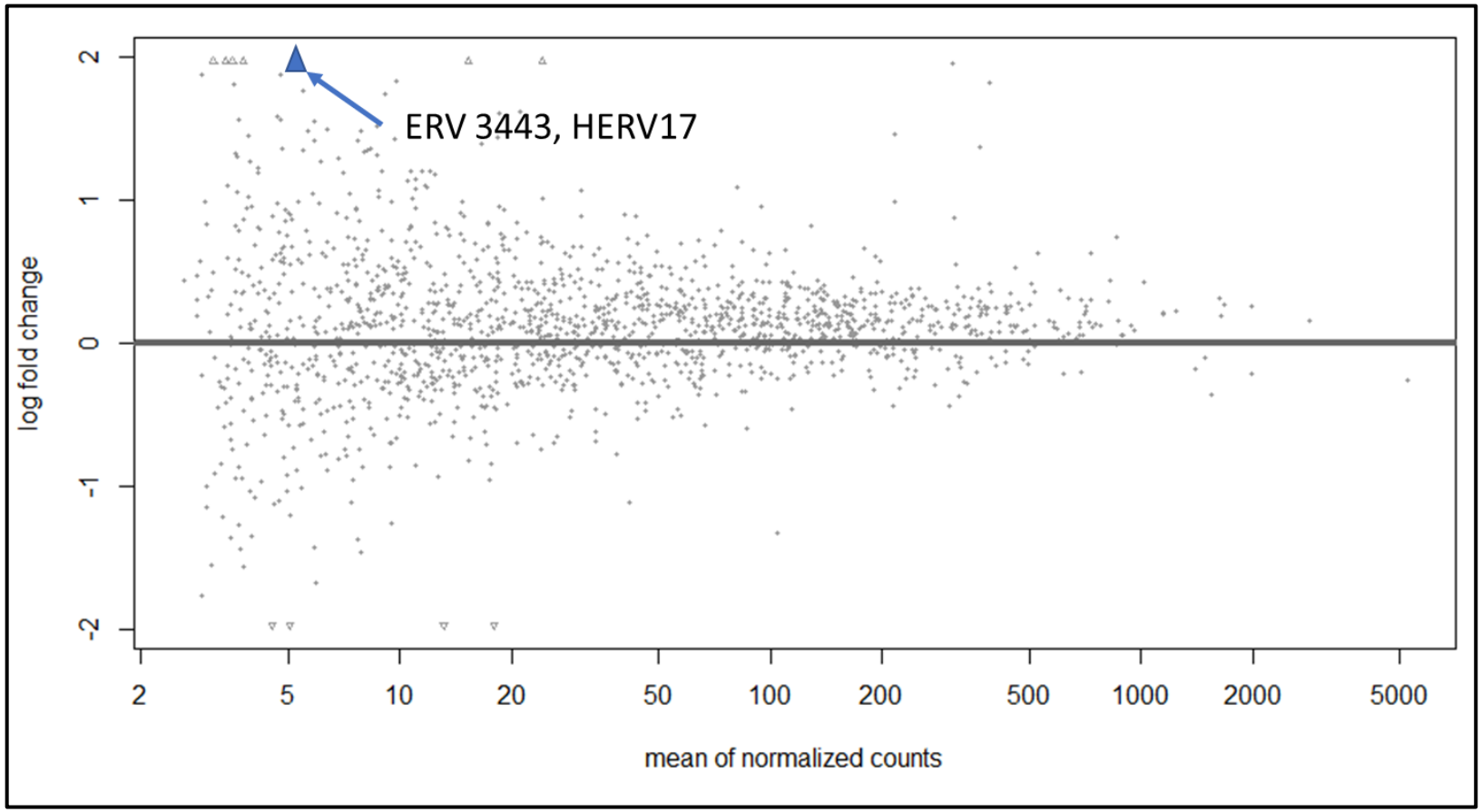


Figure S235. MA Plot of Log2 Fold Changes in Expression between Postmortem Medial Motor Cortex ALS and Non-ALS Control Samples. The MA plot in the figure above shows the distribution of Log2 fold changes against the mean of normalised counts for individual ERVs identified in ALS and non-ALS controls. Statistically significant ERVs (p-value less than 0.01) would be identified by blue points if present. The significant ERV in this graph has been highlighted by increasing the size of the blue arrow which indicates that ERV 3443 is outside of the plot range.

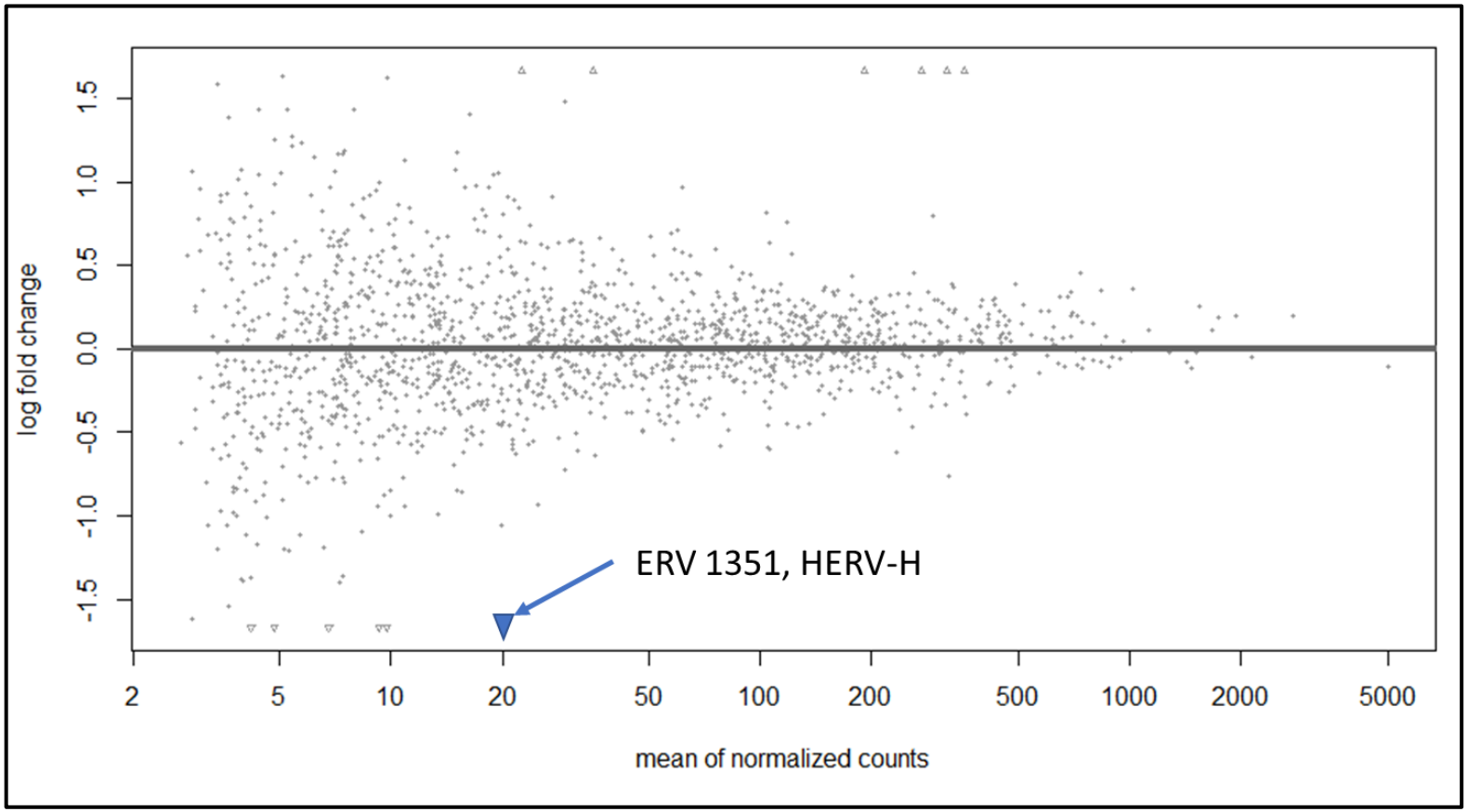


Figure S236. MA Plot of Log2 Fold Changes in Expression between Postmortem Lateral Motor Cortex ALS and Non-ALS Control Samples.
The MA plot in the figure above shows the distribution of Log2 fold changes against the mean of normalised counts for individual ERVs identified in ALS and non-ALS controls. Statistically significant ERVs (p-value less than 0.01) would be identified by blue points if present. The significant ERV in this graph has been highlighted by increasing the size of the blue arrow which indicates that ERV 1351 is outside of the plot range.

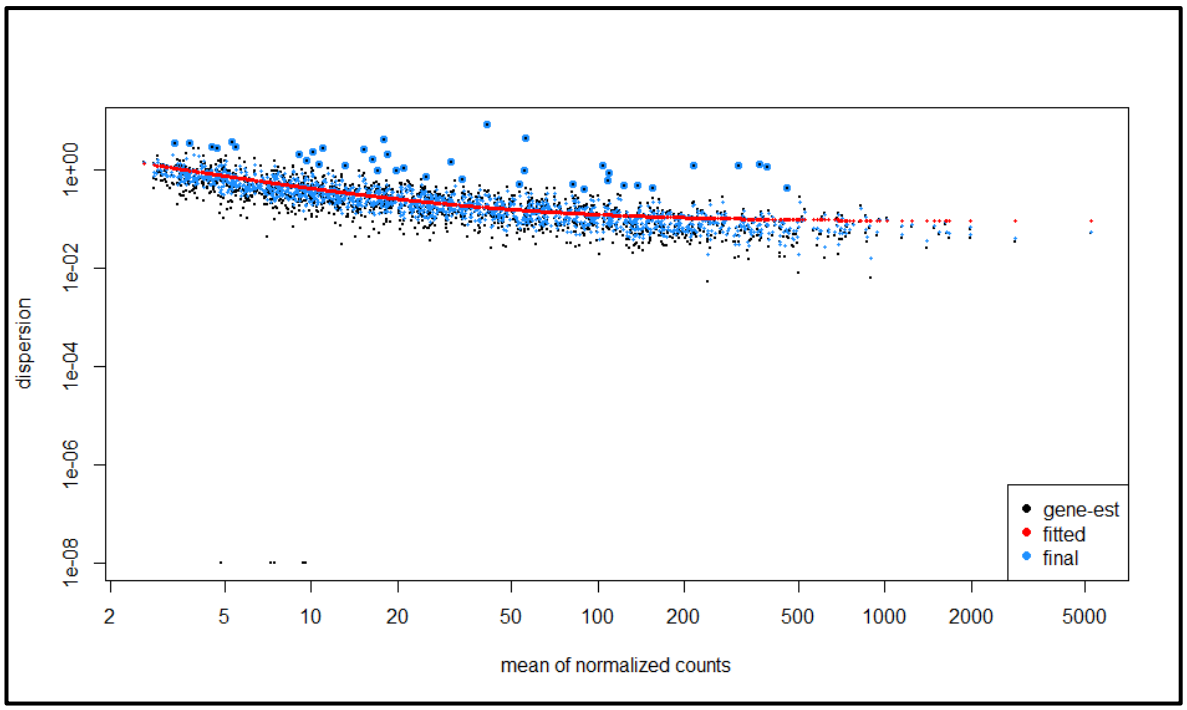


Figure S237. Dispersion Estimate Plot for Differential Expression Data Generated by DESeq2 from Medial Motor Cortex ALS and Non-ALS Controls.

The Figure above plots dispersion estimates for endogenous retrovirus differential expression data between ALS and Non-ALS control tissue from postmortem medial motor cortex samples over the mean count of those ERVs (Love, Huber and Anders, 2014). This is shown as shrinkage towards the red line (consensus) for non-outlier data points. The black points circled in blue above the main data set are defined as dispersion outliers and not shrunk towards the consensus line.

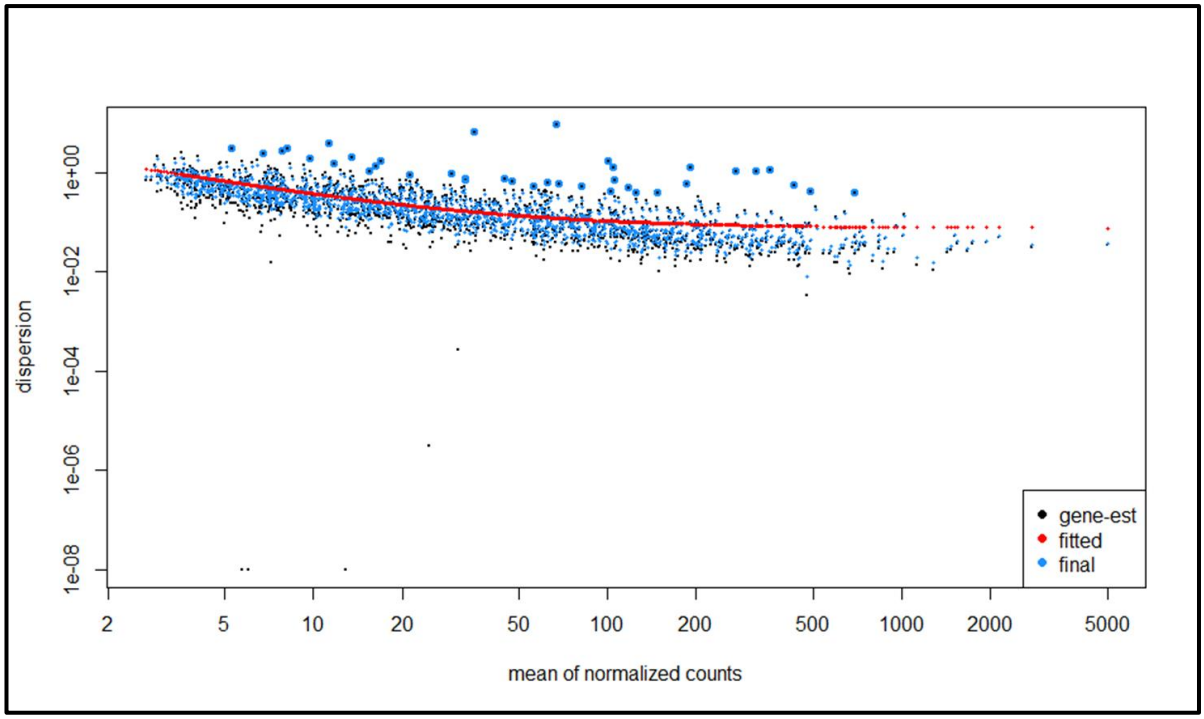


Figure S238. Dispersion Estimate Plot for Differential Expression Data Generated by DESeq2 from Lateral Motor Cortex Cortex ALS and Non-ALS Controls.

The Figure above plots dispersion estimates for endogenous retrovirus differential expression data between ALS and Non-ALS control tissue from postmortem lateral motor cortex samples over the mean count of those ERVs (Love, Huber and Anders, 2014). This is shown as shrinkage towards the red line (consensus) for non-outlier data points. The black points circled in blue above the main data set are defined as dispersion outliers and not shrunk towards the consensus line.

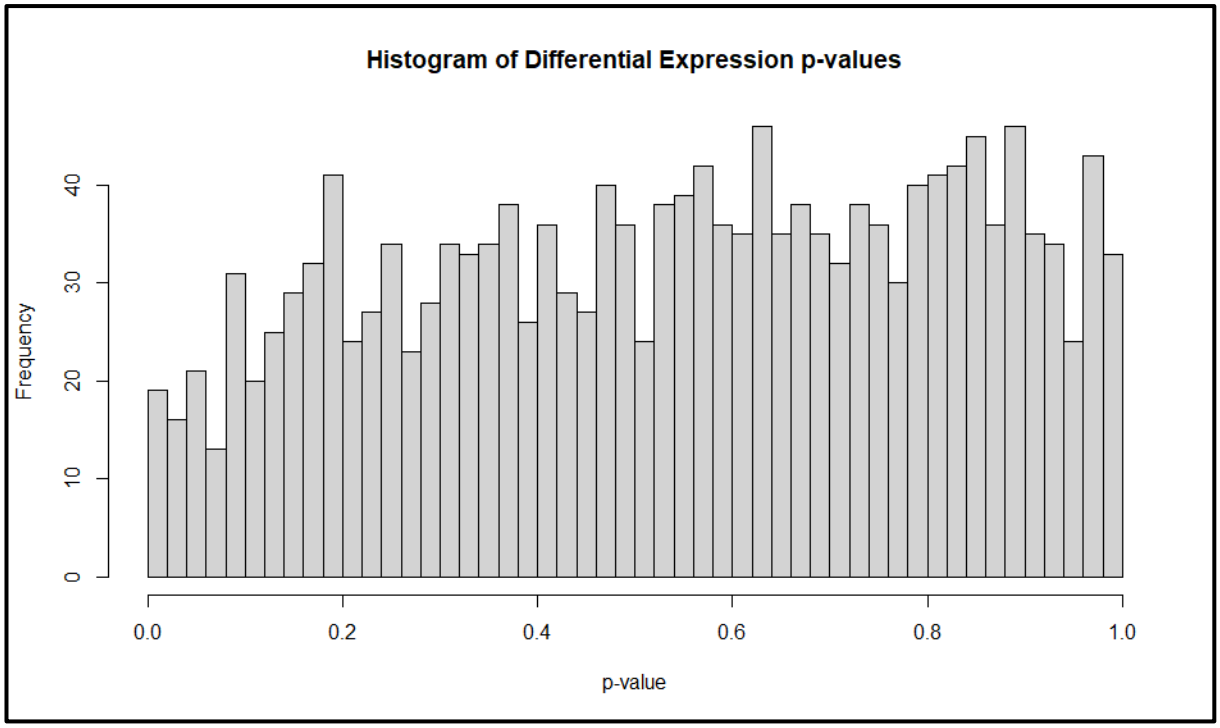


Figure S239. Histogram of p-value Frequency within DESeq2 Differential Expression Analysis of Medial Motor Cortex Tissue Samples.

The figure above displays the frequency at which a given p-value occurs within the ALS vs Non-ALS control differential expression sample set. The p-values, while varied at different values broadly shows a uniform distribution in the sample set. This indicates that the data has no statistically significant differential expression values for ERVs in ALS when compared with controls.

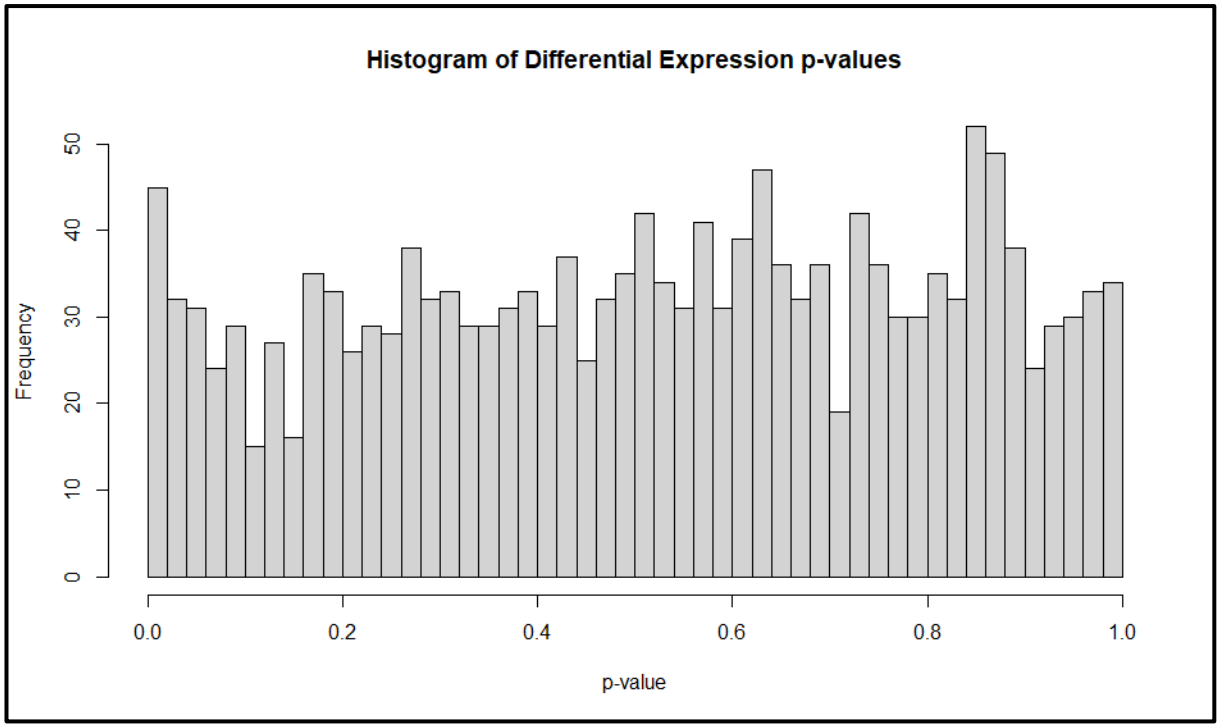


Figure S240. Histogram of p-value Frequency within DESeq2 Differential Expression Analysis of Lateral Motor Cortex Tissue Samples.

The figure above displays the frequency at which a given p-value occurs within the ALS vs Non-ALS control differential expression sample set. The p-values, while varied at different values broadly shows a uniform distribution in the sample set. This indicates that the data has no statistically significant differential expression values for ERVs in ALS when compared with controls.

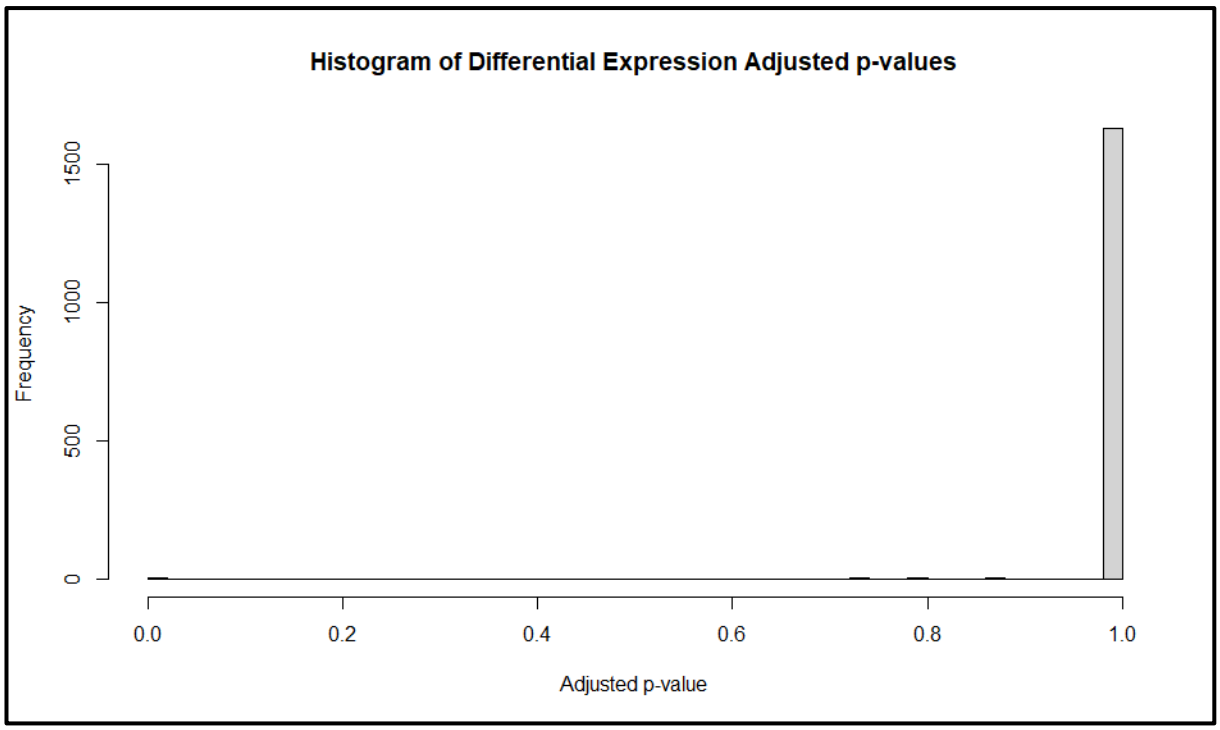


Figure S241. Histogram of Adjusted p-value Frequency within DESeq2 Differential Expression Analysis of Medial Motor Cortex Tissue Samples

The figure above displays the frequency at which a given adjusted p-value occurs within the ALS vs Non-ALS control differential expression sample set.

