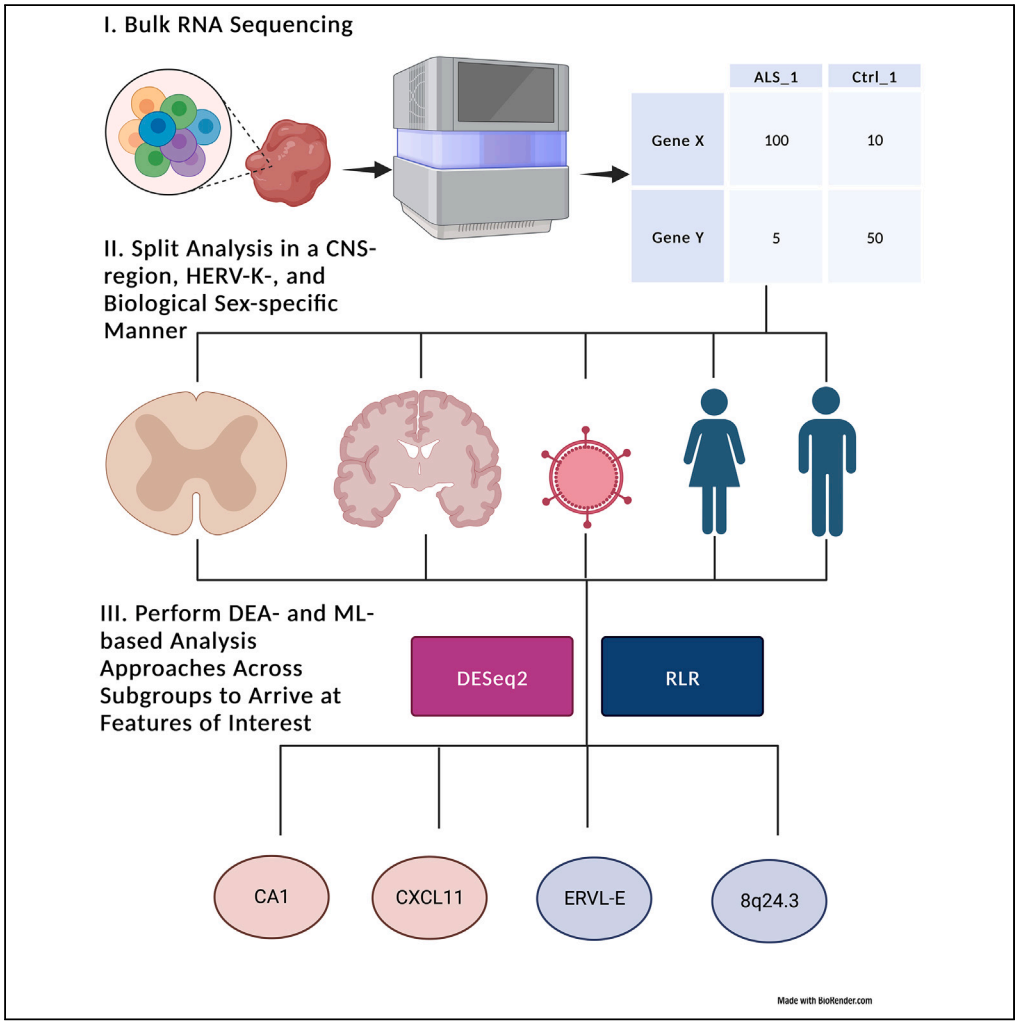
iScience

CellPress
OPEN ACCESS

Article

Endogenous retroviruses are dysregulated in ALS



Nicholas Pasternack, Tara Doucet-O'Hare, Kory Johnson, NYGC Consortium, Ole Paulsen, Avindra Nath

avindra.nath@nih.gov

Highlights

HML-2 loci 1q22 and 8p23.1 were upregulated in the spinal cord of ALS patients

A high HERV-K expressing ALS subgroup (20% of patients) was identified

8p23.1 expression is associated with the expression of ALS-associated genes

CA1, CXCL11, and ERVL-E could be important biomarkers of ALS

Pasternack et al., iScience 27, 110147
July 19, 2024
https://doi.org/10.1016/j.isci.2024.110147

iScience

CellPress
OPEN ACCESS

Article

Endogenous retroviruses are dysregulated in ALS

Nicholas Pasternack,[1,5] Tara Doucet-O'Hare,[2] Kory Johnson,[3] NYGC Consortium,[4] Ole Paulsen,[5] and Avindra Nath[1,6,*]