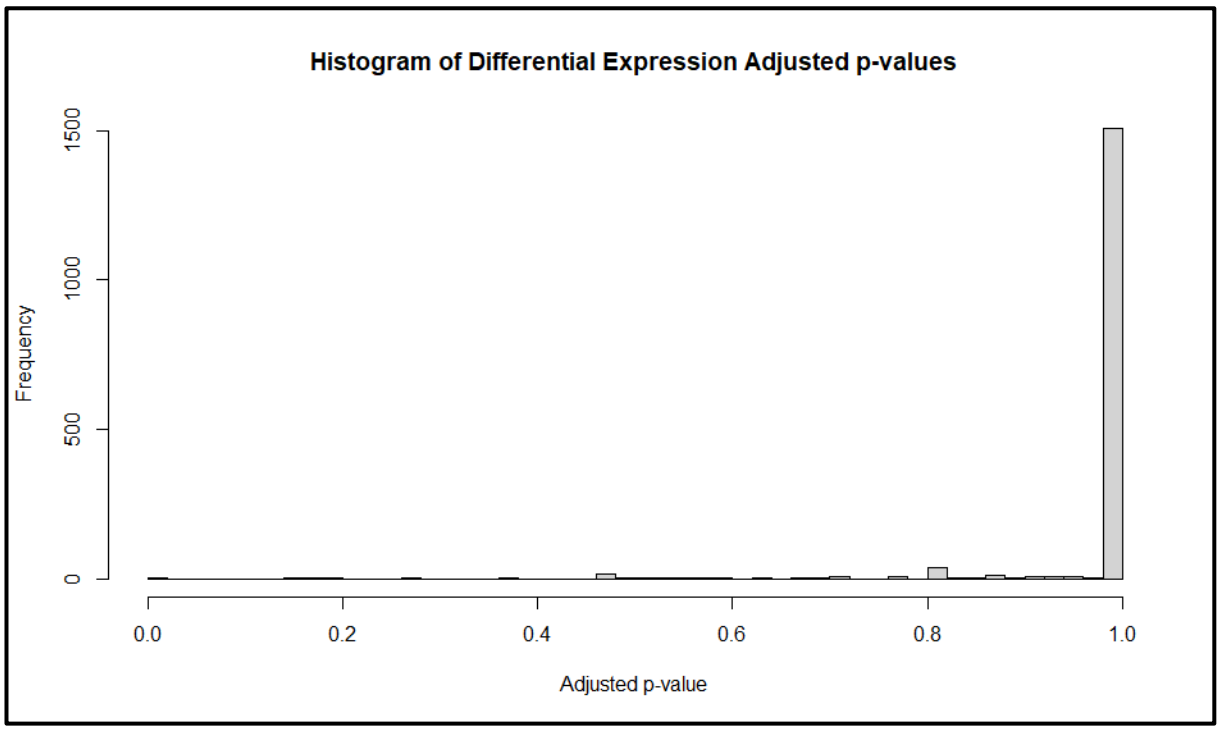


Figure S242. Histogram of Adjusted p-value Frequency within DESeq2 Differential Expression Analysis of Lateral Motor Cortex Tissue Samples

The figure above displays the frequency at which a given adjusted p-value occurs within the ALS vs Non-ALS control differential expression sample set.

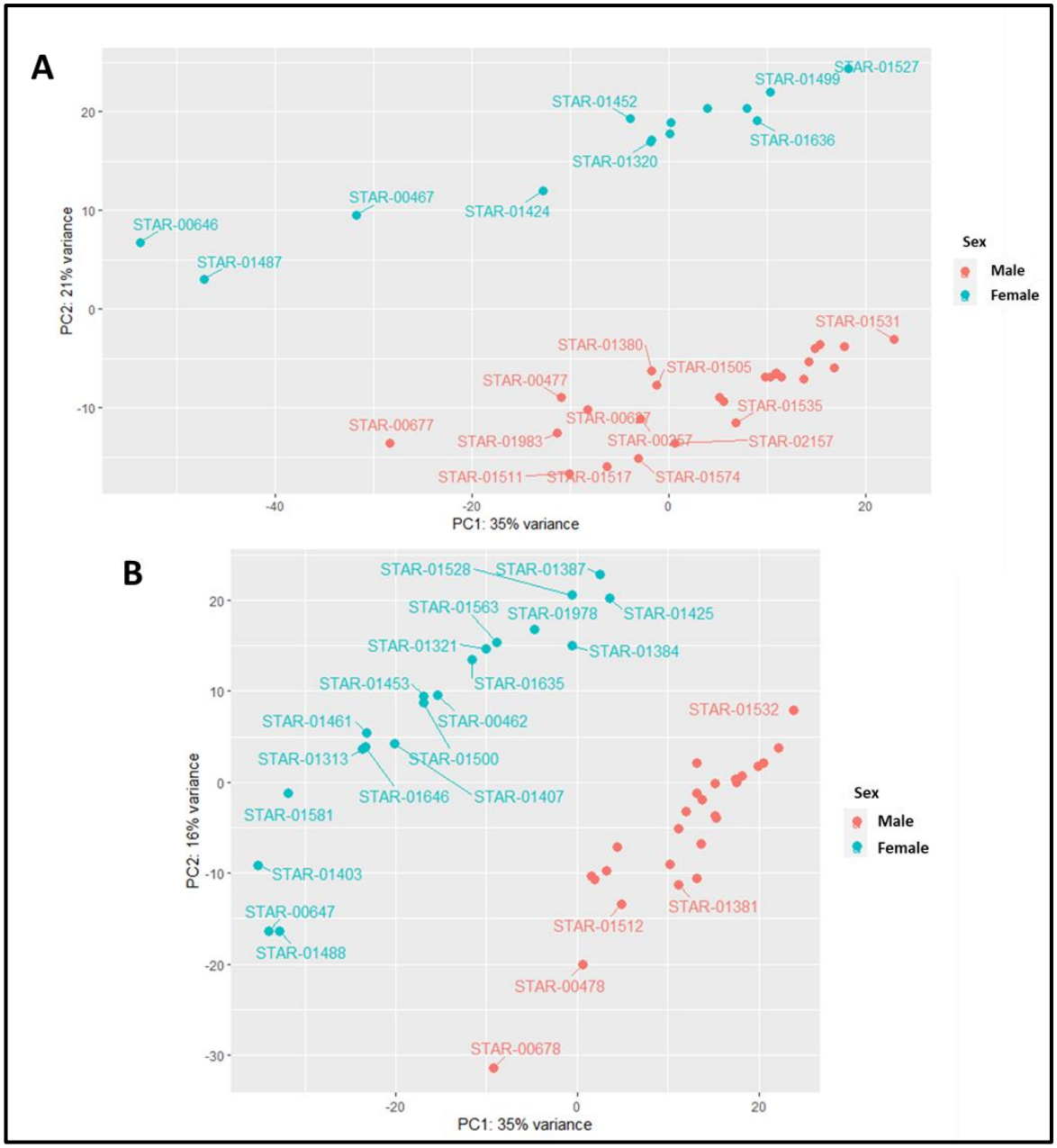


Figure S243. Principal Component Analysis (PCA) plots for DESeq2 Generated Normalised Counts from Differential Expression Analysis of Medial and Lateral Motor Cortex Tissue Showing Difference in Expression Pattern Coloured for Sex of Patient

The principal component analysis (PCA) shown in the figure above plots variation between all ERVs in each sample using normalised counts for each gene per patient sample. These principal component values are then plotted against each other to group samples together based on similar expression profiles of ERVs. In the Figure above A) shows the Male vs Female comparison for Medial Motor Cortex Tissue and B) shows the comparison for Lateral Motor Cortex Tissue.

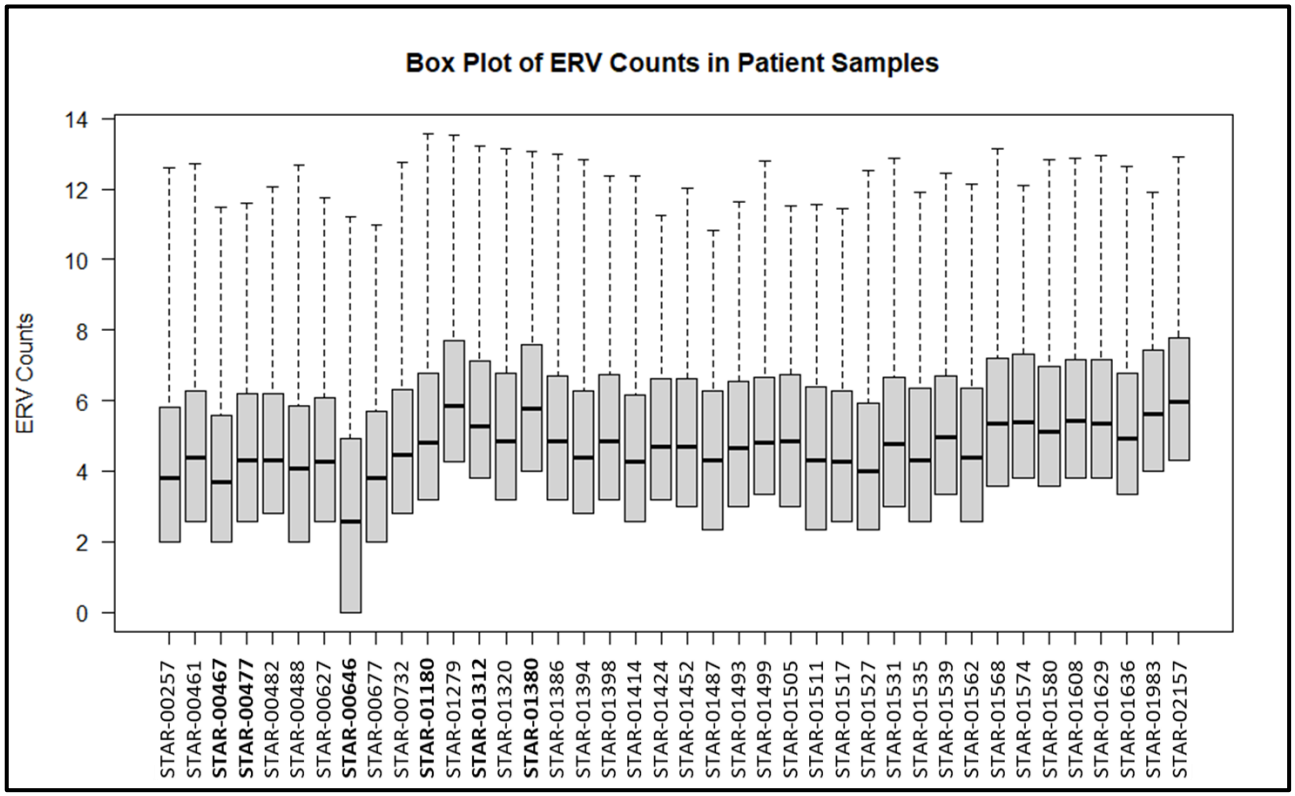


Figure S244. Box Plot of Endogenous Retrovirus Normalised Counts in Medial Motor Cortex Tissue between n=34 ALS and n=6 Non-ALS controls.

The figure above displays statistical information on the counts data for ERVs within the n=34 ALS and n=6 non-ALS control medial motor cortex sample set, inclusive of C9orf72 samples. The thick line inside each of the plots shows the median value for the data while the upper and lower end of the bars represent the interquartile ranges. The “whisker” above the plot, represented by the dotted line, shows the highest count recorded for an individual ERV within the sample. To differentiate the smaller number of control samples in the Box Plot these sample names have been typed in bold.

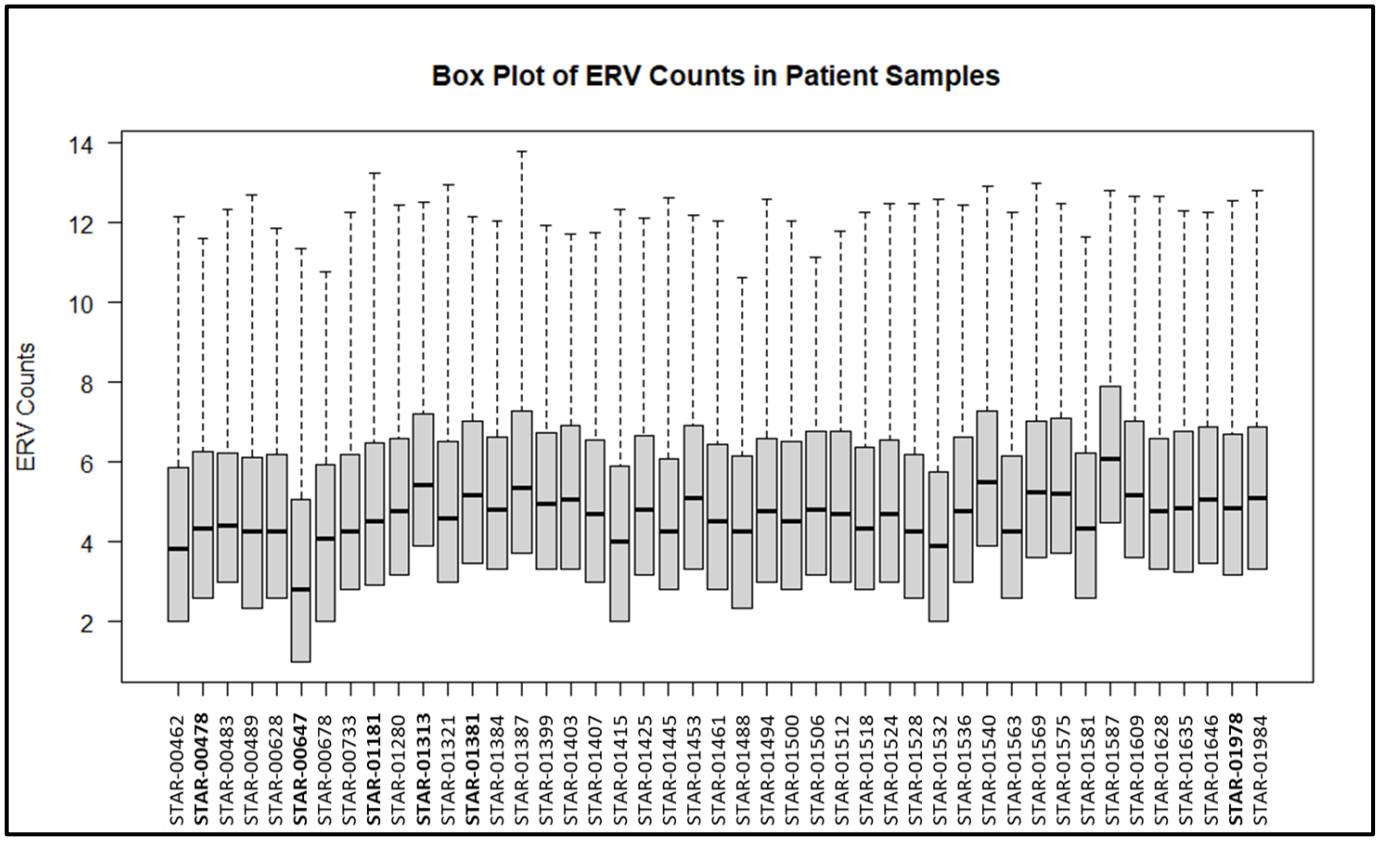


Figure S245. Box Plot of Endogenous Retrovirus Normalised Counts in Lateral Motor Cortex Tissue between n=39 ALS and n=6 Non-ALS controls.

The figure above displays statistical information on the counts data for ERVs within the n=39 ALS and n=6 non-ALS control lateral motor cortex sample set. The thick line inside each of the plots shows the median value for the data while the upper and lower end of the bars represent the interquartile ranges. The “whisker” above the plot, represented by the dotted line, shows the highest count recorded for an individual ERV within the sample. To differentiate the smaller number of control samples in the Box Plot these sample names have been typed in bold.

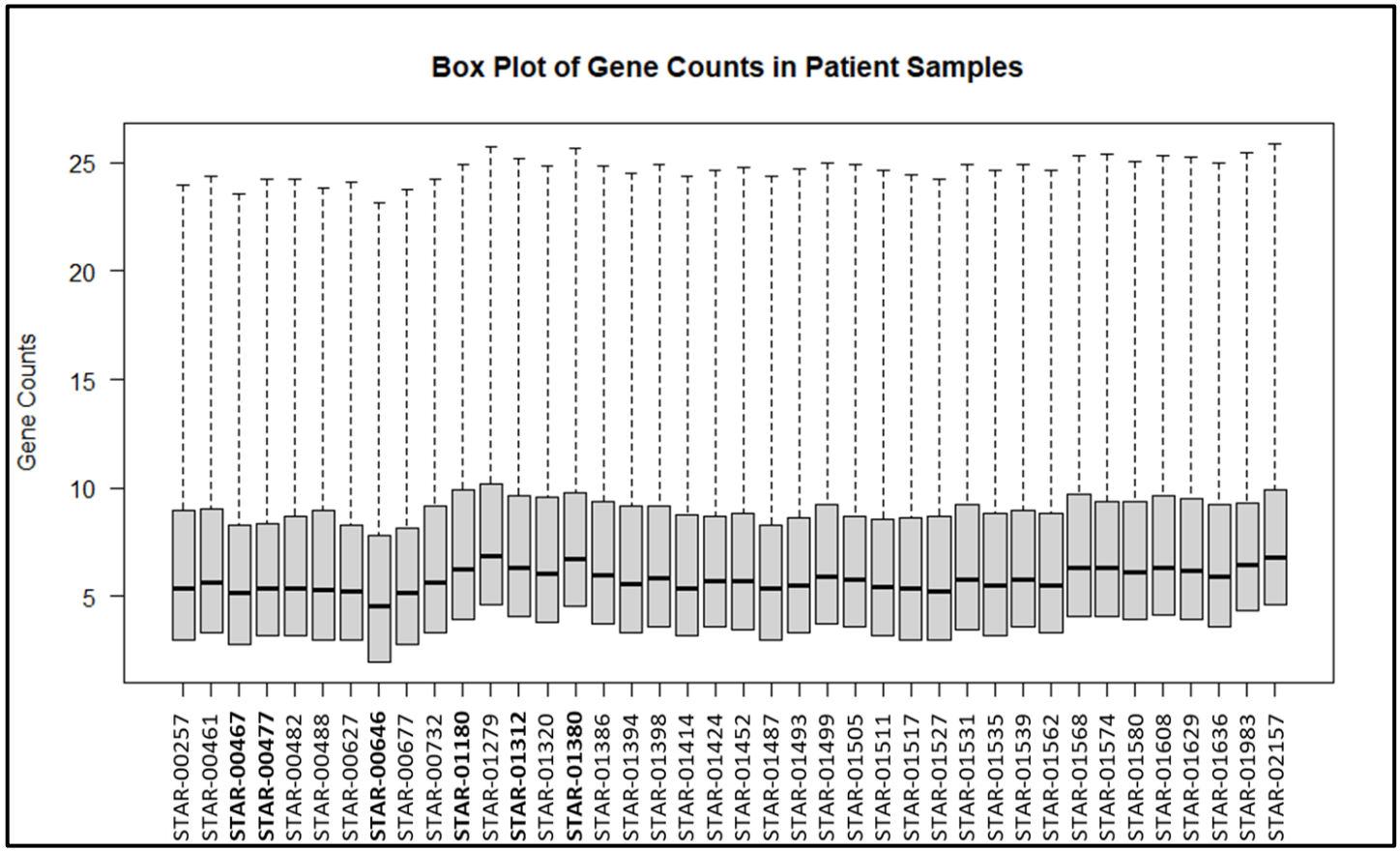


Figure S246. Box Plot of Normalised Gene Counts in Medial Motor Cortex Tissue between n=34 ALS and n=6 Non-ALS controls.

The figure above displays statistical information on the counts data for genes within the n=34 ALS and n=6 non-ALS control medial motor cortex sample set. The thick line inside each of the plots shows the median value for the data while the upper and lower end of the bars represent the interquartile ranges. The “whisker” above the plot, represented by the dotted line, shows the highest count recorded for an individual ERV within the sample. To differentiate the smaller number of control samples in the Box Plot these sample names have been typed in bold.

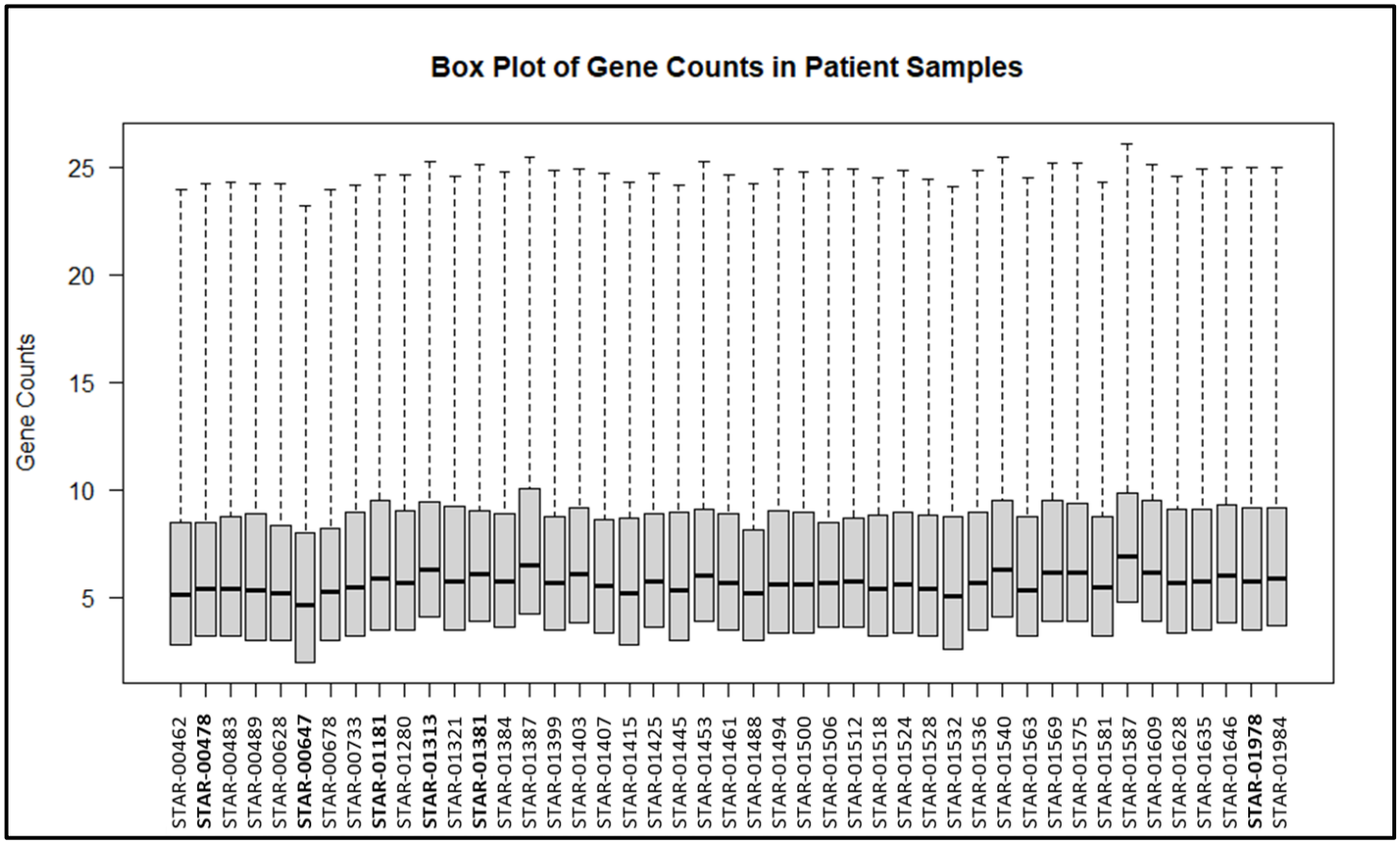


Figure S247. Box Plot of Normalised Gene Counts in Lateral Cortex Tissue between n=39 ALS and n=6 Non-ALS controls.

The figure above displays statistical information on the counts data for genes within the n=39 ALS and n=6 non-ALS control lateral motor cortex sample set. The thick line inside each of the plots shows the median value for the data while the upper and lower end of the bars represent the interquartile ranges. The “whisker” above the plot, represented by the dotted line, shows the highest count recorded for an individual ERV within the sample. To differentiate the smaller number of control samples in the Box Plot these sample names have been typed in bold.

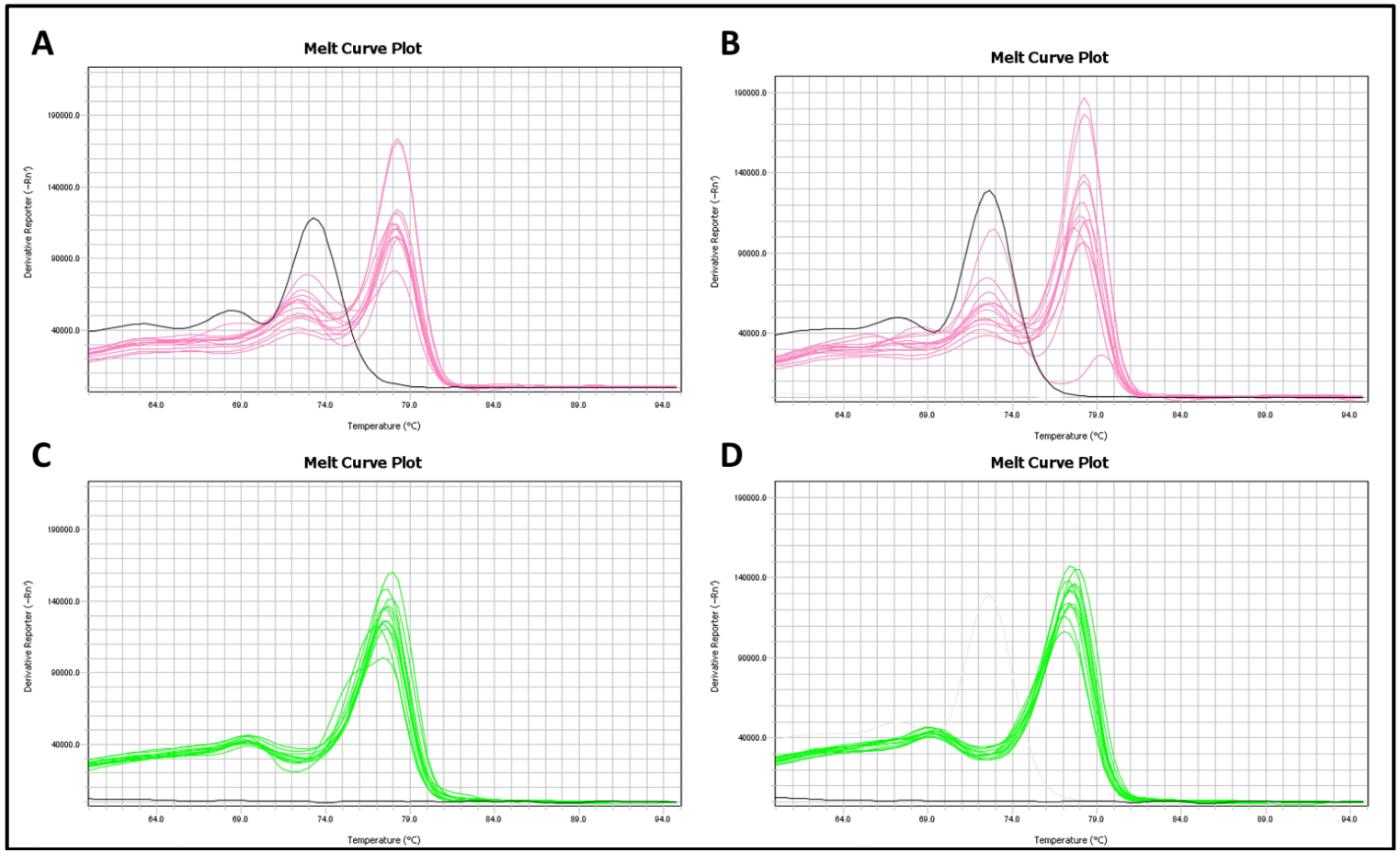


Figure S248. Melt Curve Plots for HERV-H *env* and HERV-K22 *pol* Primer Efficiency Experiments.

The figure above shows the melt curve plots for primer efficiency experiments for A&B) HERV-H *env* and C&D) HERV-K22 *pol* gene targets. These primer efficiency assays were conducted on A&C) ALS sample A151/10 and B&D) non-ALS control sample A292/09.

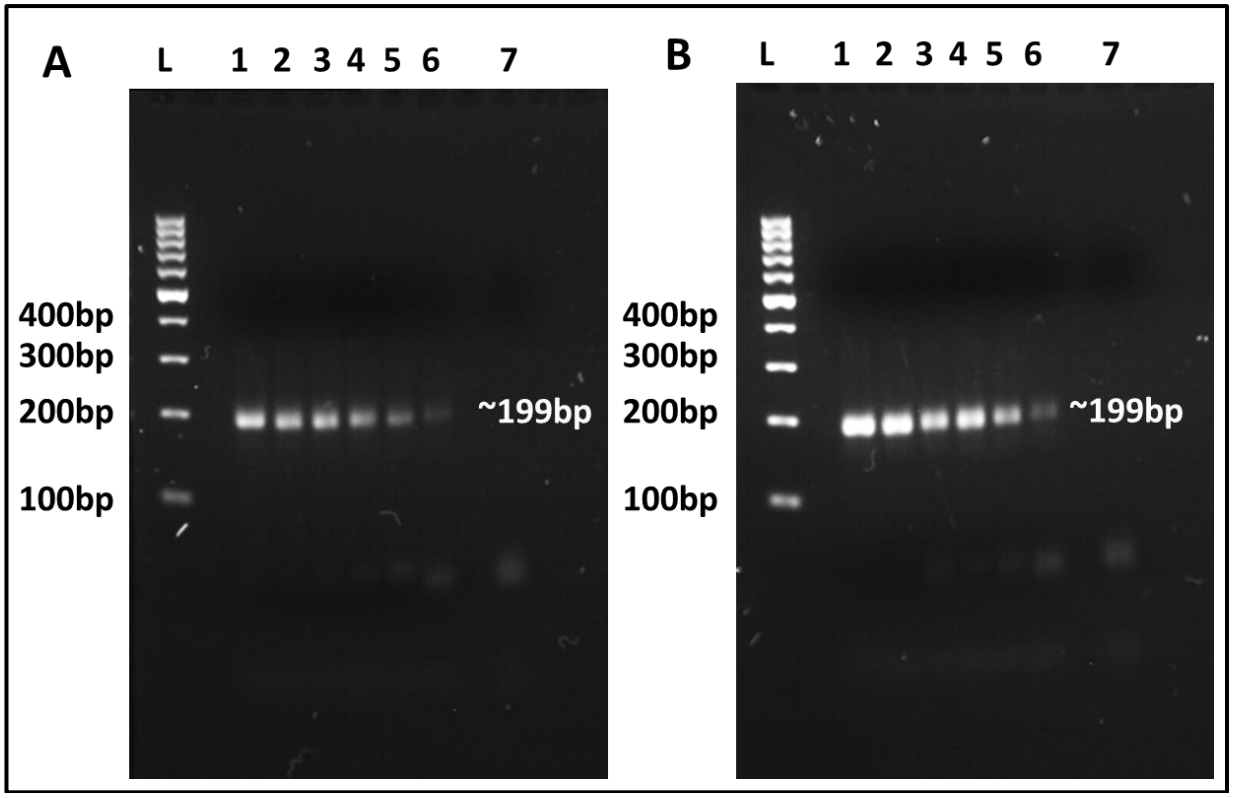


Figure S249. 2% Gel Electrophoresis Output for HERV-H *env* Primer Efficiency Assay
The figure above shows the 2% gel electrophoresis images for the amplicons produced by the primer efficiency experiments performed on ALS Sample A151/10 (B) and non-ALS control sample A292/09 (A). The numbers at the top of the images are for L) 100bp ladder, 1) undiluted cDNA, 2) cDNA at 1/2 dilution, 2) cDNA at 1/4 dilution, 3) cDNA at 1/8 dilution, 4) cDNA at 1/16 5) cDNA at 1/32 dilution, 6) cDNA at 1/64 and 7) Assay Water Control.

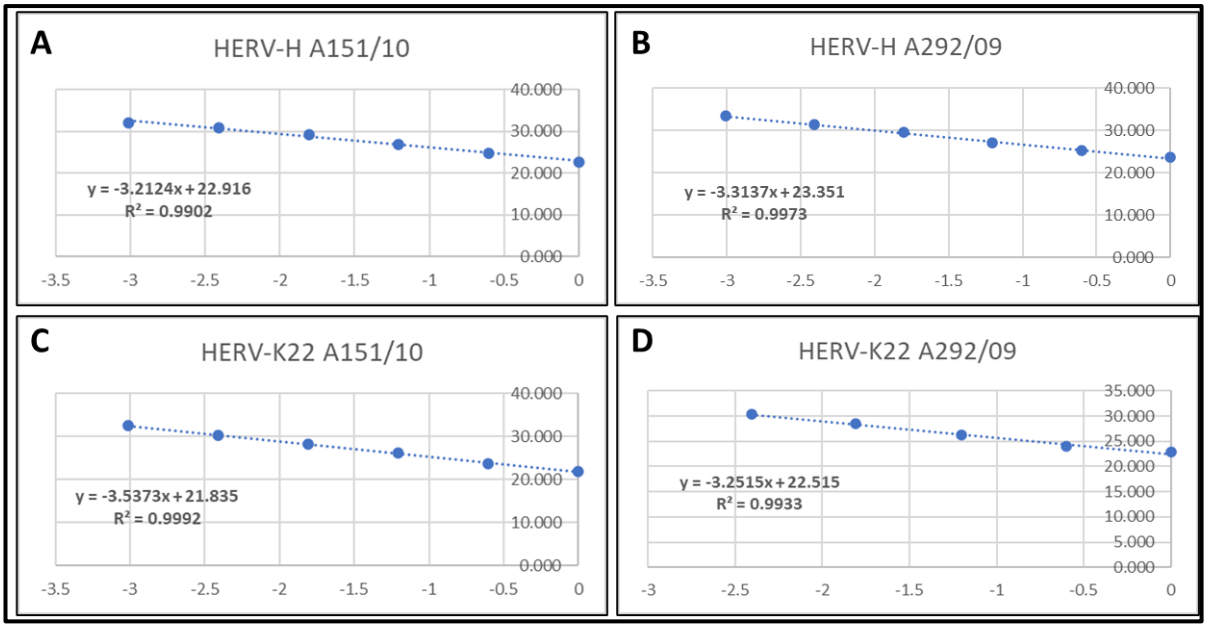


Figure S250. Amplification Efficiency Graphs for HERV-H *env* and HERV-K2 *pol* Primer Targets.

The figure above displays standard curve graphs for ALS Patient Sample A151/10 and non-ALS control sample A292/09. Primer targets for the above image are A&B) HERV-H *env* and C&D) HERV-K22 *pol*. The axes for the graphs display Ct values on the y-axis plotted against log transformed dilution factors performed on the cDNA on the x-axis.

Supplementary Table S19. Summary of Amplification Efficiency Data for HERV-K22 *pol* and HERV-H *env* primers tested on ALS and non-ALS Patient Sample.

The table below shows primer efficiency data obtained from Standard curves generated from cDNA amplification efficiency graphs shown in Figure 6.58. Efficiency Percentages were generated from the equation E = 10(-1/slope)*100.

Primer Target (Sample ID)	Slope	R^2	Efficiency
HERV-K22 *pol* (A151/10)	-3.537	0.9992	91.74%
HERV-K22 *pol* (A292/09)	-3.252	0.9933	103.03%
HERV-H *env* (A151/10)	-3.212	0.9902	104.78%
HERV-H *env* (A292/09)	-3.314	0.9973	100.34%

Supplementary Table S20. Sanger Sequencing of Target Amplicon for HERV-K22 and HERV-H Primer Sets with BLASTn Closest Species Match

The table below displays the sequences obtained for the HERV-K22 *pol* and HERV-H *env* primer targets. Sequence similarity was obtained from NCBI BLASTn web service along with the closest human match to the sequence if any.

Primer Target	Sequence Obtained From GATC Sequencing	NCBI Reference Sequence and Accession Number of Closest Match	Sequence Coverage
HERV-H *env* A292	TCTTTCTCATACACCATGAAAATCGAACCTCCCCCTC TACGCAGTTAGCCCCATCAGTCCCCATTACAACCTC TGACGGCTGCAGC	AH007766.2 Homo sapiens human endogenous retrovirus HERV-H18 5' LTR and leader region, partial sequence	79/81(98%)
HERV-H *env* A151	GATACTCAACGTTTTCTCATACACCATGAAAATCGAA CCTCCCCCTCTACGCAGTTACCCCATCAGTCCCCATT A CAACCTCTGACGGCTGATG	AH007766.2 Homo sapiens 5' long terminal repeat LTR repeat region; pol pseudogene,	90/91 (99%)
HERV-K22 *pol* A292	TTCCAAAGGTCTGGGAAATGGGGACTTT	AK098492.1 Homo sapiens cDNA FLJ25626 fis, clone STM03094	24/24 (100%)
HERV-K22 *pol* A151	TATAAGTGTGCAGTTTGCAACCTATAGGGGCCCCTC T CAAAA	No Significant Similarity to Sequence	

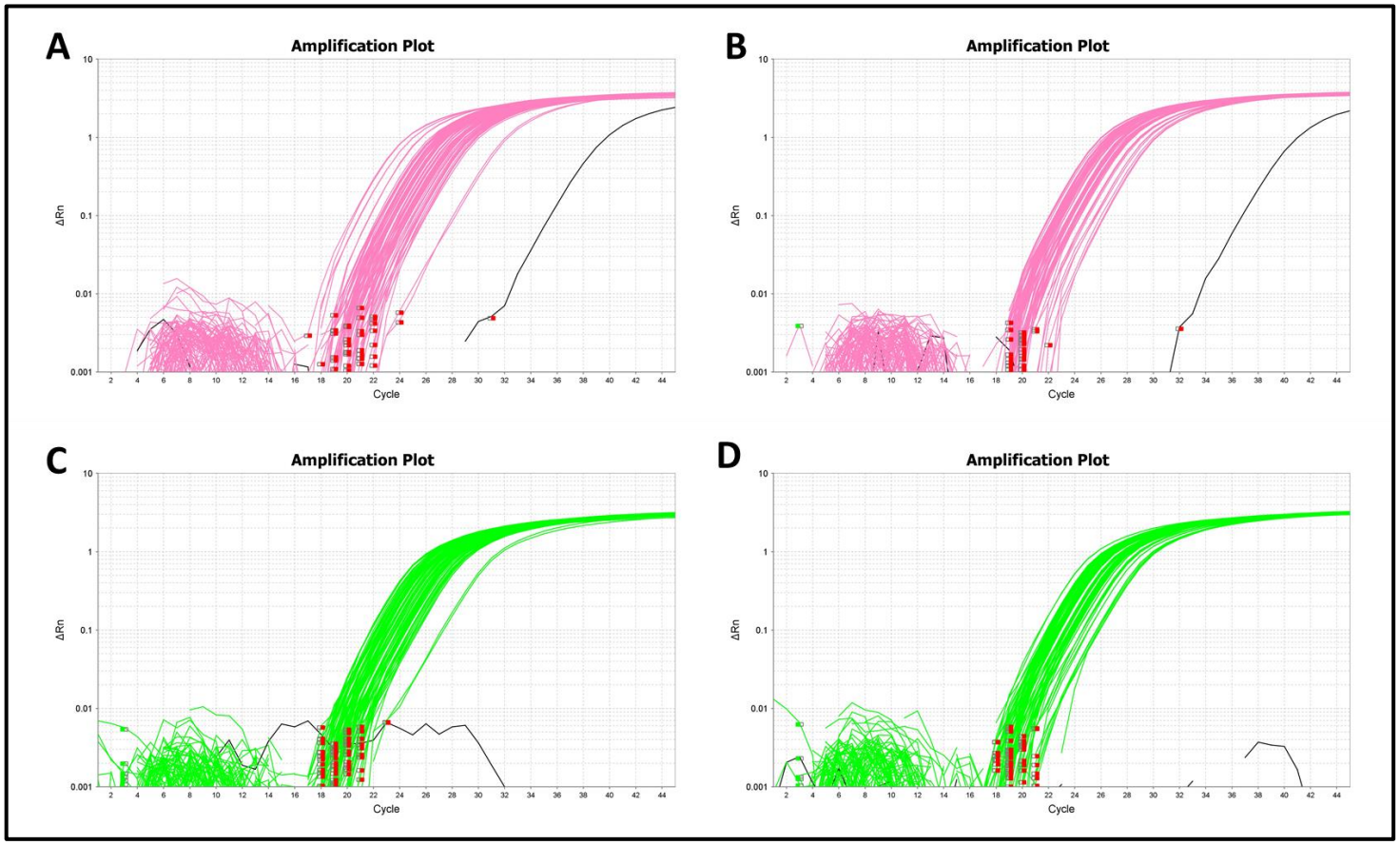


Figure S251. Amplification Plots for HERV-H *env* and HERV-K22 *pol* Gene Targets on n=54 ALS and n=37 non-ALS Control Postmortem Premotor Cortex Brain Tissue Samples.

The figure above shows the amplification plots for A&B) HERV-H *env* and C&D) HERV-K22 *pol* gene targets. The samples are split between two assays with samples 1-46 on Assay 1 (A&C) and 47-91 on Assay 2 (B&D). The black lines in each image represent water control reactions for the RT-qPCR assay. The baseline for measuring the Ct value for the samples in each gene target was ΔRn = 0.21.

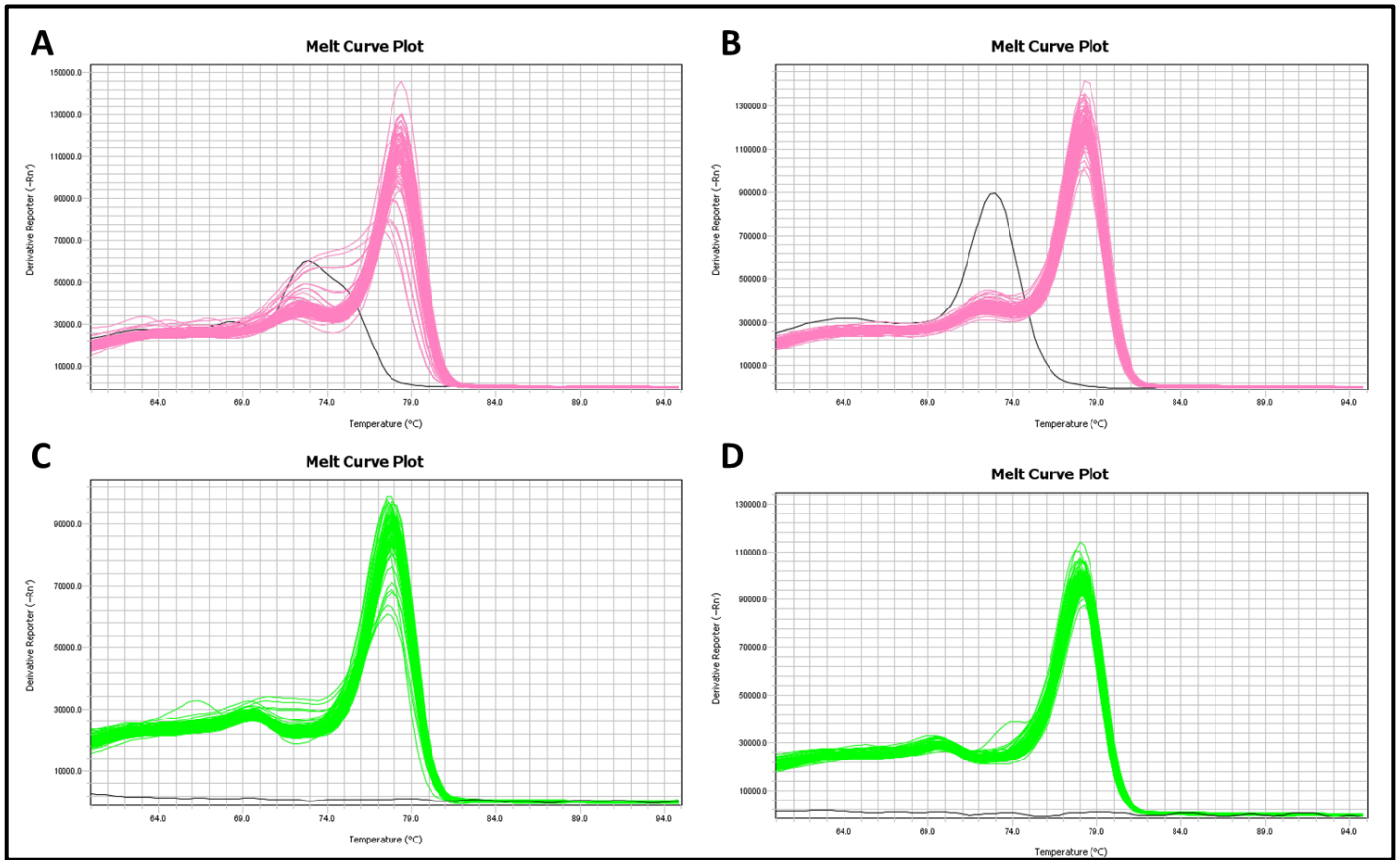


Figure S252. Melt Curve Plots for HERV-H *env* and HERV-K22 pol Gene Targets on n=54 ALS and n=37 non-ALS Control Postmortem Premotor Cortex Brain Tissue Samples.

The figure above shows the melt curve plots for A&B) HERV-H env and C&D) HERV-K22 pol gene targets. The samples are split between two assays with samples 1-46 on Assay 1 (A&C) and 47-91 on Assay 2 (B&D). The black lines in each image represent water control reactions for the RT-qPCR assay.

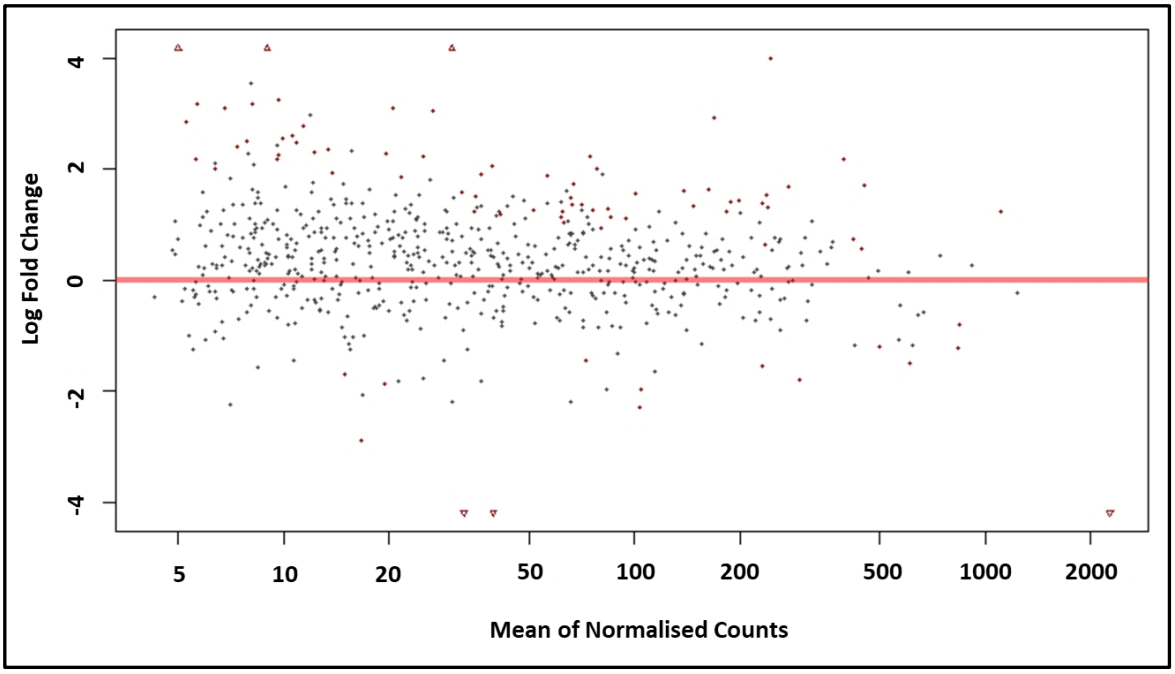


Figure S253 MA Plot of Log2 Fold Changes in Expression between Peripheral Blood Mononuclear Cells in ALS and Non-ALS Controls.

The MA plot in the figure above shows the distribution of Log2 fold changes against the mean of normalised counts for individual ERVs identified in ALS and non-ALS controls. Statistically significant ERVs (p-value less than 0.01) would be identified by red points if present. While there are more red points than appear in Table 7.1 this is due to these additional points being outside of the adjusted p-value cut-off of $q<0.05$.

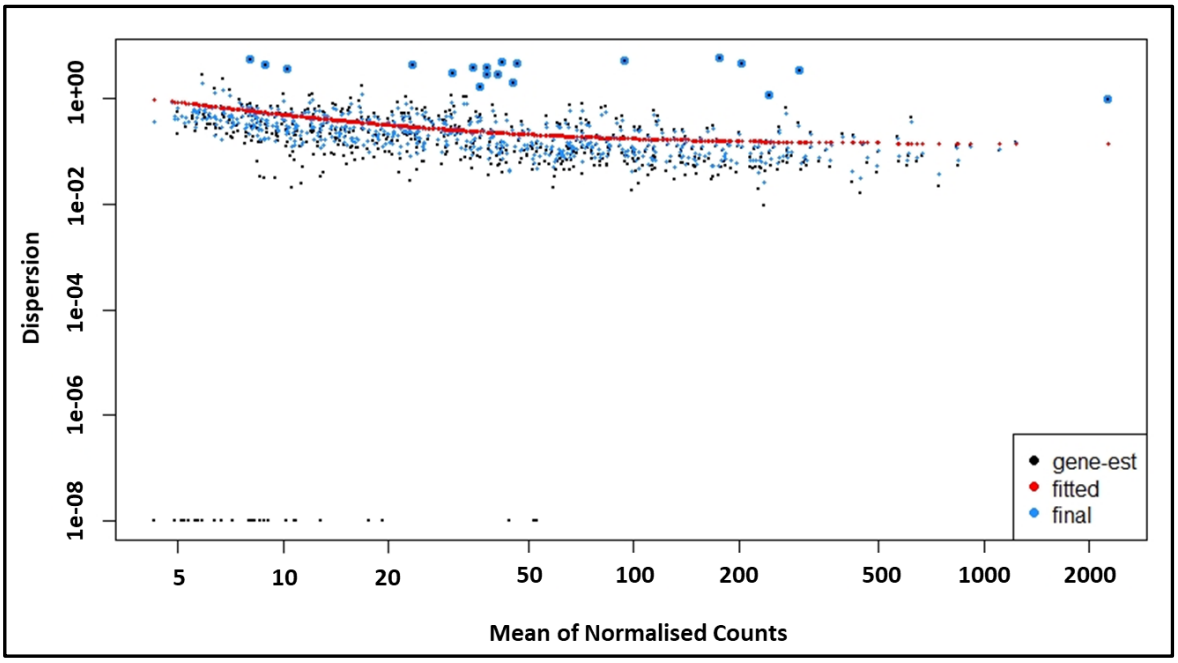


Figure S254. Dispersion Estimate Plot for Differential Expression Data Generated by DESeq2 from Peripheral Blood Mononuclear Cells taken from ALS and Non-ALS Control samples.

The Figure above plots dispersion estimates for endogenous retrovirus differential expression data between ALS and Non-ALS control tissue from Peripheral Blood Mononuclear Cells samples over the mean of normalised counts, relating to how high these are expressed, of those ERVs (Love, Huber and Anders, 2014). This is shown as shrinkage towards the red line (consensus) for non-outlier data points. The black points circled in blue above the main data set are defined as dispersion outliers and not shrunk towards the consensus line.