SUMMARY

Amyotrophic lateral sclerosis (ALS) is a universally fatal neurodegenerative disease with no cure. Human endogenous retroviruses (HERVs) have been implicated in its pathogenesis but their relevance to ALS is not fully understood. We examined bulk RNA-seq data from almost 2,000 ALS and unaffected control samples derived from the cortex and spinal cord. Using different methods of feature selection, including differential expression analysis and machine learning, we discovered that transcription of HERV-K loci 1q22 and 8p23.1 were significantly upregulated in the spinal cord of individuals with ALS. Additionally, we identified a subset of ALS patients with upregulated HERV-K expression in the cortex and spinal cord. We also found the expression of HERV-K loci 19q11 and 8p23.1 was correlated with protein coding genes previously implicated in ALS and dysregulated in ALS patients in this study. These results clarify the association of HERV-K and ALS and highlight specific genes in the pathobiology of late-stage ALS.

INTRODUCTION

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease that leads to the death of both upper motor neurons (UMN) and lower motor neurons (LMN). Current FDA-approved drugs have neuroprotective properties but have only a minimal impact on disease progression. The heritability and impact of genetic predisposition toward developing ALS in patients with no known familial cause of disease (sporadic ALS) is still a matter of ongoing research. However, the heritability of sporadic ALS has been reported to be between 17% and 37%.[1–3] On the other hand, we may be significantly underestimating ALS heritability since commonly used next generation sequencing (NGS) technologies and downstream analysis pipelines can fail to identify certain types of genetic variations such as structural variants (SV), particularly in genomic regions with highly repetitive sequences.[4] ALS is more common in males, with a male-to-female incidence ratio of 3:1, and sex hormones may mediate ALS pathophysiology.[5]

Human endogenous retroviruses (HERVs) are repetitive elements that comprise up to 8% of the human genome.[6] HERVs possess a similar structure to exogenous retroviruses: *gag*, *pro*, *pol*, and *env* genes sandwiched by two long terminal repeats (LTRs) (especially LTR5_Hs). Detecting HERVs in NGS contexts is challenging, given the ambiguous nature of aligned reads using standard analysis pipelines.[7] However, recent advancements in analysis methods have improved our ability to measure these features in bulk- and single-cell RNA sequencing (sc-RNA) and genomic data.[8–10]

A HERV of particular interest in the field of neurodegenerative diseases is HERV-K (named for the lysine, or K, tRNA primer). HERV-K is a member of the betaretrovirus-like endogenous retroviruses and has 11 subtypes, human endogenous MMTV-like (HML) 1 through 11.[11] HML-2 integrated into the human genome relatively recently in evolutionary history and some loci have been shown to be polymorphic and have functional open reading frames as well as a complete virus with the potential to replicate in humans.[12] Moreover, HERV-K *pol* transcripts were found to be more highly expressed in cortical postmortem brain specimens from ALS patients compared to healthy controls and controls with other systemic and neurodegenerative diseases.[13] Additionally, HERV-K expression was found to be correlated with TAR DNA-binding protein 43 (TDP-43) accumulation, a known hallmark pathology of ALS.[13] In multiple sclerosis and chronic kidney disease, significant differences in the expression of HERVs between patients and controls at the RNA-seq level have been validated via other methods such as quantitative polymerase chain reaction (qPCR) and *in situ* hybridization.[14,15]