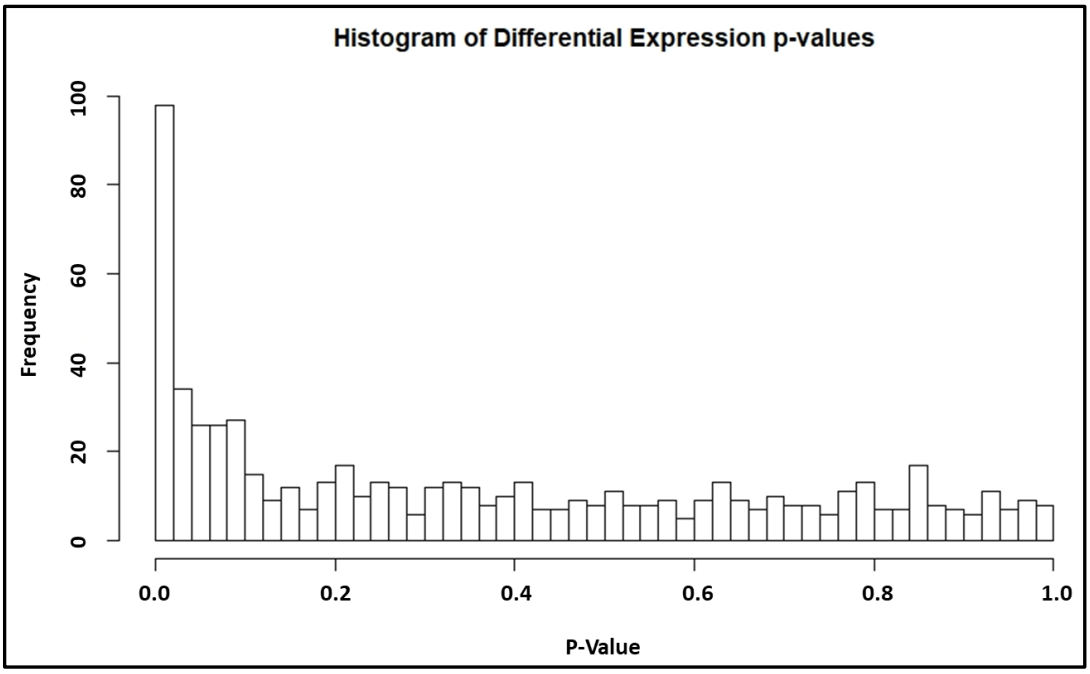


Figure S255. Histogram of P-Value Frequency within DESeq2 Differential Expression Analysis

The figure above displays the frequency at which a given p-value occurs within the ALS vs Non-ALS control differential expression sample set. The p-value distribution shows a definitive conservative distribution (weighted towards p=0.0).

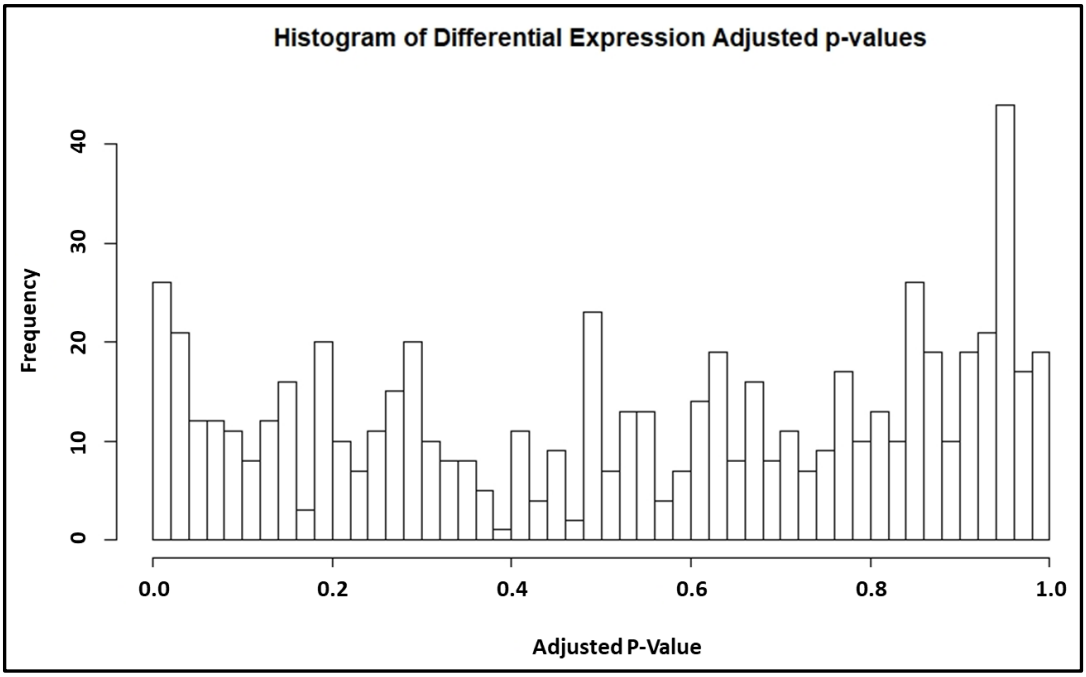


Figure S256. Histogram of Adjusted P-Value Frequency within DESeq2 Differential Expression Analysis

The figure above displays the frequency at which a given adjusted p-value occurs within the ALS vs Non-ALS control differential expression sample set. The adjusted p-values show a largely uniform distribution, despite the clear peak in the data at 1.0.

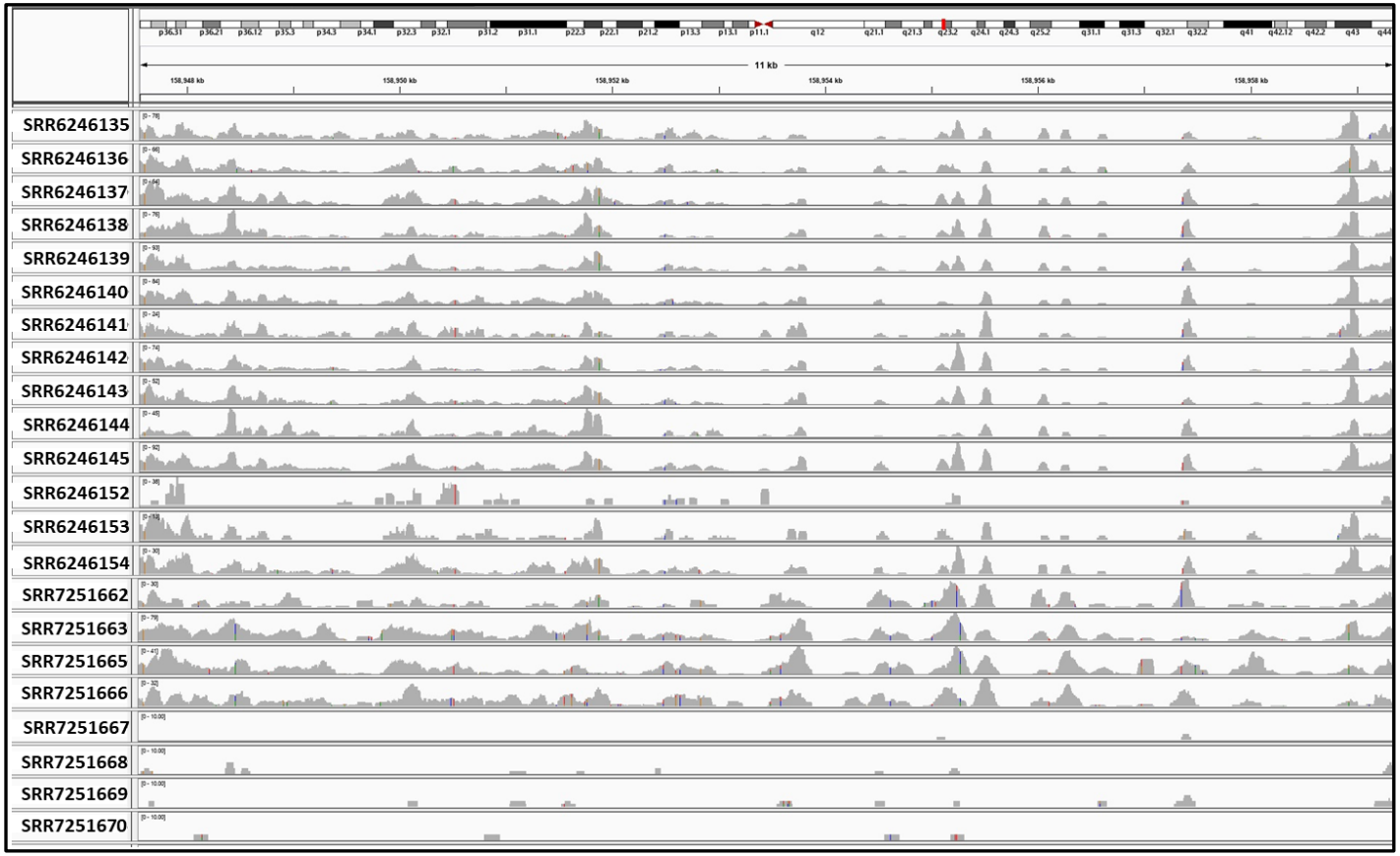


Figure S257. Read Alignment Coverage for ERVMap 6078 (HERV-K22)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 6078 (Chromosome 1, locus q23.1), identified as HERV-K22. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 1 with the red bar indicating the ERV location within the locus SRR6246135- SRR6246145 and SRR7251662- SRR7251666 are ALS and the rest of the samples are non-ALS Controls.

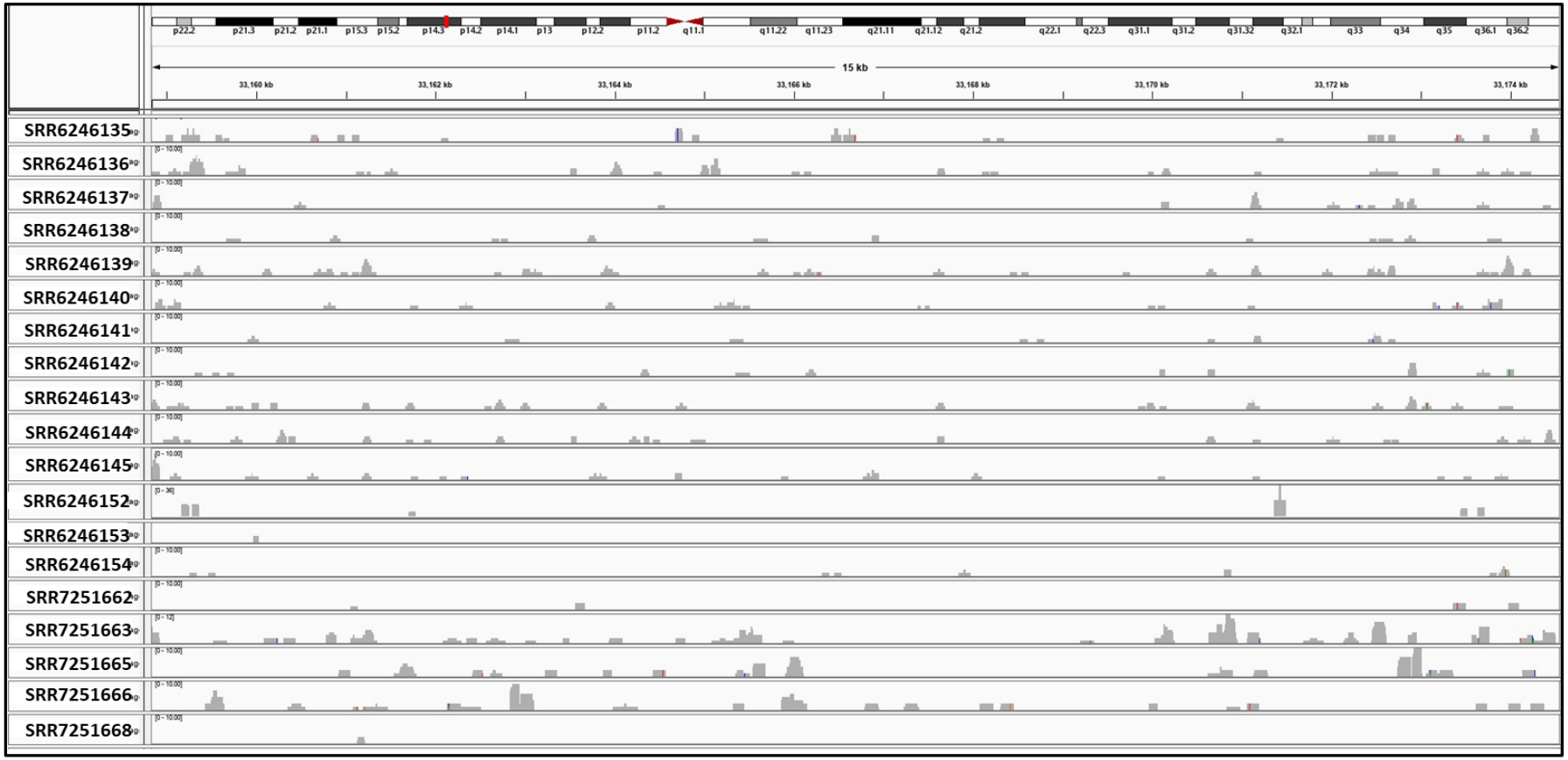


Figure S258. Read Alignment Coverage for ERVMap 2458 (MER57A)

The figure above shows the RNA sequencing reads alignment to the genomic region associated with ERVMap 2458 (Chromosome 7, locus p14.3), identified as MER57A. This information has been visualised using the Integrated Genome Viewer (IGV) program with the bar above the sample coverage information shows the locus layout of chromosome 7 with the red bar indicating the ERV location within the locus. SRR6246135- SRR6246145 and SRR7251662- SRR7251666 are ALS and the rest of the samples are non-ALS Controls.

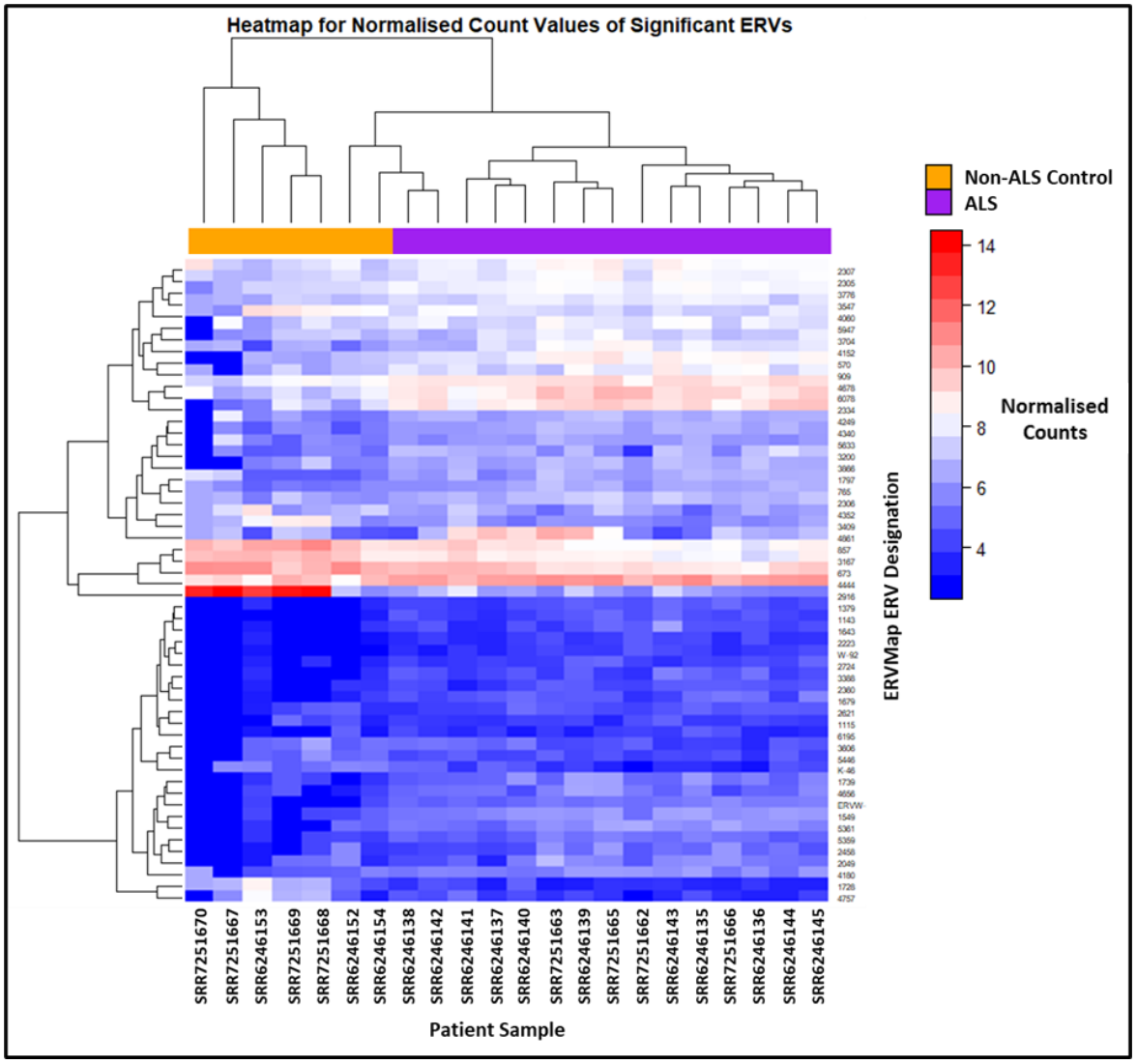


Figure S259. Heatmap of Normalised counts for Statistically Significant ERVs in Peripheral Blood Mononuclear Cells from n-15 ALS and n=7 Non-ALS Control Samples.

The heatmap displayed in the figure above shows the normalised counts data for ERVs identified by the ERVmap.bed file with low expressed ERV members filtered out. The rows and columns are hierarchically clustered to group together samples and ERVs with similar expression profiles based on normalised counts data generated from DESeq2. Also included in the cells above the counts matrix identifies those samples which are from the ALS (purple) and non-ALS control (orange) sample sets.

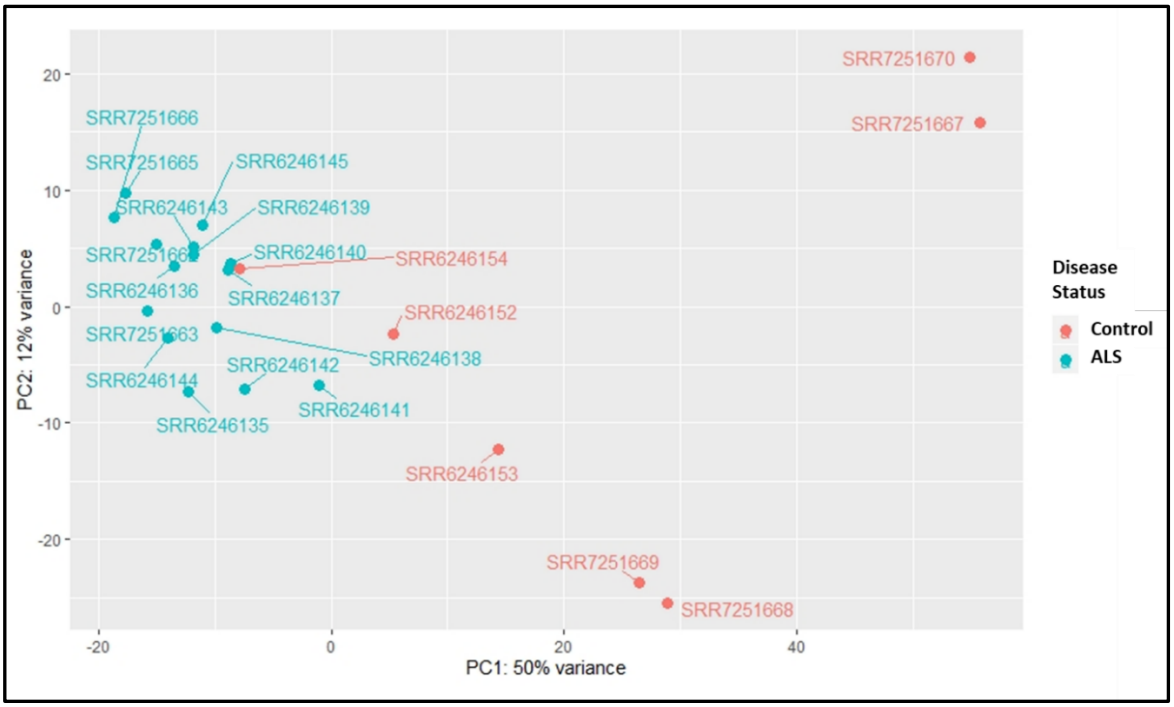


Figure S260. Principle Component Analysis (PCA) plots for DESeq2 Generated Normalised Counts from Differential Expression Analysis of Peripheral Blood Mononuclear Cells from ALS and Non-ALS Controls.

The principal component analysis (PCA) shown in the figure above plots variation between all ERVs in each sample using normalised counts for each gene per patient sample. These principal component values are then plotted against each other to group samples together based on similar expression profiles of ERVs. This PCA plot shows the distribution of ALS and Control Samples from the patient group.

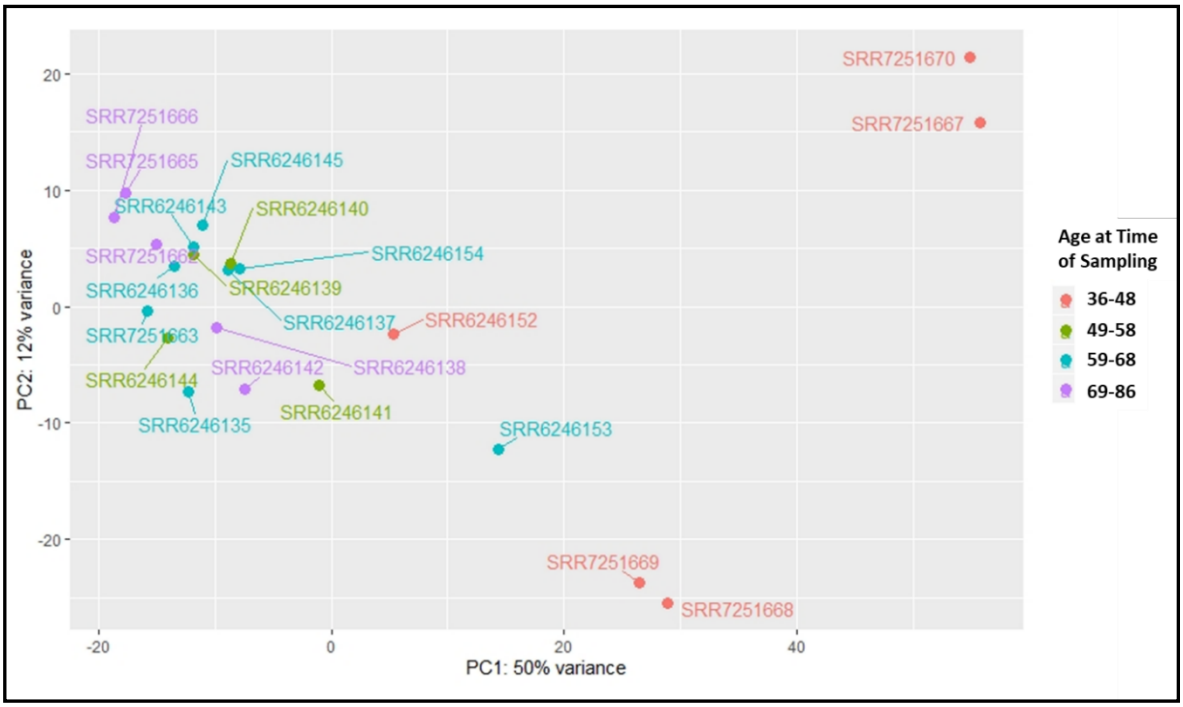


Figure S270. PCA Plot of ALS and non-ALS Control Peripheral Blood Mononuclear Cell Samples Coloured by Patient Age at time of Sampling.

The principal component analysis (PCA) shown in the figure above plots variation between all ERVs in each sample using normalised counts for each gene per patient sample. These principal component values are then plotted against each other to group samples together based on similar expression profiles of ERVs.

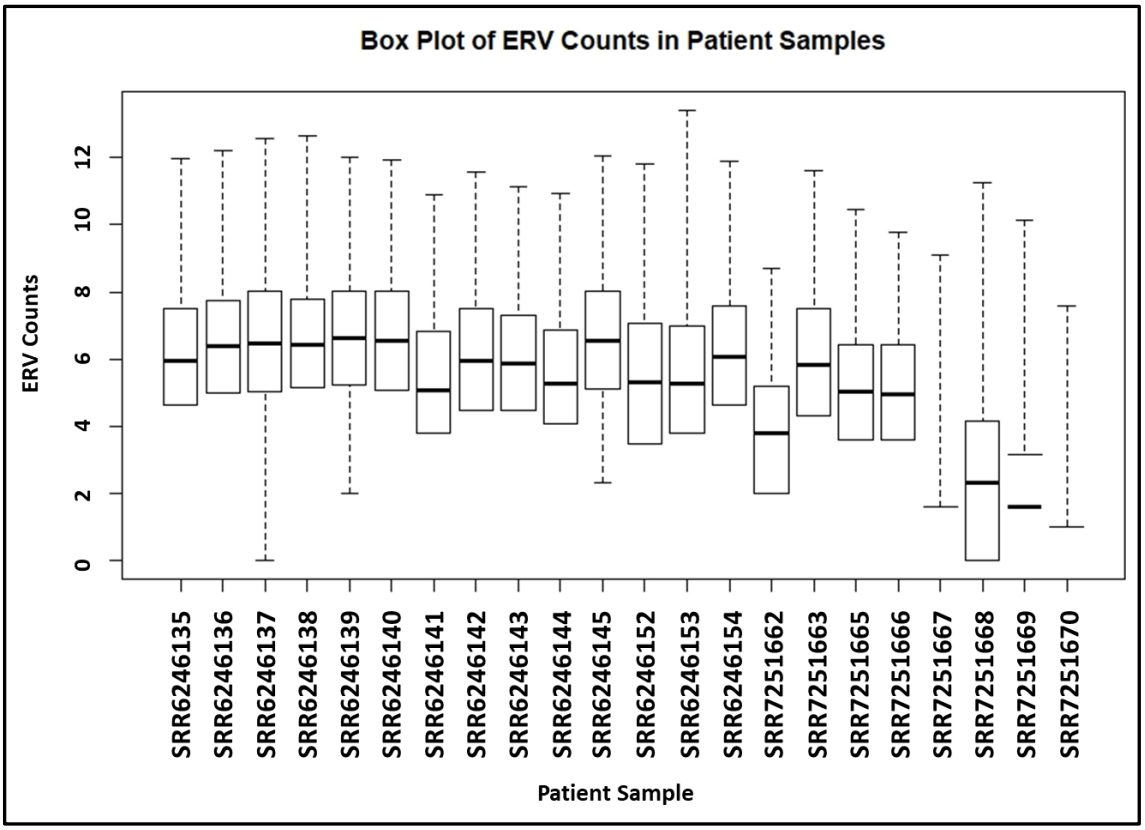


Figure S271. Box Plot of Endogenous Retrovirus Normalised Counts between n=15 ALS and n=7 Non-ALS controls.

The figure above displays statistical information on the counts data for ERVs within the n=15 ALS and n=7 non-ALS control sample set. The thick line inside each of the plots shows the median value for the date while the upper and lower end of the bars represent the interquartile ranges. The “whisker” above the plot, represented by the dotted line, shows the highest count recorded for an individual ERV within the sample.

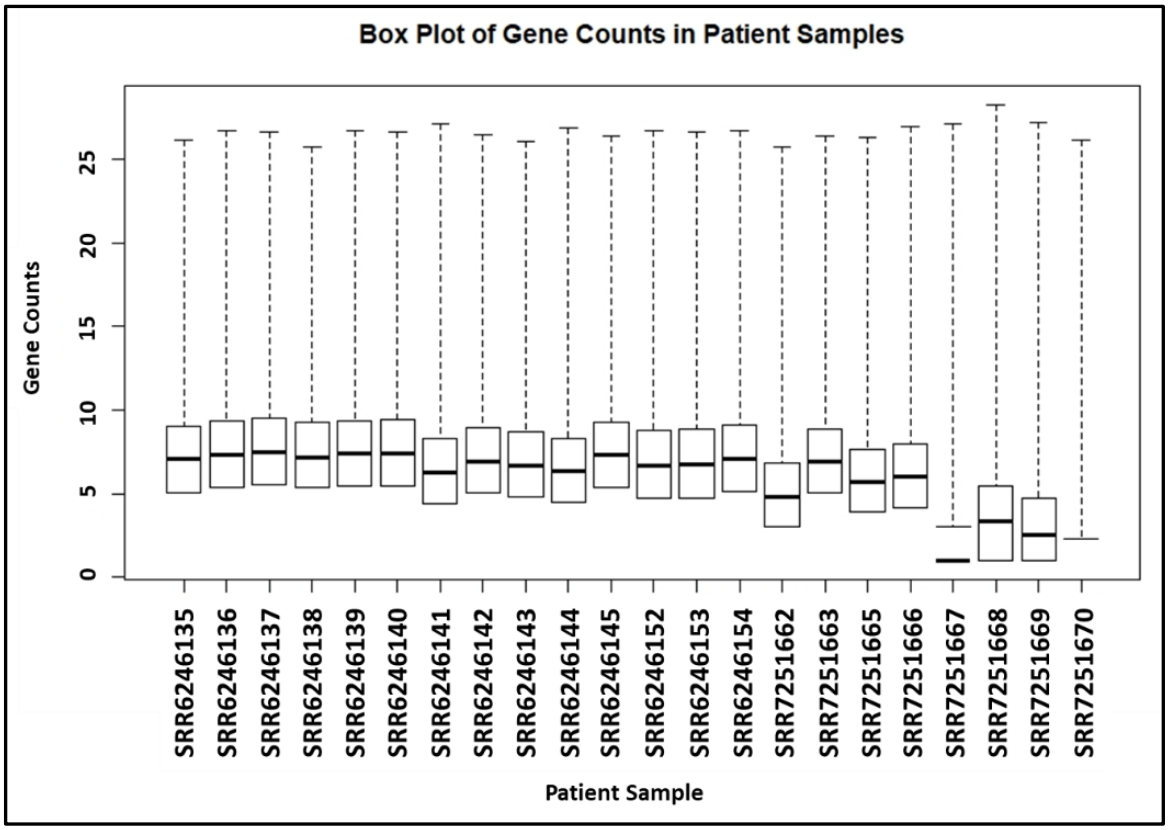
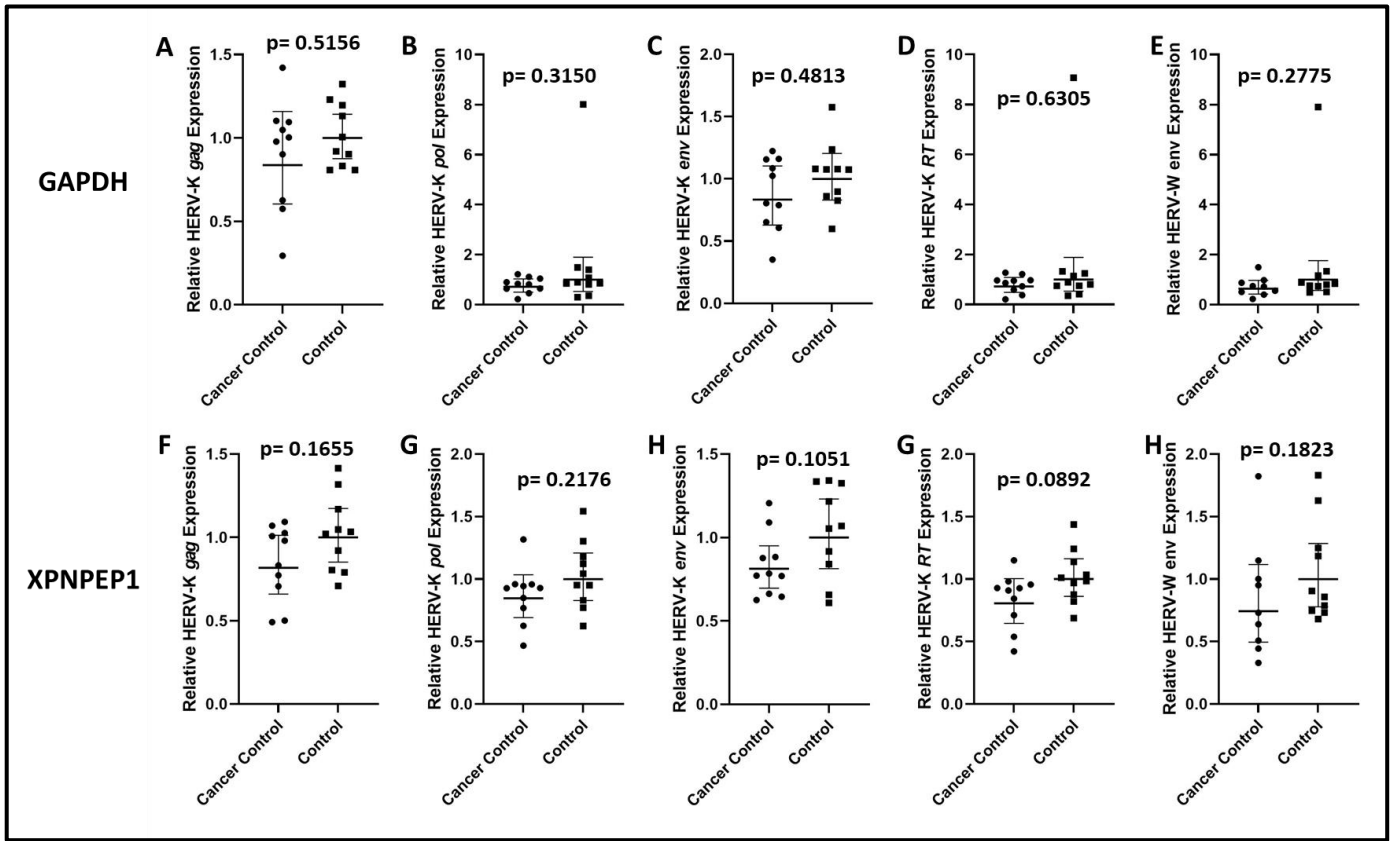


Figure S272. Box Plot of Endogenous Retrovirus Normalised Counts between n=15 ALS and n=7 Non-ALS controls.

The figure above displays statistical information on the counts data for ERVs within the n=15 ALS and n=7 non-ALS control sample set. The thick line inside each of the plots shows the median value for the date while the upper and lower end of the bars represent the interquartile ranges. The “whisker” above the plot, represented by the dotted line, shows the highest count recorded for an individual ERV within the sample.



Supplementary Figure S273. 2^-ΔΔCt Differential Expression levels for HERV-W *env* HERV-K *gag, pol, env* and *RT* gene transcripts in n=10 Cancer Control and n=10 no-Cancer Control Cases.

The graphs displayed in the figure above show 2^-ΔΔCt Differential Expression levels of A) & F) HERV-K *gag,* B) & G) HERV-K *pol,* C) & H) HERV-K *env* D) & I) HERV-K *RT* transcripts and E) & J) HERV-W *env* in cancer control and no-cancer control cases. The data is normalised either against GAPDH (A-E) or XPNPEP1 (F-J), the horizontal lines and error bars represent the geometric mean for the data set and its 95% confidence interval. *p*-values for all gene transcripts are >0.05 indicating a lack of statistical significance.

Supplementary Table S21. Binary Logistic Regression Analysis of HERV-K3 *pol* Differential Expression Using the Pfaffl Method

The combined table below shows the R^2 model summaries for the binary regression followed by the p-value significance (Sig.) that each variable is related to the difference between ALS and Non-ALS control samples.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	12.346[a]	.537	.715

Variables in the Equation

		B	S.E.	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	RIN	1.558	1.521	1	.306	4.751	.241	93.656
	Sex(1)	1.757	1.570	1	.263	5.795	.267	125.732
	Pfaffl	2.593	1.483	1	.080	13.366	.730	244.684
	Constant	-15.482	11.477	1	.177	.000		

a. Variable(s) entered on step 1: RIN, Sex, Pfaffl.

Supplementary Table S22. Binary Logistic Regression Analysis of HERV-K3 *pol* 2^-ΔΔCt Using a Single Reference Gene, GAPDH

The combined table below shows the R^2 model summaries for the binary regression followed by the p-value significance (Sig.) that each variable is related to the difference between ALS and Non-ALS control samples.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	9.777[a]	.592	.790

Variables in the Equation

		B	S.E.	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	RIN	2.737	2.591	1	.291	15.443	.096	2479.021
	Sex(1)	1.945	1.712	1	.256	6.994	.244	200.408
	GAPDH	5.139	3.634	1	.157	170.527	.137	211539.001
	Constant	-26.890	21.795	1	.217	.000		

a. Variable(s) entered on step 1: RIN, Sex, GAPDH.

Supplementary Table S23. Binary Logistic Regression Analysis of HERV-K3 *pol* 2^-ΔΔCt Using a Single Reference Gene, XPNPEP1

The combined table below shows the R^2 model summaries for the binary regression followed by the p-value significance (Sig.) that each variable is related to the difference between ALS and Non-ALS control samples.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	14.094[a]	.494	.659

Variables in the Equation

		B	S.E.	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	RIN	1.191	1.259	1	.344	3.291	.279	38.840
	Sex(1)	1.649	1.483	1	.266	5.202	.284	95.115
	XPNPEP1	1.627	.921	1	.077	5.088	.836	30.948
	Constant	-11.756	9.002	1	.192	.000		

a. Variable(s) entered on step 1: RIN, Sex, XPNPEP1.

References.

Achiron, A. and Gurevich, M. (2006) 'Peripheral blood gene expression signature mirrors central nervous system disease: The model of multiple sclerosis', *Autoimmunity Reviews*. Elsevier, pp. 517–522. doi: 10.1016/j.autrev.2006.02.009.

Adrião, A. *et al.* (2021) 'Identification of a novel mutation in MEF2C gene in an atypical patient with frontotemporal lobar degeneration', *Neurological Sciences 2021*. Springer, pp. 1–8. doi: 10.1007/S10072-021-05269-0.

Agoni, L., Guha, C. and Lenz, J. (2013) 'Detection of Human Endogenous Retrovirus K (HERV-K) Transcripts in Human Prostate Cancer Cell Lines', *Frontiers in Oncology*. Frontiers, 3, p. 180. doi: 10.3389/fonc.2013.00180.

Ajroud-Driss, S. and Siddique, T. (2015) 'Sporadic and hereditary amyotrophic lateral sclerosis (ALS)', *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*. Elsevier, 1852(4), pp. 679–684. doi: 10.1016/j.bbadis.2014.08.010.

Alba-Ferrara, L. M. and de Erausquin, G. A. (2013) 'What does anisotropy measure? Insights from increased and decreased anisotropy in selective fiber tracts in schizophrenia', *Frontiers in Integrative Neuroscience*. Frontiers, 7, p. 9. doi: 10.3389/fnint.2013.00009.

Alfahad, T. and Nath, A. (2013) 'Retroviruses and amyotrophic lateral sclerosis', *Antiviral Research*. Elsevier, pp. 180–187. doi: 10.1016/j.antiviral.2013.05.006.

Andersen, C. L., Jensen, J. L. and Ørntoft, T. F. (2004) 'Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets.', *Cancer research*. American Association for Cancer Research, 64(15), pp. 5245–50. doi: 10.1158/0008-5472.CAN-04-0496.

Ando, R. *et al.* (2015) 'Human T-lymphotropic Virus Type-I (HTLV-I)-associated Myelopathy with Bulbar Palsy-type Amyotrophic Lateral Sclerosis-like Symptoms', *Internal Medicine*. The Japanese Society of Internal Medicine, 54(9), pp. 1105–1107. doi: 10.2169/internalmedicine.54.3660.

Andrés-Benito, P. *et al.* (2017) 'Amyotrophic lateral sclerosis, gene deregulation in the anterior horn of the spinal cord and frontal cortex area 8: implications in frontotemporal lobar degeneration.', *Aging*. Impact Journals, LLC, 9(3), pp. 823–851. doi: 10.18632/aging.101195.

Angela Pérez-Novo, C. *et al.* (2005) 'Impact of RNA quality on reference gene expression stability', *BioTechniques*. Future Science Ltd London, UK , 39(1), pp. 52–56. doi: 10.2144/05391BM05.

Antony, J. M. *et al.* (2011) 'Human endogenous retroviruses and multiple sclerosis: Innocent bystanders or disease determinants?', *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*. Elsevier, 1812(2), pp. 162–176. doi: 10.1016/J.BBADIS.2010.07.016.

Arru, G. *et al.* (2018a) 'Humoral immunity response to human endogenous retroviruses K/W differentiates between amyotrophic lateral sclerosis and other neurological diseases', *European Journal of Neurology*. John Wiley & Sons, Ltd (10.1111), 25(8), pp. 1076-e84. doi: 10.1111/ene.13648.

Arru, G. *et al.* (2018b) 'Humoral immunity response to human endogenous retroviruses K/W differentiates between amyotrophic lateral sclerosis and other neurological diseases', *European Journal of Neurology*. John Wiley & Sons, Ltd, 25(8), pp. 1076-e84. doi: 10.1111/ENE.13648.

Arru, G. *et al.* (2021) 'HERV-K Modulates the Immune Response in ALS Patients', *Microorganisms*. Multidisciplinary Digital Publishing Institute (MDPI), 9(8). doi: 10.3390/MICROORGANISMS9081784.

Atamian, H. S. and Kaloshian, I. (2012) 'Construction of RNA-Seq Libraries from Large and Microscopic Tissues for the Illumina Sequencing Platform', in. Humana Press, Totowa, NJ, pp. 47–57. doi: 10.1007/978-1-61779-839-9_3.

Baccarella, A. *et al.* (2018) 'Empirical assessment of the impact of sample number and read depth on RNA-Seq analysis workflow performance', *BMC Bioinformatics*. BioMed Central Ltd., 19(1), p. 423. doi: 10.1186/s12859-018-2445-2.

Balendra, R. and Isaacs, A. M. (2018) 'C9orf72-mediated ALS and FTD: multiple pathways to disease', *Nature Reviews Neurology*. Nature Publishing Group, pp. 544–558. doi: 10.1038/s41582-018-0047-2.

Balestrieri, E. *et al.* (2012) 'HERVs Expression in Autism Spectrum Disorders', *PLoS ONE*. Public Library of Science, 7(11), p. e48831. doi: 10.1371/journal.pone.0048831.

Balestrieri, E. *et al.* (2015) 'Transcriptional Activity of Human Endogenous Retroviruses in Human Peripheral Blood Mononuclear Cells', *BioMed Research International*. Hindawi, 2015, pp. 1–9. doi: 10.1155/2015/164529.

Bannert, N. and Kurth, R. (2006) 'The Evolutionary Dynamics of Human Endogenous Retroviral Families', *Annual Review of Genomics and Human Genetics*. Annual Reviews, 7(1), pp. 149–173. doi: 10.1146/annurev.genom.7.080505.115700.

Barber, R. D. *et al.* (2005) 'GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues', *Physiological Genomics*. American Physiological Society, 21(3), pp. 389–395. doi: 10.1152/physiolgenomics.00025.2005.

Bartram, I. *et al.* (2014) 'Low expression of T-cell transcription factor BCL11b predicts inferior survival in adult standard risk T-cell acute lymphoblastic leukemia patients', *Journal of Hematology and Oncology*. BioMed Central Ltd., 7(1), p. 51. doi: 10.1186/s13045-014-0051-y.

Bashratyan, R. *et al.* (2017) 'Type 1 diabetes pathogenesis is modulated by spontaneous autoimmune responses to endogenous retrovirus antigens in NOD mice', *European Journal of Immunology*. Wiley-Blackwell, 47(3), pp. 575–584. doi: 10.1002/eji.201646755.

Bateman, A. *et al.* (2017) 'UniProt: the universal protein knowledgebase', *Nucleic Acids Research*. Oxford University Press, 45(D1), pp. D158–D169. doi: 10.1093/nar/gkw1099.

Behjati, S. and Tarpey, P. S. (2013) 'What is next generation sequencing?', *Archives of disease in childhood. Education and practice edition*. BMJ Publishing Group, 98(6), pp. 236–8. doi: 10.1136/archdischild-2013-304340.

Belshaw, R. *et al.* (2005) 'High Copy Number in Human Endogenous Retrovirus Families is Associated with Copying Mechanisms in Addition to Reinfection', *Molecular Biology and Evolution*. Oxford Academic, 22(4), pp. 814–817. doi: 10.1093/MOLBEV/MSI088.

Benachenhou, F. *et al.* (2013) 'Conserved structure and inferred evolutionary history of long terminal repeats (LTRs)', *Mobile DNA 2013 4:1*. BioMed Central, 4(1), pp. 1–16. doi: 10.1186/1759-8753-4-5.

Bernardo, V., Ribeiro Pinto, L. F. and Albano, R. M. (2013) 'Gene expression analysis by real-time PCR: Experimental demonstration of PCR detection limits', *Analytical Biochemistry*. Academic Press, 432(2), pp. 131–133. doi: 10.1016/J.AB.2012.09.029.

Bhardwaj, N. *et al.* (2014) 'HIV-1 Infection Leads to Increased Transcription of Human Endogenous Retrovirus HERV-K (HML-2) Proviruses In Vivo but Not to Increased Virion Production', *Journal of Virology*. American Society for Microbiology, 88(19), pp. 11108–11120. doi: 10.1128/JVI.01623-14.

Bhardwaj, N. *et al.* (2015) 'Differential Expression of HERV-K (HML-2) Proviruses in Cells

and Virions of the Teratocarcinoma Cell Line Tera-1', *Viruses*. Multidisciplinary Digital Publishing Institute, 7(3), pp. 939–968. doi: 10.3390/v7030939.

Bhat, R. K. *et al.* (2014) 'Human endogenous retrovirus-K(II) envelope induction protects neurons during HIV/AIDS', *PLoS ONE*. Edited by B. W. Banfield. Public Library of Science, 9(7), p. e97984. doi: 10.1371/journal.pone.0097984.

Bhattarai, A. *et al.* (2019) 'Serial assessment of iron in the motor cortex in limb-onset Amyotrophic Lateral Sclerosis using Quantitative Susceptibility Mapping', *bioRxiv*. Cold Spring Harbor Laboratory, p. 865709. doi: 10.1101/865709.

Bhetariya, P. J., Kriesel, J. D. and Fischer, K. F. (2017) 'Analysis of Human Endogenous Retrovirus Expression in Multiple Sclerosis Plaques.', *Journal of Emerging Diseases and Virology*. NIH Public Access, 3(2). doi: 10.16966/2473-1846.133.

Billingsley, K. J. *et al.* (2019) 'Analysis of repetitive element expression in the blood and skin of patients with Parkinson's disease identifies differential expression of satellite elements', *Scientific Reports*. Nature Publishing Group, 9(1), p. 4369. doi: 10.1038/s41598-019-40869-z.

Boldogkői, Z. *et al.* (2019) 'Long-Read Sequencing – A Powerful Tool in Viral Transcriptome Research', *Trends in Microbiology*. Elsevier Ltd, pp. 578–592. doi: 10.1016/j.tim.2019.01.010.

Borba, F. *et al.* (2019) 'Distinct patterns of cerebellar damage in sporadic and C9ORF72-related ALS', *Neurology*, 92(15 Supplement).

Bowen, L. N. *et al.* (2016) 'HIV-associated motor neuron disease', *Neurology*. American Academy of Neurology, 87(17), pp. 1756–1762. doi: 10.1212/WNL.0000000000003258.

Brazma, A. *et al.* (2001) 'Minimum information about a microarray experiment (MIAME)—toward standards for microarray data', *Nature Genetics*, 29(4), pp. 365–371. doi: 10.1038/ng1201-365.

Brodziak, A. *et al.* (2012) 'The role of human endogenous retroviruses in the pathogenesis of autoimmune diseases', *Medical Science Monitor*, 18(6), pp. RA80–RA88. doi: 10.12659/MSM.882892.

Brudek, T. *et al.* (2007) 'Activation of endogenous retrovirus reverse transcriptase in multiple sclerosis patient lymphocytes by inactivated HSV-1, HHV-6 and VZV', *Journal of Neuroimmunology*. Elsevier, 187(1–2), pp. 147–155. doi: 10.1016/j.jneuroim.2007.04.003.

Brudek, T. *et al.* (2009) 'B cells and monocytes from patients with active multiple sclerosis

exhibit increased surface expression of both HERV-H Env and HERV-W Env, accompanied by increased seroreactivity', *Retrovirology 2009 6:1*. BioMed Central, 6(1), pp. 1–13. doi: 10.1186/1742-4690-6-104.

Bustin, S. A. *et al.* (2009) 'The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments.', *Clinical chemistry*. Clinical Chemistry, 55(4), pp. 611–22. doi: 10.1373/clinchem.2008.112797.

Bustin, S. A. and Wittwer, C. T. (2017) 'MIQE: A Step Toward More Robust and Reproducible Quantitative PCR', *Clinical Chemistry*. American Association for Clinical Chemistry Inc., 63(9), pp. 1537–1538. doi: 10.1373/clinchem.2016.268953.

Buzdin, A. *et al.* (2003) 'Human-specific subfamilies of HERV-K (HML-2) long terminal repeats: Three master genes were active simultaneously during branching of hominoid lineages', *Genomics*. Academic Press, 81(2), pp. 149–156. doi: 10.1016/S0888-7543(02)00027-7.

Buzdin, A. *et al.* (2006) 'At least 50% of human-specific HERV-K (HML-2) long terminal repeats serve in vivo as active promoters for host nonrepetitive DNA transcription', *Journal of virology*. American Society for Microbiology, 80(21), pp. 10752–10762. doi: 10.1128/JVI.00871-06.

Buzdin, A. A., Prassolov, V. and Garazha, A. V (2017) 'Friends-Enemies: Endogenous Retroviruses Are Major Transcriptional Regulators of Human DNA.', *Frontiers in chemistry*. Frontiers Media SA, 5, p. 35. doi: 10.3389/fchem.2017.00035.

Byun, J. M. *et al.* (2021) 'The clinical significance of HERV-H LTR –associating 2 expression in cervical adenocarcinoma', *Medicine*. Wolters Kluwer Health, 100(1), p. e23691. doi: 10.1097/MD.0000000000023691.

Cai, C. *et al.* (2010) 'Is human blood a good surrogate for brain tissue in transcriptional studies?', *BMC Genomics 2010 11:1*. BioMed Central, 11(1), pp. 1–15. doi: 10.1186/1471-2164-11-589.

Carter, T. A. *et al.* (2021) 'Mosaic cis-regulatory evolution drives transcriptional partitioning of HERVH endogenous retrovirus in the human embryo', *bioRxiv*. Cold Spring Harbor Laboratory, p. 2021.07.08.451617. doi: 10.1101/2021.07.08.451617.

Chatzou, M. *et al.* (2016) 'Multiple sequence alignment modeling: methods and applications', *Briefings in Bioinformatics*. Oxford Academic, 17(6), pp. 1009–1023. doi: 10.1093/BIB/BBV099.

Chen, J., Foroozesh, M. and Qin, Z. (2019) 'Transactivation of human endogenous retroviruses by tumor viruses and their functions in virus-associated malignancies', *Oncogenesis*. Nature Publishing Group, 8(1), p. 6. doi: 10.1038/s41389-018-0114-y.

Chen, X. and Li, D. (2018) 'ERVcaller: Identifying and genotyping non-reference unfixed endogenous retroviruses (ERVs) and other transposable elements (TEs) using next-generation sequencing data', *bioRxiv*.

Cherrier, T. *et al.* (2013) 'CTIP2 is a negative regulator of P-TEFb.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 110(31), pp. 12655–60. doi: 10.1073/pnas.1220136110.

Cismasiu, V. B. *et al.* (2008) 'BCL11B is a general transcriptional repressor of the HIV-1 long terminal repeat in T lymphocytes through recruitment of the NuRD complex', *Virology*. Academic Press, 380(2), pp. 173–181. doi: 10.1016/J.VIROL.2008.07.035.

Coffin, J. M., Varmus, H. and Hughes, S. H. (2002) 'Retrotransposons, Endogenous Retroviruses, and the Evolution of Retroelements', in *Retroviruses*. 1st edn. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press. Available at: https://www.ncbi.nlm.nih.gov/books/NBK19468/.

Cohen, C. J., Lock, W. M. and Mager, D. L. (2009) 'Endogenous retroviral LTRs as promoters for human genes: A critical assessment', *Gene*. Elsevier, pp. 105–114. doi: 10.1016/j.gene.2009.06.020.

Cohen, T. J., Lee, V. M. Y. and Trojanowski, J. Q. (2011) 'TDP-43 functions and pathogenic mechanisms implicated in TDP-43 proteinopathies.', *Trends in molecular medicine*. NIH Public Access, 17(11), pp. 659–67. doi: 10.1016/j.molmed.2011.06.004.

Conesa, A. *et al.* (2016) 'A survey of best practices for RNA-seq data analysis', *Genome Biology*. BioMed Central Ltd. doi: 10.1186/s13059-016-0881-8.

Cosottini, M. *et al.* (2012) 'Structural and functional evaluation of cortical motor areas in Amyotrophic Lateral Sclerosis', *Experimental Neurology*. Academic Press, 234(1), pp. 169–180. doi: 10.1016/J.EXPNEUROL.2011.12.024.

Costagli, M. *et al.* (2016) 'Magnetic susceptibility in the deep layers of the primary motor cortex in Amyotrophic Lateral Sclerosis', *NeuroImage: Clinical*. Elsevier Inc., 12, pp. 965–969. doi: 10.1016/j.nicl.2016.04.011.

Coulson, D. T. R. *et al.* (2008) 'Identification of valid reference genes for the normalization of RT qPCR gene expression data in human brain tissue.', *BMC molecular biology*. BioMed

Central, 9, p. 46. doi: 10.1186/1471-2199-9-46.
Curis, E. *et al.* (2019) 'Selecting reference genes in RT-qPCR based on equivalence tests: a network based approach', *Scientific Reports*. Nature Research, 9(1), pp. 1–8. doi: 10.1038/s41598-019-52217-2.
Cytlak, U. *et al.* (2018) 'Ikaros family zinc finger 1 regulates dendritic cell development and function in humans', *Nature Communications 2018 9:1*. Nature Publishing Group, 9(1), pp. 1–10. doi: 10.1038/s41467-018-02977-8.
D'Erchia, A. M. *et al.* (2017) 'Massive transcriptome sequencing of human spinal cord tissues provides new insights into motor neuron degeneration in ALS', *Scientific Reports*. Nature Publishing Group, 7(1), pp. 1–20. doi: 10.1038/s41598-017-10488-7.
Dadon-Nachum, M., Melamed, E. and Offen, D. (2011) 'The "dying-back" phenomenon of motor neurons in ALS', *Journal of Molecular Neuroscience*. Humana Press Inc, pp. 470–477. doi: 10.1007/s12031-010-9467-1.
Dean, B., Udawela, M. and Scarr, E. (2016) 'Validating reference genes using minimally transformed qpcr data: Findings in human cortex and outcomes in schizophrenia', *BMC Psychiatry*. BioMed Central, 16(1), p. 154. doi: 10.1186/s12888-016-0855-0.
Delic, V. *et al.* (2018) 'Discrete mitochondrial aberrations in the spinal cord of sporadic ALS patients', *Journal of Neuroscience Research*. John Wiley and Sons Inc., 96(8), pp. 1353–1366. doi: 10.1002/jnr.24249.
Dembny, P. *et al.* (2020) 'Human endogenous retrovirus HERV-K(HML-2) RNA causes neurodegeneration through Toll-like receptors', *JCI Insight*. American Society for Clinical Investigation, 5(7). doi: 10.1172/jci.insight.131093.
Dervan, E. *et al.* (2021) 'Ancient Adversary - HERV-K (HML-2) in Cancer', *Frontiers in oncology*. Front Oncol, 11. doi: 10.3389/FONC.2021.658489.
Derveaux, S., Vandesompele, J. and Hellemans, J. (2010) 'How to do successful gene expression analysis using real-time PCR', *Methods*. Academic Press, 50(4), pp. 227–230. doi: 10.1016/J.YMETH.2009.11.001.
Desplats, P. *et al.* (2013) 'Molecular and pathologic insights from latent HIV-1 infection in the human brain.', *Neurology*. Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology, 80(15), pp. 1415–23. doi: 10.1212/WNL.0b013e31828c2e9e.
Dieffenbach, C. W., Lowe, T. M. and Dveksler, G. S. (1993) 'General concepts for PCR primer design.', *PCR methods and applications*, 3(3), pp. S30–S37. doi:

10.1101/gr.3.3.S30.
van Dijk, E. L. *et al.* (2014) 'Ten years of next-generation sequencing technology', *Trends in Genetics*. Elsevier Current Trends, 30(9), pp. 418–426. doi: 10.1016/J.TIG.2014.07.001.
Dolei, A. *et al.* (2015) 'The aliens inside human DNA: HERV-W/MSRV/syncytin-1 endogenous retroviruses and neurodegeneration', *The Journal of Infection in Developing Countries*. Journal of Infection in Developing Countries, 9(06), pp. 577–587. doi: 10.3855/jidc.6916.
Dolei, A. *et al.* (2019) 'Expression of HERV genes as possible biomarker and target in neurodegenerative diseases', *International Journal of Molecular Sciences*. MDPI AG. doi: 10.3390/ijms20153706.
Douville, R. *et al.* (2011) 'Identification of active loci of a human endogenous retrovirus in neurons of patients with amyotrophic lateral sclerosis', *Annals of Neurology*. NIH Public Access, 69(1), pp. 141–151. doi: 10.1002/ana.22149.
Douville, R. N. and Nath, A. (2014) 'Human endogenous retroviruses and the nervous system', *Handbook of Clinical Neurology*. Elsevier, 123, pp. 465–485. doi: 10.1016/B978-0-444-53488-0.00022-5.
Douville, R. N. and Nath, A. (2017) 'Human Endogenous Retrovirus-K and TDP-43 Expression Bridges ALS and HIV Neuropathology.', *Frontiers in microbiology*. Frontiers Media SA, 8, p. 1986. doi: 10.3389/fmicb.2017.01986.
Downey, R. F. *et al.* (2015) 'Human endogenous retrovirus K and cancer: Innocent bystander or tumorigenic accomplice?', *International journal of cancer*. Int J Cancer, 137(6), pp. 1249–1257. doi: 10.1002/IJC.29003.
Durrenberger, P. F. *et al.* (2010) 'Effects of Antemortem and Postmortem Variables on Human Brain mRNA Quality: A BrainNet Europe Study', *Journal of Neuropathology & Experimental Neurology*. Oxford University Press, 69(1), pp. 70–81. doi: 10.1097/NEN.0b013e3181c7e32f.
Durrenberger, P. F. *et al.* (2012) 'Selection of novel reference genes for use in the human central nervous system: a BrainNet Europe Study', *Acta Neuropathol*, 124, pp. 893–903. doi: 10.1007/s00401-012-1027-z.
Edgar, R. C. (2004) 'MUSCLE: Multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Research*. Oxford University Press, 32(5), pp. 1792–1797. doi: 10.1093/nar/gkh340.

Ehinger, J. K. *et al.* (2015) 'Mitochondrial dysfunction in blood cells from amyotrophic lateral sclerosis patients', *Journal of Neurology*. Dr. Dietrich Steinkopff Verlag GmbH and Co. KG, 262(6), pp. 1493–1503. doi: 10.1007/s00415-015-7737-0.

Ehlhardt, S. *et al.* (2006) 'Human endogenous retrovirus HERV-K(HML-2) Rec expression and transcriptional activities in normal and rheumatoid arthritis synovia.', *The Journal of rheumatology*. The Journal of Rheumatology, 33(1), pp. 16–23. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16395745 (Accessed: 27 March 2018).

Eisen, A. *et al.* (2017) 'Cortical influences drive amyotrophic lateral sclerosis.', *Journal of neurology, neurosurgery, and psychiatry*. BMJ Publishing Group Ltd, 88(11), pp. 917–924. doi: 10.1136/jnnp-2017-315573.

Eisenberg, E. and Levanon, E. Y. (2013) 'Human housekeeping genes, revisited', *Trends in Genetics*. Elsevier Current Trends, pp. 569–574. doi: 10.1016/j.tig.2013.05.010.

Fernández, M., Sierra-Arregui, T. and Peñagarikano, O. (2019) 'The Cerebellum and Autism: More than Motor Control', in *Behavioral Neuroscience*. IntechOpen. doi: 10.5772/intechopen.85897.

Fleige, S. and Pfaffl, M. W. (2006) 'RNA integrity and the effect on the real-time qRT-PCR performance', *Molecular Aspects of Medicine*. Pergamon, 27(2–3), pp. 126–139. doi: 10.1016/J.MAM.2005.12.003.

Frank, J. A. and Feschotte, C. (2017) 'Co-option of endogenous viral sequences for host cell function', *Current Opinion in Virology*. Elsevier, pp. 81–89. doi: 10.1016/j.coviro.2017.07.021.

Fujinami, R. S. and Libbey, J. E. (1999) 'Endogenous retroviruses: Are they the cause of multiple sclerosis?', *Trends in Microbiology*. Elsevier, pp. 263–264. doi: 10.1016/S0966-842X(99)01532-2.

Gajecka, M. (2016) 'Unrevealed mosaicism in the next-generation sequencing era.', *Molecular genetics and genomics : MGG*. Springer, 291(2), pp. 513–30. doi: 10.1007/s00438-015-1130-7.

Gallego Romero, I. *et al.* (2014) 'RNA-seq: impact of RNA degradation on transcript quantification', *BMC Biology*. BioMed Central Ltd., 12(1), p. 42. doi: 10.1186/1741-7007-12-42.

Garcia-Montojo, M. *et al.* (2013) 'The DNA Copy Number of Human Endogenous Retrovirus-W (MSRV-Type) Is Increased in Multiple Sclerosis Patients and Is Influenced by

Gender and Disease Severity', *PLoS ONE*. Public Library of Science, 8(1). doi: 10.1371/journal.pone.0053623.

Garcia-Montojo, M. *et al.* (2018) 'Human endogenous retrovirus-K (HML-2): a comprehensive review', *Critical Reviews in Microbiology*. Taylor & Francis, pp. 1–24. doi: 10.1080/1040841X.2018.1501345.

Garcia-Montojo, M. *et al.* (2021) 'Inhibition of HERV-K (HML-2) in amyotrophic lateral sclerosis patients on antiretroviral therapy', *Journal of the Neurological Sciences*. Elsevier, 423, p. 117358. doi: 10.1016/J.JNS.2021.117358.

García-Montojo, M. *et al.* (2014) 'HERV-W polymorphism in chromosome X is associated with multiple sclerosis risk and with differential expression of MSRV', *Retrovirology*. BioMed Central, 11(1), p. 2. doi: 10.1186/1742-4690-11-2.

Garson, J. A. *et al.* (2019) 'Quantitative analysis of human endogenous retrovirus-K transcripts in postmortem premotor cortex fails to confirm elevated expression of HERV-K RNA in amyotrophic lateral sclerosis', *Acta Neuropathologica Communications*. BioMed Central, 7(1), p. 45. doi: 10.1186/s40478-019-0698-2.

Gemmell, P., Hein, J. and Katzourakis, A. (2019) 'The Exaptation of HERV-H: Evolutionary Analyses Reveal the Genomic Features of Highly Transcribed Elements ', *Frontiers in Immunology* , p. 1339. Available at: https://www.frontiersin.org/article/10.3389/fimmu.2019.01339.

Genome Reference Consortium (2020) *Human Genome Assembly GRCh37.p13 - Genome Reference Consortium*. Available at: https://www.ncbi.nlm.nih.gov/grc/human (Accessed: 7 April 2020).

Gifford, R. J. *et al.* (2018) 'Nomenclature for endogenous retrovirus (ERV) loci', *Retrovirology*. BioMed Central Ltd. doi: 10.1186/s12977-018-0442-1.

Giménez-Orenga, K. and Oltra, E. (2021) 'Human Endogenous Retrovirus as Therapeutic Targets in Neurologic Disease', *Pharmaceuticals*. Multidisciplinary Digital Publishing Institute (MDPI), 14(6). doi: 10.3390/PH14060495.

Golan, M. *et al.* (2008) 'Human Endogenous Retrovirus (HERV-K) Reverse Transcriptase as a Breast Cancer Prognostic Marker', *Neoplasia*. Neoplasia Press, 10(6), pp. 521-IN2. doi: 10.1593/neo.07986.

Gold, J. *et al.* (2019) 'Safety and Tolerability of Triumeq in Amyotrophic Lateral Sclerosis: The Lighthouse Trial'. Available at:

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3347916 (Accessed: 25 April 2019).
Gontier, G. *et al.* (2018) 'Tet2 Rescues Age-Related Regenerative Decline and Enhances Cognitive Function in the Adult Mouse Brain', *Cell Reports*. Cell Press, 22(8), pp. 1974–1981. doi: 10.1016/J.CELREP.2018.02.001.
Gonzalez-Cao, M. *et al.* (2016) 'Human endogenous retroviruses and cancer.', *Cancer biology & medicine*. Chinese Anti-Cancer Association, 13(4), pp. 483–488. doi: 10.20892/j.issn.2095-3941.2016.0080.
Grandi, N. *et al.* (2017) 'Identification of a novel HERV-K(HML10): Comprehensive characterization and comparative analysis in non-human primates provide insights about HML10 proviruses structure and diffusion', *Mobile DNA*. BioMed Central, 8(1), p. 15. doi: 10.1186/s13100-017-0099-7.
Grandi, N. and Tramontano, E. (2017) 'Type W human endogenous retrovirus (HERV-W) integrations and their mobilization by L1 machinery: Contribution to the human transcriptome and impact on the host physiopathology', *Viruses*. Multidisciplinary Digital Publishing Institute, p. 162. doi: 10.3390/v9070162.
Grandi, N. and Tramontano, E. (2018) 'HERV Envelope Proteins: Physiological Role and Pathogenic Potential in Cancer and Autoimmunity', *Frontiers in Microbiology*. Frontiers, 9(462), p. 462. doi: 10.3389/fmicb.2018.00462.
Gray, L. R. *et al.* (2019) 'HIV-1 Rev interacts with HERV-K RcREs present in the human genome and promotes export of unspliced HERV-K proviral RNA', *Retrovirology*. BioMed Central Ltd., 16(1), p. 40. doi: 10.1186/s12977-019-0505-y.
Gröger, V. and Cynis, H. (2018) 'Human Endogenous Retroviruses and Their Putative Role in the Development of Autoimmune Disorders Such as Multiple Sclerosis', *Frontiers in microbiology*. Frontiers Media SA, 9(February), p. 265. doi: 10.3389/fmicb.2018.00265.
Guide to Performing Relative Quantitation of Gene Expression Using Real-Time Quantitative PCR (2004). Available at: https://www.gu.se/digitalAssets/1125/1125331_ABI_-_Guide_Relative_Quantification_using_realtime_PCR.pdf (Accessed: 20 December 2018).
Guo, Y. *et al.* (2017) 'Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis', *Genomics*. Academic Press Inc., 109(2), pp. 83–90. doi: 10.1016/j.ygeno.2017.01.005.
Hahn, S. *et al.* (2008) 'Serological Response to Human Endogenous Retrovirus K in

Melanoma Patients Correlates with Survival Probability', *AIDS Research and Human Retroviruses*. Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801-5215 USA, 24(5), pp. 717–723. doi: 10.1089/aid.2007.0286.

Han, Y. *et al.* (2015) 'Advanced applications of RNA sequencing and challenges', *Bioinformatics and Biology Insights*. Libertas Academica Ltd., 9, pp. 29–46. doi: 10.4137/BBI.S28991.

Hanke, K. *et al.* (2013) 'The Rec protein of HERV-K(HML-2) upregulates androgen receptor activity by binding to the human small glutamine-rich tetratricopeptide repeat protein (hSGT)', *International Journal of Cancer*. Wiley-Blackwell, 132(3), pp. 556–567. doi: 10.1002/ijc.27693.

Hanke, K., Hohn, O. and Bannert, N. (2016) 'HERV-K(HML-2), a seemingly silent subtenant - but still waters run deep', *APMIS*, pp. 67–87. doi: 10.1111/apm.12475.

Harris, J. L., Reeves, T. M. and Phillips, L. L. (2009) 'Injury modality, survival interval, and sample region are critical determinants of qRT-PCR reference gene selection during long-term recovery from brain trauma.', *Journal of neurotrauma*. Mary Ann Liebert, Inc., 26(10), pp. 1669–81. doi: 10.1089/neu.2009.0875.

Heid, C. A. *et al.* (1996) 'Real time quantitative PCR.', *Genome Research*. Cold Spring Harbor Laboratory Press, 6(10), pp. 986–994. doi: 10.1101/gr.6.10.986.

Hera, B. de la and Urcelay, E. (2016) 'HERVs in Multiple Sclerosis — From Insertion to Therapy', in *Advances in Molecular Retrovirology*. InTech. doi: 10.5772/61726.

HF, Y. *et al.* (2011) 'Polyomavirus enhancer activator 3 protein promotes breast cancer metastatic progression through Snail-induced epithelial-mesenchymal transition', *The Journal of pathology*. J Pathol, 224(1), pp. 78–89. doi: 10.1002/PATH.2859.

Higgins, D. (1997) 'Multiple Sequence Alignment', *Genetic Databases*. Academic Press, pp. 165–183. doi: 10.1016/B978-012101625-8/50010-4.

Higuchi, R. *et al.* (1992) 'Simultaneous Amplification and Detection of Specific DNA Sequences', *Bio/Technology*. Nature Publishing Group, 10(4), pp. 413–417. doi: 10.1038/nbt0492-413.

Hindson, B. J. *et al.* (2011) 'High-throughput droplet digital PCR system for absolute quantitation of DNA copy number', *Analytical Chemistry*. American Chemical Society, 83(22), pp. 8604–8610. doi: 10.1021/AC202028G/SUPPL_FILE/AC202028G_SI_001.PDF.

Hohn, O., Hanke, K. and Bannert, N. (2013) 'HERV-K(HML-2), the Best Preserved Family of

HERVs: Endogenization, Expression, and Implications in Health and Disease', *Frontiers in Oncology*. Frontiers Media SA, 3, p. 246. doi: 10.3389/fonc.2013.00246.
Hosaka, T. *et al.* (2019) 'Extracellular RNAs as biomarkers of sporadic amyotrophic lateral sclerosis and other neurodegenerative diseases', *International Journal of Molecular Sciences*. MDPI AG. doi: 10.3390/ijms20133148.
Hu, W.-S. and Hughes, S. H. (2012) 'HIV-1 reverse transcription', *Cold Spring Harbor perspectives in medicine*. Cold Spring Harbor Laboratory Press, 2(10), p. a006882. doi: 10.1101/cshperspect.a006882.
Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2009a) 'Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists', *Nucleic acids research*. Nucleic Acids Res, 37(1), pp. 1–13. doi: 10.1093/NAR/GKN923.
Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2009b) 'Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources', *Nature protocols*. Nat Protoc, 4(1), pp. 44–57. doi: 10.1038/NPROT.2008.211.
Huggett, J. *et al.* (2005) 'Real-time RT-PCR normalisation; strategies and considerations', *Genes & Immunity*. Nature Publishing Group, 6(4), pp. 279–284. doi: 10.1038/sj.gene.6364190.
Hughes, J. F. and Coffin, J. M. (2004) 'Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: Implications for human and viral evolution', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 101(6), pp. 1668–1672. doi: 10.1073/pnas.0307885100.
Ibba, G. *et al.* (2018) 'Disruption by SaCas9 Endonuclease of HERV-Kenv, a Retroviral Gene with Oncogenic and Neuropathogenic Potential, Inhibits Molecules Involved in Cancer and Amyotrophic Lateral Sclerosis.', *Viruses*. Multidisciplinary Digital Publishing Institute (MDPI), 10(8). doi: 10.3390/v10080412.
IFIT2 Gene - GeneCards | IFIT2 Protein | IFIT2 Antibody (no date). Available at: https://www.genecards.org/cgi-bin/carddisp.pl?gene=IFIT2 (Accessed: 30 March 2022).
illumina (2011) 'MiSeq ™ System FAQs'. Available at: www.illumina.com/company/legal.html. (Accessed: 24 April 2019).
Ishihara, T. *et al.* (2022) 'Endogenous human retrovirus-K is not increased in the affected tissues of Japanese ALS patients', *Neuroscience Research*. Elsevier. doi: 10.1016/J.NEURES.2022.01.009.

Jain, M. *et al.* (2018) 'Nanopore sequencing and assembly of a human genome with ultra-long reads', *Nature Biotechnology*. Nature Publishing Group, 36(4), pp. 338–345. doi: 10.1038/nbt.4060.

James Knierim (2000) *Chapter 3: Motor Cortex*, *Neuroscience Online*. Available at: https://nba.uth.tmc.edu/neuroscience/m/s3/chapter03.html (Accessed: 2 August 2019).

James Knierim (2018) *Motor Cortex (Section 3, Chapter 3) Neuroscience Online: An Electronic Textbook for the Neurosciences | Department of Neurobiology and Anatomy - The University of Texas Medical School at Houston*, *Neuroscience Online*. Available at: https://nba.uth.tmc.edu/neuroscience/m/s3/chapter03.html (Accessed: 3 December 2018).

Jamuar, S. S., D'Gama, A. M. and Walsh, C. A. (2016) 'Somatic Mosaicism and Neurological Diseases', *Genomics, Circuits, and Pathways in Clinical Neuropsychiatry*. Academic Press, pp. 179–199. doi: 10.1016/B978-0-12-800105-9.00012-3.

Jia, Y. *et al.* (2004) 'An association study between polymorphisms in three genes of 14-3-3 (tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein) family and paranoid schizophrenia in northern Chinese population.', *European Psychiatry*, 19(6). Available at: https://www.sciencedirect.com/science/article/pii/S0924933804001841?via%3Dihub (Accessed: 26 October 2018).

Jia, Y. (2012) 'Real-Time PCR', *Methods in Cell Biology*. Academic Press, 112, pp. 55–68. doi: 10.1016/B978-0-12-405914-6.00003-2.

Johanning, G. L. *et al.* (2017) 'Expression of human endogenous retrovirus-K is strongly associated with the basal-like breast cancer phenotype', *Scientific Reports*. Nature Publishing Group, 7, p. 41960. doi: 10.1038/srep41960.

Johnson, A. A. *et al.* (2012) 'The Role of DNA Methylation in Aging, Rejuvenation, and Age-Related Disease', *Rejuvenation Research*. Mary Ann Liebert, Inc., 15(5), p. 483. doi: 10.1089/REJ.2012.1324.

Jones, A. R. *et al.* (2021) 'A HML6 endogenous retrovirus on chromosome 3 is upregulated in amyotrophic lateral sclerosis motor cortex', *Scientific Reports 2021 11:1*. Nature Publishing Group, 11(1), pp. 1–10. doi: 10.1038/s41598-021-93742-3.

Jones, R. B. *et al.* (2012) 'HERV-K-specific T cells eliminate diverse HIV-1/2 and SIV primary isolates.', *The Journal of clinical investigation*. American Society for Clinical Investigation,

122(12), pp. 4473–89. doi: 10.1172/JCI64560.
de Jonge, H. J. M. *et al.* (2007) 'Evidence based selection of housekeeping genes.', *PloS one*. Public Library of Science, 2(9), p. e898. doi: 10.1371/journal.pone.0000898.
Ju, J. *et al.* (2006) 'Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 103(52), pp. 19635–40. doi: 10.1073/pnas.0609513103.
Katoh, I. and Kurata, S. (2013) 'Association of Endogenous Retroviruses and Long Terminal Repeats with Human Disorders', *Frontiers in Oncology*. Frontiers, 3, p. 234. doi: 10.3389/fonc.2013.00234.
Katsura, Y. and Asai, S. (2019) 'Evolutionary Medicine of Retroviruses in the Human Genome', *American Journal of the Medical Sciences*. Elsevier B.V., pp. 384–388. doi: 10.1016/j.amjms.2019.09.007.
Kiaei, L. and Kiaei, M. (2021) 'RNA as a source of biomarkers for amyotrophic lateral sclerosis', *Metabolic Brain Disease*. Springer, pp. 1–6. doi: 10.1007/S11011-021-00738-Z/FIGURES/2.
Kiernan, M. C. *et al.* (2011) 'Amyotrophic lateral sclerosis', in *The Lancet*. Elsevier, pp. 942–955. doi: 10.1016/S0140-6736(10)61156-7.
Kirschneck, C. *et al.* (2017) 'Valid gene expression normalization by RT-qPCR in studies on hPDL fibroblasts with focus on orthodontic tooth movement and periodontitis', *Scientific Reports*. Nature Publishing Group, 7(1), p. 14751. doi: 10.1038/s41598-017-15281-0.
Kiskinis, E. *et al.* (2014) 'Pathways disrupted in human ALS motor neurons identified through genetic correction of mutant SOD1', *Cell Stem Cell*. Cell Press, 14(6), pp. 781–795. doi: 10.1016/j.stem.2014.03.004.
Kleiveland, C. (2015) 'Peripheral blood mononuclear cells', in *The Impact of Food Bioactives on Health: In Vitro and Ex Vivo Models*. Springer International Publishing, pp. 161–167. doi: 10.1007/978-3-319-16104-4_15.
de Kok, J. B. *et al.* (2005) 'Normalization of gene expression measurements in tumor tissues: comparison of 13 endogenous control genes', *Laboratory Investigation*. Nature Publishing Group, 85(1), pp. 154–159. doi: 10.1038/labinvest.3700208.
Kongsbak, M. *et al.* (2013) 'The Vitamin D Receptor and T Cell Function', *Frontiers in Immunology*. Frontiers, 4(JUN), p. 148. doi: 10.3389/FIMMU.2013.00148.

Koppelkamm, A. *et al.* (2010) 'Validation of adequate endogenous reference genes for the normalisation of qPCR gene expression data in human post mortem tissue.', *International journal of legal medicine*. Springer-Verlag, 124(5), pp. 371–380. doi: 10.1007/s00414-010-0433-9.

Koppelkamm, A. *et al.* (2011) 'RNA integrity in post-mortem samples: Influencing parameters and implications on RT-qPCR assays', *International Journal of Legal Medicine*. Springer-Verlag, 125(4), pp. 573–580. doi: 10.1007/s00414-011-0578-1.

Kovanda, A. *et al.* (2018) 'Differential expression of microRNAs and other small RNAs in muscle tissue of patients with ALS and healthy age-matched controls', *Scientific Reports*. Nature Publishing Group, 8(1), pp. 1–15. doi: 10.1038/s41598-018-23139-2.

Kozera, B. and Rapacz, M. (2013) 'Reference genes in real-time PCR', *Journal of Applied Genetics*. Springer Berlin Heidelberg, 54(4), pp. 391–406. doi: 10.1007/s13353-013-0173-x.

Krzysztalowska-Wawrzyniak, M. *et al.* (2011) 'The distribution of human endogenous retrovirus K-113 in health and autoimmune diseases in Poland', *Rheumatology*. Oxford University Press, 50(7), pp. 1310–1314. doi: 10.1093/rheumatology/ker022.

Kuang, J. *et al.* (2018) 'An overview of technical considerations when using quantitative real-time PCR analysis of gene expression in human exercise research', *PLOS ONE*. Edited by R. Kalendar. Public Library of Science, 13(5), p. e0196438. doi: 10.1371/journal.pone.0196438.

Kukurba, Kimberly R and Montgomery, S. B. (2015) 'RNA Sequencing and Analysis.', *Cold Spring Harbor protocols*. NIH Public Access, 2015(11), pp. 951–69. doi: 10.1101/pdb.top084970.

Kukurba, Kimberly R. and Montgomery, S. B. (2015) 'RNA sequencing and analysis', *Cold Spring Harbor Protocols*. Cold Spring Harbor Laboratory Press, 2015(11), pp. 951–969. doi: 10.1101/pdb.top084970.

Kumar, S. *et al.* (2016) 'MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets Downloaded from', *Mol. Biol. Evol*, 33(7), pp. 1870–1874. doi: 10.1093/molbev/msw054.

Küry, P. *et al.* (2018) 'Human Endogenous Retroviruses in Neurological Diseases', *Trends in Molecular Medicine*. doi: 10.1016/j.molmed.2018.02.007.

van der Kuyl, A. C. (2012) 'HIV infection and HERV expression: a review.', *Retrovirology*.

BioMed Central, 9, p. 6. doi: 10.1186/1742-4690-9-6.

Lander, E. S. *et al.* (2001) 'Initial sequencing and analysis of the human genome', *Nature*. Nature Publishing Group, 409(6822), pp. 860–921. doi: 10.1038/35057062.

Laska, M. J. *et al.* (2012) 'Expression of HERV-Fc1, a human endogenous retrovirus, is increased in patients with active multiple sclerosis.', *Journal of virology*. American Society for Microbiology Journals, 86(7), pp. 3713–22. doi: 10.1128/JVI.06723-11.

Lederer, C. W. *et al.* (2007) 'Pathways and genes differentially expressed in the motor cortex of patients with sporadic amyotrophic lateral sclerosis', *BMC Genomics*. BioMed Central, 8(1), pp. 1–26. doi: 10.1186/1471-2164-8-26/FIGURES/7.

Lee, J.-U., Kim, L.-K. and Choi, J.-M. (2018) 'Revisiting the Concept of Targeting NFAT to Control T Cell Immunity and Autoimmune Diseases', *Frontiers in Immunology*. Frontiers, 9(NOV), p. 2747. doi: 10.3389/FIMMU.2018.02747.

Lee, Y. N., Malim, M. H. and Bieniasz, P. D. (2008) 'Hypermutation of an Ancient Human Retrovirus by APOBEC3G', *Journal of Virology*. American Society for Microbiology, 82(17), pp. 8762–8770. doi: 10.1128/JVI.00751-08.

Leib-Mösch, C. *et al.* (1993) 'Genomic distribution and transcription of solitary HERV-K LTRs', *Genomics*. Academic Press, 18(2), pp. 261–269. doi: 10.1006/geno.1993.1464.

Lemaitre, C. *et al.* (2014) 'The HERV-K Human Endogenous Retrovirus Envelope Protein Antagonizes Tetherin Antiviral Activity', *Journal of Virology*. American Society for Microbiology, 88(23), pp. 13626–13637. doi: 10.1128/JVI.02234-14.

Lennon, M. J. *et al.* (2016) 'Bcl11b: A New Piece to the Complex Puzzle of Amyotrophic Lateral Sclerosis Neuropathogenesis?', *Neurotoxicity Research*. Springer US, 29(2), pp. 201–207. doi: 10.1007/s12640-015-9573-5.

Lennon, M. J. *et al.* (2017) 'Bcl11b—A Critical Neurodevelopmental Transcription Factor—Roles in Health and Disease', *Frontiers in Cellular Neuroscience*. Frontiers, 11, p. 89. doi: 10.3389/fncel.2017.00089.

Levet, S. *et al.* (2017) 'An ancestral retroviral protein identified as a therapeutic target in type-1 diabetes.', *JCI insight*. American Society for Clinical Investigation, 2(17). doi: 10.1172/jci.insight.94387.

Li, C. and Zhang, J. (2019) 'Stop-codon read-through arises largely from molecular errors and is generally nonadaptive', *PLoS genetics*. Public Library of Science, 15(5), pp. e1008141–e1008141. doi: 10.1371/journal.pgen.1008141.

Li, F. *et al.* (2019) 'Transcription of human endogenous retroviruses in human brain by RNA-seq analysis', *PLoS ONE*. Edited by K. Ruprecht. Public Library of Science, 14(1), p. e0207353. doi: 10.1371/journal.pone.0207353.

Li, J. *et al.* (2016) 'Comparison of microarray and RNA-Seq analysis of mRNA expression in dermal mesenchymal stem cells', *Biotechnology Letters*. Springer Netherlands, 38(1), pp. 33–41. doi: 10.1007/s10529-015-1963-5.

Li, S. *et al.* (2014) 'Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study', *Nature Biotechnology*. Nature Publishing Group, 32(9), pp. 915–925. doi: 10.1038/nbt.2972.

Li, W. *et al.* (2015) 'Human endogenous retrovirus-K contributes to motor neuron disease.', *Science translational medicine*. American Association for the Advancement of Science, 7(307), p. 307ra153. doi: 10.1126/scitranslmed.aac8201.

Li, Y. *et al.* (2022) 'Human endogenous retrovirus K (HERV-K) env in neuronal extracellular vesicles: a new biomarker of motor neuron disease', *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*. Taylor and Francis Ltd., 23(1–2), pp. 100–107. doi: 10.1080/21678421.2021.1936061/SUPPL_FILE/IAFD_A_1936061_SM9151.DOCX.

Liew, C. C. *et al.* (2006) 'The peripheral blood transcriptome dynamically reflects system wide biology: A potential diagnostic tool', *Journal of Laboratory and Clinical Medicine*, 147(3), pp. 126–132. doi: 10.1016/j.lab.2005.10.005.

Liu, E. Y., Russ, J. and Lee, E. B. (2020) 'Neuronal Transcriptome from C9orf72 Repeat Expanded Human Tissue is Associated with Loss of C9orf72 Function', *Free neuropathology*. NIH Public Access, 1. Available at: /pmc/articles/PMC7470232/ (Accessed: 19 March 2022).

Livak, K. J. and Schmittgen, T. D. (2001) 'Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method', *Methods*. Academic Press, 25(4), pp. 402–408. doi: 10.1006/METH.2001.1262.

Logroscino, G. *et al.* (2010) 'Incidence of amyotrophic lateral sclerosis in Europe', *Journal of Neurology, Neurosurgery and Psychiatry*. NIH Public Access, 81(4), pp. 385–390. doi: 10.1136/jnnp.2009.183525.

Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*. BioMed Central Ltd., 15(12), p. 550. doi: 10.1186/s13059-014-0550-8.

Lowe, R. *et al.* (2017) 'Transcriptomics technologies', *PLoS Computational Biology*. Public Library of Science, 13(5). doi: 10.1371/journal.pcbi.1005457.
Lu, H., Giordano, F. and Ning, Z. (2016) 'Oxford Nanopore MinION Sequencing and Genome Assembly', *Genomics, Proteomics & Bioinformatics*. Elsevier, 14(5), pp. 265–279. doi: 10.1016/J.GPB.2016.05.004.
Lu, X. *et al.* (2014) 'The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity', *Nature Structural & Molecular Biology 2014 21:4*. Nature Publishing Group, 21(4), pp. 423–425. doi: 10.1038/nsmb.2799.
Macfarlane, C. M. and Badge, R. M. (2015) 'Genome-wide amplification of proviral sequences reveals new polymorphic HERV-K(HML-2) proviruses in humans and chimpanzees that are absent from genome assemblies', *Retrovirology*. BioMed Central, 12(1). doi: 10.1186/S12977-015-0162-8.
Magiorkinis, G., Belshaw, R. and Katzourakis, A. (2013) '"There and back again": revisiting the pathophysiological roles of human endogenous retroviruses in the post-genomic era.', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*. The Royal Society, 368(1626), p. 20120504. doi: 10.1098/rstb.2012.0504.
Majumder, V. *et al.* (2018) 'TDP-43 as a potential biomarker for amyotrophic lateral sclerosis: A systematic review and meta-analysis', *BMC Neurology*. BioMed Central Ltd., 18(1). doi: 10.1186/s12883-018-1091-7.
Mameli, G. *et al.* (2009) 'Novel reliable real-time PCR for differential detection of MSRVenv and syncytin-1 in RNA and DNA from patients with multiple sclerosis', *Journal of Virological Methods*. Elsevier, 161(1), pp. 98–106. doi: 10.1016/j.jviromet.2009.05.024.
Mameli, G. *et al.* (2013) 'Activation of MSRV-type endogenous retroviruses during infectious mononucleosis and Epstein-Barr virus latency: The missing link with multiple sclerosis?', *PLoS ONE*. Public Library of Science, 8(11), p. e78474. doi: 10.1371/journal.pone.0078474.
Manca, M. A. *et al.* (2022) 'HERV-K and HERV-H Env Proteins Induce a Humoral Response in Prostate Cancer Patients', *Pathogens*. MDPI, 11(1). doi: 10.3390/PATHOGENS11010095/S1.
Mangeney, M. *et al.* (2001) 'The full-length envelope of an HERV-H human endogenous retrovirus has immunosuppressive properties', *Journal of General Virology*. Society for General Microbiology, 82(10), pp. 2515–2518. doi: 10.1099/0022-1317-82-10-

2515/CITE/REFWORKS.

Manghera, M. and Douville, R. N. (2013) 'Endogenous retrovirus-K promoter: a landing strip for inflammatory transcription factors?', *Retrovirology 2013 10:1*. BioMed Central, 10(1), pp. 1–11. doi: 10.1186/1742-4690-10-16.

Manghera, M., Ferguson-Parry, J. and Douville, R. N. (2016) 'TDP-43 regulates endogenous retrovirus-K viral protein accumulation', *Neurobiology of Disease*. Academic Press, 94, pp. 226–236. doi: 10.1016/J.NBD.2016.06.017.

Manto, M. *et al.* (2012) 'Consensus paper: Roles of the cerebellum in motor control-the diversity of ideas on cerebellar involvement in movement', in *Cerebellum*. NIH Public Access, pp. 457–487. doi: 10.1007/s12311-011-0331-9.

Marini, C. *et al.* (2018) 'Interplay between spinal cord and cerebral cortex metabolism in amyotrophic lateral sclerosis', *Brain*. Oxford University Press, 141(8), pp. 2272–2279. doi: 10.1093/brain/awy152.

Mason, M. J. *et al.* (2014) 'Low HERV-K(C4) copy number is associated with type 1 diabetes.', *Diabetes*. American Diabetes Association, 63(5), pp. 1789–95. doi: 10.2337/db13-1382.

Mayer, J. *et al.* (2018) 'Transcriptional profiling of HERV-K(HML-2) in amyotrophic lateral sclerosis and potential implications for expression of HML-2 proteins', *Molecular Neurodegeneration*. BioMed Central, 13(1), p. 39. doi: 10.1186/s13024-018-0275-3.

McCormick, A. L. *et al.* (2008) 'Quantification of reverse transcriptase in ALS and elimination of a novel retroviral candidate.', *Neurology*. American Academy of Neurology, 70(4), pp. 278–83. doi: 10.1212/01.wnl.0000297552.13219.b4.

Mejzini, R. *et al.* (2019) 'ALS Genetics, Mechanisms, and Therapeutics: Where Are We Now?', *Frontiers in Neuroscience*. Frontiers Media S.A. doi: 10.3389/fnins.2019.01310.

Melnick, M. *et al.* (2021) 'Application of a bioinformatic pipeline to RNA-seq data identifies novel virus-like sequence in human blood', *G3: Genes|Genomes|Genetics*. Oxford University Press, 11(9). doi: 10.1093/G3JOURNAL/JKAB141.

El Mendili, M. M. *et al.* (2014) 'Multi-parametric spinal cord MRI as potential progression marker in amyotrophic lateral sclerosis', *PLoS ONE*. Public Library of Science, 9(4). doi: 10.1371/journal.pone.0095516.

Messeguer, X. *et al.* (2002) 'PROMO: detection of known transcription regulatory elements using species-tailored searches', *Bioinformatics*. Oxford Academic, 18(2), pp.

333–334. doi: 10.1093/BIOINFORMATICS/18.2.333.
Monde, K. *et al.* (2012) 'Human Endogenous Retrovirus K Gag Coassembles with HIV-1 Gag and Reduces the Release Efficiency and Infectivity of HIV-1', *Journal of Virology*. American Society for Microbiology, 86(20), pp. 11194–11208. doi: 10.1128/JVI.00301-12.
Montesion, M. *et al.* (2017) 'Mechanisms of HERV-K (HML-2) Transcription during Human Mammary Epithelial Cell Transformation'. doi: 10.1128/JVI.01258-17.
Mootha, V. K. *et al.* (2003) 'PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes', *Nature Genetics 2003 34:3*. Nature Publishing Group, 34(3), pp. 267–273. doi: 10.1038/ng1180.
Morandi, E. *et al.* (2017) 'The association between human endogenous retroviruses and multiple sclerosis: A systematic review and meta-analysis', *PLoS ONE*. Edited by K. Ruprecht. Public Library of Science, p. e0172415. doi: 10.1371/journal.pone.0172415.
Morandi, E., Tarlinton, R. E. and Gran, B. (2015) 'Multiple sclerosis between genetics and infections: Human endogenous retroviruses in monocytes and macrophages', *Frontiers in Immunology*. Frontiers Media SA, p. 647. doi: 10.3389/fimmu.2015.00647.
Morris, G. *et al.* (2019) 'Do Human Endogenous Retroviruses Contribute to Multiple Sclerosis, and if So, How?', *Molecular neurobiology*. 2018/07/25. Springer US, 56(4), pp. 2590–2605. doi: 10.1007/s12035-018-1255-x.
Mortelmans, K., Wang-Johanning, F. and Johanning, G. L. (2016) 'The role of human endogenous retroviruses in brain development and function', *APMIS*. Wiley/Blackwell (10.1111), pp. 105–115. doi: 10.1111/apm.12495.
Nagy, C. *et al.* (2015) 'Effects of Postmortem Interval on Biomolecule Integrity in the Brain', *Journal of Neuropathology & Experimental Neurology*. Narnia, 74(5), pp. 459–469. doi: 10.1097/NEN.0000000000000190.
Nardo, G. *et al.* (2011) 'Amyotrophic Lateral Sclerosis Multiprotein Biomarkers in Peripheral Blood Mononuclear Cells', *PLoS ONE*. Edited by H. Pant. Public Library of Science, 6(10), p. e25545. doi: 10.1371/journal.pone.0025545.
Nath, A. *et al.* (2015) 'First international workshop on human endogenous retroviruses and diseases, HERVs & disease 2015', *Mobile DNA*. BioMed Central, 6(1), p. 20. doi: 10.1186/s13100-015-0051-7.
Naumova, O. Y. *et al.* (2013) 'Gene expression in the human brain: the current state of the study of specificity and spatiotemporal dynamics.', *Child development*. NIH Public Access,

84(1), pp. 76–88. doi: 10.1111/cdev.12014.

Nevalainen, T. *et al.* (2018) 'Aging-associated patterns in the expression of human endogenous retroviruses', *PLOS ONE*. Edited by K. Roemer. Public Library of Science, 13(12), p. e0207407. doi: 10.1371/journal.pone.0207407.

Nguyen, H. P., Van Broeckhoven, C. and van der Zee, J. (2018) 'ALS Genes in the Genomic Era and their Implications for FTD', *Trends in Genetics*, 28 March. doi: 10.1016/j.tig.2018.03.001.

Nógrádi, A. and Vrbová, G. (2013) 'Anatomy and Physiology of the Spinal Cord'. Landes Bioscience.

Norris, F. H. (1977) 'Current status of the search for virus in amyotrophic lateral sclerosis (ALS).', *Neurologia, neurocirugia, psiquiatria*, 18(2-3 Suppl), pp. 443–454. Available at: http://www.ncbi.nlm.nih.gov/pubmed/99678 (Accessed: 27 March 2018).

Ochoa Thomas, E. *et al.* (2020) 'Awakening the dark side: retrotransposon activation in neurodegenerative disorders', *Current Opinion in Neurobiology*. Elsevier Ltd, pp. 65–72. doi: 10.1016/j.conb.2020.01.012.

Oeckl, P. *et al.* (2020) 'Proteomics in cerebrospinal fluid and spinal cord suggests UCHL1, MAP2 and GPNMB as biomarkers and underpins importance of transcriptional pathways in amyotrophic lateral sclerosis', *Acta Neuropathologica*. Springer, 139(1), pp. 119–134. doi: 10.1007/s00401-019-02093-x.

Olney, K. C. *et al.* (2020) 'Reference genome and transcriptome informed by the sex chromosome complement of the sample increase ability to detect sex differences in gene expression from RNA-Seq data', *Biology of Sex Differences*. BioMed Central, 11(1), pp. 1–18. doi: 10.1186/S13293-020-00312-9/FIGURES/5.

Oluwole, S. O. A. *et al.* (2007) 'Elevated levels of transcripts encoding a human retroviral envelope protein (syncytin) in muscles from patients with motor neuron disease', *Amyotrophic Lateral Sclerosis*, 8(2), pp. 67–72. doi: 10.1080/17482960600864207.

Oturai, D. B. *et al.* (2016) 'Identification of Suitable Reference Genes for Peripheral Blood Mononuclear Cell Subset Studies in Multiple Sclerosis', *Scandinavian Journal of Immunology*. Blackwell Publishing Ltd, 83(1), pp. 72–80. doi: 10.1111/sji.12391.

Paquin, M. E. *et al.* (2018) 'Spinal cord gray matter atrophy in amyotrophic lateral sclerosis', *American Journal of Neuroradiology*. American Society of Neuroradiology, 39(1), pp. 184–192. doi: 10.3174/ajnr.A5427.

Patel, H., Dobson, R. J. B. and Newhouse, S. J. (2019) 'A meta-Analysis of Alzheimer's disease brain transcriptomic data', *Journal of Alzheimer's Disease*. IOS Press, 68(4), pp. 1635–1656. doi: 10.3233/JAD-181085.

Penna, I. *et al.* (2011) 'Selection of candidate housekeeping genes for normalization in human postmortem brain samples', *International Journal of Molecular Sciences*. Multidisciplinary Digital Publishing Institute (MDPI), 12(9), pp. 5461–5470. doi: 10.3390/ijms12095461.

Perron, H. *et al.* (2009) 'Endogenous retroviral genes, Herpesviruses and gender in Multiple Sclerosis', *Journal of the Neurological Sciences*. Elsevier, 286(1–2), pp. 65–72. doi: 10.1016/J.JNS.2009.04.034.

Petriccione, M. *et al.* (2015) 'Reference gene selection for normalization of RT-qPCR gene expression data from Actinidia deliciosa leaves infected with Pseudomonas syringae pv. Actinidiae', *Scientific Reports*. Nature Publishing Group, 5, p. 16961. doi: 10.1038/srep16961.

Pfaffl, Michael W. (2001) 'A new mathematical model for relative quantification in real-time RT–PCR', *Nucleic Acids Research*. Oxford University Press, 29(9), p. e45. doi: 10.1093/NAR/29.9.E45.

Pfaffl, M. W. (2001) 'A new mathematical model for relative quantification in real-time RT-PCR.', *Nucleic Acids Research*. Oxford University Press, 29(9), p. e45. doi: 10.1093/nar/29.9.e45.

Pfaffl, M. W. *et al.* (2004) 'Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper – Excel-based tool using pair-wise correlations', *Biotechnology Letters*. Kluwer Academic Publishers, 26(6), pp. 509–515. doi: 10.1023/B:BILE.0000019559.84305.47.

Pfaffl, M. W., Vandesompele, J. and Kubista, M. (2009) *Real-Time PCR: Current Technology and Applications.*, *Real-Time PCR: Current Technology and Applications.* Caister Academic Press. doi: 10.1016/j.molimm.2014.08.001.

Phan, K. *et al.* (2021) 'Pathological manifestation of human endogenous retrovirus K in frontotemporal dementia', *Communications medicine*. NIH Public Access, 1(1). doi: 10.1038/S43856-021-00060-W.

Picardi, E. *et al.* (2012) 'A Novel Computational Strategy to Identify A-to-I RNA Editing Sites by RNA-Seq Data: De Novo Detection in Human Spinal Cord Tissue', *PLoS ONE*. Public

Library of Science, 7(9). doi: 10.1371/journal.pone.0044184.
Pisano, M. P. *et al.* (2019a) 'Comprehensive Characterization of the Human Endogenous Retrovirus HERV-K(HML-6) Group: Overview of Structure, Phylogeny, and Contribution to the Human Genome', *Journal of Virology*. American Society for Microbiology (ASM), 93(16). doi: 10.1128/JVI.00110-19.
Pisano, M. P. *et al.* (2019b) 'Comprehensive Characterization of the Human Endogenous Retrovirus HERV-K(HML-6) Group: Overview of Structure, Phylogeny, and Contribution to the Human Genome', *Journal of Virology*. American Society for Microbiology (ASM), 93(16). doi: 10.1128/JVI.00110-19.
Pollard, M. O. *et al.* (2018) 'Long reads: their purpose and place', *Human Molecular Genetics*. Narnia, 27(R2), pp. R234–R241. doi: 10.1093/hmg/ddy177.
Potter, S. C. *et al.* (2018) 'HMMER web server: 2018 update', *Nucleic Acids Research*. Oxford Academic, 46(W1), pp. W200–W204. doi: 10.1093/NAR/GKY448.
Preston, B. D., Poiesz, B. J. and Loeb, L. A. (1988) 'Fidelity of HIV-1 reverse transcriptase.', *Science (New York, N.Y.)*, 242(4882), pp. 1168–71. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2460924 (Accessed: 22 April 2018).
Prudencio, M. *et al.* (2015) 'Distinct brain transcriptome profiles in C9orf72-associated and sporadic ALS', *Nature Neuroscience*. Nature Publishing Group, 18(8), pp. 1175–1182. doi: 10.1038/nn.4065.
Prudencio, M. *et al.* (2017) 'Repetitive element transcripts are elevated in the brain of C9orf72 ALS/FTLD patients', *Human Molecular Genetics*. Oxford University Press, 26(17), pp. 3421–3431. doi: 10.1093/hmg/ddx233.
Purves, D. *et al.* (2001) 'The Premotor Cortex', in *Neuroscience*. Sinauer Associates. doi: 10.1007/s11199-009-9615-7.
Qiu, T. *et al.* (2019) 'Precentral degeneration and cerebellar compensation in amyotrophic lateral sclerosis: A multimodal MRI analysis', *Human Brain Mapping*. John Wiley and Sons Inc., 40(12), pp. 3464–3474. doi: 10.1002/hbm.24609.
Quail, M. *et al.* (2012) 'A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers', *BMC Genomics*. BioMed Central, 13(1), p. 341. doi: 10.1186/1471-2164-13-341.
Rahman, M. R. *et al.* (2019) 'Identification of molecular signatures and pathways common to blood cells and brain tissue of amyotrophic lateral sclerosis patients', *Informatics in*

Medicine Unlocked. Elsevier Ltd, 16. doi: 10.1016/j.imu.2019.100193.

Rebollo, R., Romanish, M. T. and Mager, D. L. (2012) 'Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes', *Annual Review of Genetics*. Annual Reviews, 46(1), pp. 21–42. doi: 10.1146/annurev-genet-110711-155621.

Reiman, M. *et al.* (2017) 'Effects of RNA integrity on transcript quantification by total RNA sequencing of clinically collected human placental samples', *The FASEB Journal*. FASEB, 31(8), pp. 3298–3308. doi: 10.1096/fj.201601031RR.

Rhoads, A. and Au, K. F. (2015) 'PacBio Sequencing and Its Applications', *Genomics, Proteomics & Bioinformatics*. Elsevier, 13(5), pp. 278–289. doi: 10.1016/J.GPB.2015.08.002.

Roberts, J. D., Bebenek, K. and Kunkel, T. A. (1988) 'The accuracy of reverse transcriptase from HIV-1.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 242(4882), pp. 1171–3. doi: 10.1126/science.2460925.

Röhn, G. *et al.* (2018) 'ACTB and SDHA are suitable endogenous reference genes for gene expression studies in human astrocytomas using quantitative RT-PCR', *Technology in Cancer Research and Treatment*. SAGE Publications Inc., 17. doi: 10.1177/1533033818802318.

Rollins, B. *et al.* (2010) 'Analysis of whole genome biomarker expression in blood and brain', *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, 153(4), pp. 919–936. doi: 10.1002/ajmg.b.31062.

Romano, G., Klima, R. and Feiguin, F. (2020) 'TDP-43 prevents retrotransposon activation in the Drosophila motor system through regulation of Dicer-2 activity', *BMC Biology*. BioMed Central, 18(1). doi: 10.1186/S12915-020-00816-1.

Rotem, N. *et al.* (2017) 'ALS along the Axons - Expression of coding and noncoding RNA differs in axons of ALS models', *Scientific Reports*. Nature Publishing Group, 7(1), pp. 1–17. doi: 10.1038/srep44500.

Rowland, L. P. and Shneider, N. A. (2001) 'Amyotrophic Lateral Sclerosis', *New England Journal of Medicine*. Massachusetts Medical Society, 344(22), pp. 1688–1700. doi: 10.1056/NEJM200105313442207.

Rutter, L. *et al.* (2019) 'Visualization methods for differential expression analysis', *BMC Bioinformatics*. BioMed Central Ltd., 20(1), p. 458. doi: 10.1186/s12859-019-2968-1.

Rydbirk, R. *et al.* (2016) 'Assessment of brain reference genes for RT-qPCR studies in

neurodegenerative diseases', *Scientific Reports*. Nature Publishing Group, 6(1), p. 37116. doi: 10.1038/srep37116.

Safe, S. *et al.* (2014) 'Transcription factor Sp1, also known as specificity protein 1 as a therapeutic target', *http://dx.doi.org/10.1517/14728222.2014.914173*. Taylor & Francis, 18(7), pp. 759–769. doi: 10.1517/14728222.2014.914173.

Saito, T. *et al.* (2017) 'Upregulation of Human Endogenous Retrovirus-K Is Linked to Immunity and Inflammation in Pulmonary Arterial Hypertension.', *Circulation*. American Heart Association, Inc., 136(20), pp. 1920–1935. doi: 10.1161/CIRCULATIONAHA.117.027589.

Savage, A. L. *et al.* (2018) 'Retrotransposons in the development and progression of amyotrophic lateral sclerosis.', *Journal of neurology, neurosurgery, and psychiatry*. BMJ Publishing Group Ltd, p. jnnp-2018-319210. doi: 10.1136/jnnp-2018-319210.

Schmahmann, J. D. and Caplan, D. (2006) 'Cognition, emotion and the cerebellum.', *Brain : a journal of neurology*, pp. 290–292. doi: 10.1093/brain/awh729.

Schmitt, K. *et al.* (2015) 'HERV-K(HML-2) Rec and Np9 transcripts not restricted to disease but present in many normal human tissues', *Mobile DNA*. BioMed Central, 6(1), p. 4. doi: 10.1186/s13100-015-0035-7.

Schmittgen, T. D. and Livak, K. J. (2008) 'Analyzing real-time PCR data by the comparative CT method', *Nature Protocols*. Nature Publishing Group, 3(6), pp. 1101–1108. doi: 10.1038/nprot.2008.73.

Schultz, J. *et al.* (1998) 'SMART, a simple modular architecture research tool: Identification of signaling domains', *Proceedings of the National Academy of Sciences of the United States of America*, 95(11), pp. 5857–5864. doi: 10.1073/PNAS.95.11.5857/ASSET/7CA4C22C-EFFB-4036-BB0D-263ECE5D1AB3/ASSETS/GRAPHIC/PQ0980595003.JPEG.

Scientific, T. F. (2017) 'Ion 520™ & Ion 530™ Kit – Chef USER GUIDE', pp. 1–95. Available at: https://www.thermofisher.com/document-connect/document-connect.html?url=https%3A%2F%2Fassets.thermofisher.com%2FTFS-Assets%2FLSG%2Fmanuals%2FMAN0015805_Ion520_530ExTKit_UG.pdf&title=VXNlciBHdWlkZTogSW9uIDUyMCBhbmQgSW9uIDUzMCBFeFQgS2l0IC0gQ2hlZg== (Accessed: 24 April 2019).

Sebastián-Martín, A., Barrioluengo, V. and Menéndez-Arias, L. (2018) 'Transcriptional

inaccuracy threshold attenuates differences in RNA-dependent DNA synthesis fidelity between retroviral reverse transcriptases', *Scientific Reports 2018 8:1*. Nature Publishing Group, 8(1), pp. 1–13. doi: 10.1038/s41598-017-18974-8.

Sexton, C. E., Tillett, R. L. and Han, M. V. (2021) 'The essential but enigmatic regulatory role of HERVH in pluripotency', *Trends in Genetics*. Elsevier Current Trends. doi: 10.1016/J.TIG.2021.07.007.

Shin, W. *et al.* (2013) 'Human-Specific HERV-K Insertion Causes Genomic Variations in the Human Genome', *PLoS ONE*. Edited by R. Cordaux. Public Library of Science, 8(4), p. e60605. doi: 10.1371/journal.pone.0060605.

Shpyleva, S. *et al.* (2018) 'Overexpression of LINE-1 Retrotransposons in Autism Brain', *Molecular Neurobiology*. Springer US, 55(2), pp. 1740–1749. doi: 10.1007/s12035-017-0421-x.

Siddique, T. and Ajroud-Driss, S. (2011) 'Familial amyotrophic lateral sclerosis, a historical perspective.', *Acta myologica : myopathies and cardiomyopathies : official journal of the Mediterranean Society of Myology*. Pacini Editore, 30(2), pp. 117–20. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22106714 (Accessed: 3 December 2018).

Sidransky, E. (2006) 'Heterozygosity for a Mendelian disorder as a risk factor for complex disease', *Clinical Genetics*. John Wiley & Sons, Ltd, 70(4), pp. 275–282. doi: 10.1111/J.1399-0004.2006.00688.X.

Siegel, A. and Sapru, H. N. (2019) *Essential Neuroscience*. Fourth. Philadelphia: Wolters Kluwer.

Sievers, F. and Higgins, D. G. (2018) 'Clustal Omega for making accurate alignments of many protein sequences', *Protein Science*. John Wiley & Sons, Ltd, 27(1), pp. 135–145. doi: 10.1002/pro.3290.

Silver, N. *et al.* (2006) 'Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR.', *BMC molecular biology*. BioMed Central, 7, p. 33. doi: 10.1186/1471-2199-7-33.

Simoneau, J. *et al.* (2021) 'Current RNA-seq methodology reporting limits reproducibility', *Briefings in Bioinformatics*. Oxford Academic, 22(1), pp. 140–145. doi: 10.1093/BIB/BBZ124.

Simula, E. R. *et al.* (2021) 'TDP-43 and HERV-K Envelope-Specific Immunogenic Epitopes Are Recognized in ALS Patients', *Viruses*. Multidisciplinary Digital Publishing Institute

(MDPI), 13(11). doi: 10.3390/V13112301.
Singh, S. K. (2007) 'Endogenous retroviruses: suspects in the disease world', *Future Microbiology*, 2(3), pp. 269–275. doi: 10.2217/17460913.2.3.269.
Slokar, G. and Hasler, G. (2016) 'Human endogenous retroviruses as pathogenic factors in the development of schizophrenia', *Frontiers in Psychiatry*. Frontiers Media SA, p. 183. doi: 10.3389/fpsyt.2015.00183.
Smith, C. J. and Osborn, A. M. (2009) 'Advantages and limitations of quantitative PCR (Q-PCR)-based approaches in microbial ecology', *FEMS Microbiology Ecology*. John Wiley & Sons, Ltd (10.1111), 67(1), pp. 6–20. doi: 10.1111/j.1574-6941.2008.00629.x.
Smyth, L. C. D. *et al.* (2018) 'Markers for human brain pericytes and smooth muscle cells', *Journal of Chemical Neuroanatomy*. Elsevier, 92, pp. 48–60. doi: 10.1016/J.JCHEMNEU.2018.06.001.
Sonntag, K.-C. *et al.* (2016) 'Limited predictability of postmortem human brain tissue quality by RNA integrity numbers.', *Journal of neurochemistry*. NIH Public Access, 138(1), pp. 53–9. doi: 10.1111/jnc.13637.
Sreedharan, S. P., Kumar, A. and Giridhar, P. (2018) 'Primer design and amplification efficiencies are crucial for reliability of quantitative PCR studies of caffeine biosynthetic N-methyltransferases in coffee.', *3 Biotech*. Springer, 8(11), p. 467. doi: 10.1007/s13205-018-1487-5.
Stan, A. D. *et al.* (2006) 'Human postmortem tissue: What quality markers matter?', *Brain Research*. NIH Public Access, 1123(1), pp. 1–11. doi: 10.1016/j.brainres.2006.09.025.
Steele, A. J. *et al.* (2005) 'Detection of serum reverse transcriptase activity in patients with ALS and unaffected blood relatives', *Neurology*. American Academy of Neurology, 64(3), pp. 454–458. doi: 10.1212/01.WNL.0000150899.76130.71.
Steinbaugh, M. J. *et al.* (2018) 'BcbioRNAseq: R package for bcbio rna-seq analysis [version 2; peer review: 1 approved, 1 approved with reservations]', *F1000Research*. F1000 Research Ltd, 7, p. 1976. doi: 10.12688/F1000RESEARCH.12093.2.
Stephenson, F. H. and Stephenson, F. H. (2016) 'Real-Time PCR', *Calculations for Molecular Biology and Biotechnology*. Academic Press, pp. 215–320. doi: 10.1016/B978-0-12-802211-5.00009-6.
Studer, G. *et al.* (2020) 'QMEANDisCo—distance constraints applied on model quality estimation', *Bioinformatics*. Oxford Academic, 36(6), pp. 1765–1771. doi:

10.1093/BIOINFORMATICS/BTZ828.

Subramanian, A. *et al.* (2005) 'Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles', *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), pp. 15545–15550. doi: 10.1073/PNAS.0506580102/SUPPL_FILE/06580FIG7.JPG.

Subramanian, R. P. *et al.* (2011) 'Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses.', *Retrovirology*. BioMed Central, 8(1), p. 90. doi: 10.1186/1742-4690-8-90.

Sun, Y. *et al.* (2012) 'Pseudogenes as Weaknesses of ACTB (Actb) and GAPDH (Gapdh) Used as Reference Genes in Reverse Transcription and Polymerase Chain Reactions', *PLoS ONE*. Edited by A. Rishi. Public Library of Science, 7(8), p. e41659. doi: 10.1371/journal.pone.0041659.

Suntsova, M. *et al.* (2013) 'Human-specific endogenous retroviral insert serves as an enhancer for the schizophrenia-linked gene PRODH.', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 110(48), pp. 19472–19477. doi: 10.1073/pnas.1318172110.

Suntsova, M. *et al.* (2015) 'Molecular functions of human endogenous retroviruses in health and disease', *Cellular and Molecular Life Sciences*, pp. 3653–3675. doi: 10.1007/s00018-015-1947-6.

Suntsova, M. *et al.* (2019) 'Atlas of RNA sequencing profiles for normal human tissues', *Scientific data*. NLM (Medline), 6(1), p. 36. doi: 10.1038/s41597-019-0043-4.

Svec, D. *et al.* (2015) 'How good is a PCR efficiency estimate: Recommendations for precise and robust qPCR efficiency assessments', *Biomolecular Detection and Quantification*. Elsevier, 3, pp. 9–16. doi: 10.1016/J.BDQ.2015.01.005.

Sverdlov, E. D. (1998) 'Perpetually mobile footprints of ancient infections in human genome', *FEBS Letters*. Wiley-Blackwell, pp. 1–6. doi: 10.1016/S0014-5793(98)00478-5.

Swindell, W. R. *et al.* (2019) 'ALS blood expression profiling identifies new biomarkers, patient subgroups, and evidence for neutrophilia and hypoxia', *Journal of Translational Medicine*. BioMed Central Ltd., 17(1), pp. 1–33. doi: 10.1186/s12967-019-1909-0.

Tam, O. H. *et al.* (2019) 'Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia', *Cell Reports*. Elsevier B.V., 29(5), pp. 1164-1177.e5. doi: 10.1016/j.celrep.2019.09.066.

Tam, O. H., Ostrow, L. W. and Gale Hammell, M. (2019) 'Diseases of the nERVous system: Retrotransposon activity in neurodegenerative disease', *Mobile DNA*. BioMed Central Ltd. doi: 10.1186/s13100-019-0176-1.

Tamaki, Y. *et al.* (2018) 'Elimination of TDP-43 inclusions linked to amyotrophic lateral sclerosis by a misfolding-specific intrabody with dual proteolytic signals', *Scientific Reports*. Nature Publishing Group, 8(1), p. 6030. doi: 10.1038/s41598-018-24463-3.

Taylor, C. F. *et al.* (2007) 'The minimum information about a proteomics experiment (MIAPE)', *Nature Biotechnology*. Nature Publishing Group, 25(8), pp. 887–893. doi: 10.1038/nbt1329.

Taylor, C. F. *et al.* (2008) 'Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project.', *Nature biotechnology*. NIH Public Access, 26(8), pp. 889–96. doi: 10.1038/nbt.1411.

Taylor, S. *et al.* (2010) 'A practical approach to RT-qPCR—Publishing data that conform to the MIQE guidelines', *Methods*. Academic Press, 50(4), pp. S1–S5. doi: 10.1016/J.YMETH.2010.01.005.

Taylor, S. C. *et al.* (2019) 'The Ultimate qPCR Experiment: Producing Publication Quality, Reproducible Data the First Time', *Trends in Biotechnology*. Elsevier Current Trends. doi: 10.1016/J.TIBTECH.2018.12.002.

Thomas, J., Perron, H. and Feschotte, C. (2018) 'Variation in proviral content among human genomes mediated by LTR recombination', *Mobile DNA 2018 9:1*. BioMed Central, 9(1), pp. 1–15. doi: 10.1186/S13100-018-0142-3.

Tokuyama, M. *et al.* (2018a) 'ERVmap analysis reveals genome-wide transcription of human endogenous retroviruses.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 115(50), pp. 12565–12572. doi: 10.1073/pnas.1814589115.

Tokuyama, M. *et al.* (2018b) 'ERVmap analysis reveals genome-wide transcription of human endogenous retroviruses.', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 115(50), pp. 12565–12572. doi: 10.1073/PNAS.1814589115.

Tortarolo, M. *et al.* (2017) 'Amyotrophic Lateral Sclerosis, a Multisystem Pathology: Insights into the Role of TNF α', *Mediators of Inflammation*. Hindawi Limited. doi: 10.1155/2017/2985051.

Touchberry, C. D. *et al.* (2006) 'Age-related changes in relative expression of real-time PCR housekeeping genes in human skeletal muscle.', *Journal of biomolecular techniques : JBT*. The Association of Biomolecular Resource Facilities, 17(2), pp. 157–62. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16741243 (Accessed: 26 October 2018).

Toufaily, C. *et al.* (2011) 'Activation of LTRs from different human endogenous retrovirus (HERV) families by the HTLV-1 tax protein and T-cell activators', *Viruses*. Molecular Diversity Preservation International, 3(11), pp. 2146–2159. doi: 10.3390/v3112146.

Treangen, T. J. and Salzberg, S. L. (2012) 'Repetitive DNA and next-generation sequencing: Computational challenges and solutions', *Nature Reviews Genetics*. NIH Public Access, pp. 36–46. doi: 10.1038/nrg3117.

Turner, G. *et al.* (2001) 'Insertional polymorphisms of full-length endogenous retroviruses in humans', *Current Biology*. Cell Press, 11(19), pp. 1531–1535. doi: 10.1016/S0960-9822(01)00455-9.

Umoh, M. E. *et al.* (2016) 'Comparative analysis of C9orf72 and sporadic disease in an ALS clinic population', *Neurology*. Lippincott Williams and Wilkins, 87(10), pp. 1024–1030. doi: 10.1212/WNL.0000000000003067.

Usarek, E. *et al.* (2017) 'Validation of qPCR reference genes in lymphocytes from patients with amyotrophic lateral sclerosis', *PLOS ONE*. Edited by C. Cereda. Public Library of Science, 12(3), p. e0174317. doi: 10.1371/journal.pone.0174317.

Valko, K. and Ciesla, L. (2019) 'Amyotrophic lateral sclerosis', in Longo, D. L. (ed.) *Progress in Medicinal Chemistry*, pp. 63–117. doi: 10.1016/bs.pmch.2018.12.001.

Vandesompele, J. *et al.* (2002) 'Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes.', *Genome biology*. BioMed Central, 3(7). doi: 10.1186/gb-2002-3-7-research0034.

Vargiu, L. *et al.* (2016) 'Classification and characterization of human endogenous retroviruses; mosaic forms are common.', *Retrovirology*. BioMed Central, 13, p. 7. doi: 10.1186/s12977-015-0232-y.

Vijayakumar, U. G. *et al.* (2019) 'A systematic review of suggested molecular strata, biomarkers and their tissue sources in ALS', *Frontiers in Neurology*. Frontiers Media S.A., 10(MAY). doi: 10.3389/fneur.2019.00400.

Viola, M. V *et al.* (1975) 'RNA-instructed DNA polymerase activity in a cytoplasmic particulate fraction in brains from Guamanian patients.', *The Journal of experimental*

medicine. The Rockefeller University Press, 142(2), pp. 483–94. doi: 10.1084/jem.142.2.483.

Wallace, T. A. *et al.* (2014) 'Elevated HERV-K mRNA expression in PBMC is associated with a prostate cancer diagnosis particularly in older men and smokers', *Carcinogenesis*. Carcinogenesis, 35(9), pp. 2074–2083. doi: 10.1093/CARCIN/BGU114.

Wang, B. *et al.* (2016) 'Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing.', *Nature communications*. Nature Publishing Group, 7, p. 11708. doi: 10.1038/ncomms11708.

Wang, W. *et al.* (2019) 'The CCAAT/Enhancer-Binding Protein Family: Its Roles in MDSC Expansion and Function', *Frontiers in Immunology*. Frontiers, 10, p. 1804. doi: 10.3389/FIMMU.2019.01804.

Wang, X., Huang, J. and Zhu, F. (2018) 'Human Endogenous Retroviral Envelope Protein Syncytin-1 and Inflammatory Abnormalities in Neuropsychological Diseases.', *Frontiers in psychiatry*. Frontiers Media SA, 9, p. 422. doi: 10.3389/fpsyt.2018.00422.

Weber, M. *et al.* (2021) 'Increased HERV-K(HML-2) Transcript Levels Correlate with Clinical Parameters of Liver Damage in Hepatitis C Patients', *Cells*. Multidisciplinary Digital Publishing Institute (MDPI), 10(4). doi: 10.3390/CELLS10040774.

Weis, S. *et al.* (2007) 'Quality control for microarray analysis of human brain samples: The impact of postmortem factors, RNA characteristics, and histopathology', *Journal of Neuroscience Methods*. Elsevier, 165(2), pp. 198–209. doi: 10.1016/J.JNEUMETH.2007.06.001.

Whiteford, N. *et al.* (2005) 'An analysis of the feasibility of short read sequencing.', *Nucleic acids research*. Oxford University Press, 33(19), p. e171. doi: 10.1093/nar/gni170.

Wierschke, S. *et al.* (2010) 'Evaluating reference genes to normalize gene expression in human epileptogenic brain tissues', *Biochemical and Biophysical Research Communications*. Academic Press, 403(3–4), pp. 385–390. doi: 10.1016/J.BBRC.2010.10.138.

Wildschutte, J. H. *et al.* (2016) 'From the Cover: PNAS Plus: Discovery of unfixed endogenous retrovirus insertions in diverse human populations', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 113(16), p. E2326. doi: 10.1073/PNAS.1602336113.

Wilson, D. B., Dorfman, D. M. and Orkin, S. H. (1990) 'A nonerythroid GATA-binding

protein is required for function of the human preproendothelin-1 promoter in endothelial cells', *Molecular and Cellular Biology*. American Society for Microbiology, 10(9), pp. 4854–4862. doi: 10.1128/MCB.10.9.4854-4862.1990.

Xiao, T., Li, X. and Felsenfeld, G. (2021) 'The Myc-associated zinc finger protein (MAZ) works together with CTCF to control cohesin positioning and genome organization', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 118(7). doi: 10.1073/PNAS.2023127118.

Xie, F. *et al.* (2012) 'miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs', *Plant Molecular Biology*. Springer Netherlands, 80(1), pp. 75–84. doi: 10.1007/s11103-012-9885-2.

Xue, B., Sechi, L. A. and Kelvin, D. J. (2020) 'Human Endogenous Retrovirus K (HML-2) in Health and Disease', *Frontiers in Microbiology*. Frontiers Media SA, 11. doi: 10.3389/FMICB.2020.01690.

Xue, Y. C. *et al.* (2018) 'Enteroviral Infection: The Forgotten Link to Amyotrophic Lateral Sclerosis?', *Frontiers in Molecular Neuroscience*. Frontiers, 11, p. 63. doi: 10.3389/fnmol.2018.00063.

Yamanaka, K. and Komine, O. (2018) 'The multi-dimensional roles of astrocytes in ALS', *Neuroscience Research*. Elsevier Ireland Ltd, pp. 31–38. doi: 10.1016/j.neures.2017.09.011.

Yang, C. *et al.* (2007) 'Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters', *Gene*. 2006/10/10, 389(1), pp. 52–65. doi: 10.1016/j.gene.2006.09.029.

Ye, J. *et al.* (2012) 'Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction.', *BMC bioinformatics*. BioMed Central, 13, p. 134. doi: 10.1186/1471-2105-13-134.

Yi, J.-M., Kim, H.-M. and Kim, H.-S. (2006) 'Human endogenous retrovirus HERV-H family in human tissues and cancer cells: expression, identification, and phylogeny', *Cancer Letters*. Elsevier, 231(2), pp. 228–239. doi: 10.1016/j.canlet.2005.02.001.

Yu, H.-L., Zhao, Z.-K. and Zhu, F. (2013) 'The role of human endogenous retroviral long terminal repeat sequences in human cancer (Review)', *Int J Mol Med*. Department of Medical Microbiology, School of Medicine, Wuhan University, Wuhan, Hubei 430071, P.R. China Department of Urology, Zhongnan Hospital, Wuhan University, Wuhan, Hubei

430071, P.R. China, 32(4), pp. 755–762. doi: 10.3892/ijmm.2013.1460.
Zhang, F. *et al.* (2018) 'Altered white matter microarchitecture in amyotrophic lateral sclerosis: A voxel-based meta-analysis of diffusion tensor imaging', *NeuroImage: Clinical*. Elsevier, 19, pp. 122–129. doi: 10.1016/J.NICL.2018.04.005.
Zhang, M., Liang, J. Q. and Zheng, S. (2019) 'Expressional activation and functional roles of human endogenous retroviruses in cancers', *Reviews in Medical Virology*. John Wiley & Sons, Ltd, p. e2025. doi: 10.1002/rmv.2025.
Zhang, T. *et al.* (2020) 'The E-Twenty-Six Family in Hepatocellular Carcinoma: Moving into the Spotlight', *Frontiers in Oncology*. Frontiers Media SA, 10. doi: 10.3389/FONC.2020.620352.
Zhao, J. *et al.* (2011) 'Expression of Human Endogenous Retrovirus Type K Envelope Protein is a Novel Candidate Prognostic Marker for Human Breast Cancer', *Genes & cancer*. Genes Cancer, 2(9), pp. 914–922. doi: 10.1177/1947601911431841.
Zhao, X. *et al.* (2018) 'Reference Gene Selection for Quantitative Real-Time PCR of Mycelia from Lentinula edodes under High-Temperature Stress.', *BioMed research international*. Hindawi Limited, 2018, p. 1670328. doi: 10.1155/2018/1670328.
Zhou, F. *et al.* (2015) 'Chimeric antigen receptor T cells targeting HERV-K inhibit breast cancer and its metastasis through downregulation of Ras', *OncoImmunology*. Taylor & Francis, 4(11), p. e1047582. doi: 10.1080/2162402X.2015.1047582.
Zucca, S. *et al.* (2019) 'RNA-Seq profiling in peripheral blood mononuclear cells of amyotrophic lateral sclerosis patients and controls', *Scientific Data*. Nature Publishing Groups, 6(1), p. 190006. doi: 10.1038/sdata.2019.6.
Zuo, X. *et al.* (2021) 'TDP-43 aggregation induced by oxidative stress causes global mitochondrial imbalance in ALS', *Nature Structural & Molecular Biology 2021 28:2*. Nature Publishing Group, 28(2), pp. 132–142. doi: 10.1038/s41594-020-00537-7.