There is evidence that the envelope protein of HERV-K/HML-2 (Env) has a causal role in ALS-mediated motor neuron degeneration: a study utilizing a transgenic mouse that overexpresses the HERV-K/HML-2 Env protein found this mouse model recapitulated many of the facets of ALS seen in humans including progressive motor dysfunction, muscle atrophy, and decreased activity of pyramidal neurons.[16] Moreover, the study found that TDP-43 bound to the HERV-K/HML-2 LTR and regulated HERV-K/HML-2 expression.[16] A recent study utilizing Drosophila found that HML-2 was sufficient to induce cytoplasmic aggregation of TDP-43 and ERV transmission led to TDP-43 proteinopathy.[17]

[1]Section of Infections of the Nervous System, National Institute of Neurological Disorders and Stroke (NINDS), National Institutes of Health (NIH), Bethesda, MD, USA
[2]Neuro-Oncology Branch Stem Cell Team, National Cancer Institute (NCI), National Institutes of Health (NIH), Bethesda, MD, USA
[3]Bioinformatics Unit, National Institute of Neurological Disorders and Stroke (NINDS), National Institutes of Health (NIH), Bethesda, MD, USA
[4]New York Genome Center Consortium, New York City, NY, USA
[5]Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, UK
[6]Lead contact
*Correspondence: avindra.nath@nih.gov
https://doi.org/10.1016/j.isci.2024.110147

Whether retroviruses, specifically HML-2, are associated with ALS neuropathophysiology remains an area of active debate within the field. For example, some studies failed to find an association between ALS and HML-2.[18–20]

We hypothesized that there would be a CNS region- and biological sex-specific pattern of transposable element (TE) and Ensembl gene (ENSG) expression. Moreover, we hypothesized that there would be a subset of ALS patients with higher expression of HERV-K/HML-2 transcripts and that this subpopulation could be predicted based on transcriptomic data.

RESULTS

Differential expression analysis in cortex across all ALS patients and controls reveals SLC13A4, CXCL, and IGKC to be relevant for disease

We analyzed bulk RNA sequencing (RNA-seq) data from 476 individuals (46% female) including 297 ALS patients (47% female) and 179 unaffected controls (45% female). About 15% of ALS patients were positive for a *C9ORF72* mutation, 2% were positive for a *SOD1* mutation, and 1% were positive for a *FUS* mutation. In terms of clinical presentation, about 62% of ALS patients had limb onset, 34% had bulbar onset, and 3% had axial onset. Multiple samples from different CNS regions were collected from the same individual, so the total number of samples analyzed was 1,710. There were 1,250 ALS samples (54% CTX) and 460 control samples (69% CTX). Of the individuals in this dataset that did not have ALS, about 55% had no known neurological disease, 42% had a neurological disease other than ALS, 1% had a motor neuron disease other than ALS, and 2% had no data on their disease status. Of the controls with no known neurological disease, over half had unknown comorbidities, however, tumors and cardiovascular disease were relatively common. Since TEs are poorly detected with standard alignment and enumeration techniques, a tool designed to specifically detect TE features, TEtranscripts, was used[9] to quantify TE-related RNA species. Samples were pooled by CNS region (CTX or SC) and male and female samples were analyzed together and separately. Differential expression analysis (DEA) was performed using DESeq2.[21] Features that were associated with sex (based on DESeq2 adjusted *p*-value) were removed in DEAs performed on both male and female samples.

There were 704 upregulated ERV loci and 1,239 downregulated ERV loci in ALS patients relative to controls. Of these, there were 15 significantly dysregulated HERV-K features (three upregulated and 12 downregulated) including two downregulated Env-coding loci (Figure 1A). The Env-coding loci could only encode a truncated Env protein. The non-Env-coding HERV-K loci mostly related to HERV-K3, HERV-K9, and HERV-K11 features. There were a total of 8 HERV-K Env-coding features (Table S1) significantly dysregulated across all CTX DEAs.

In terms of ENSGs, there were 2,490 downregulated genes, and 1,894 upregulated. *SLC13A4* (Solute Carrier Family 13 Member 4), *IGHA1* (Immunoglobulin Heavy Constant Alpha 1), *IGKC* (Immunoglobulin Kappa Constant), and *CXCL8* (Interleukin 8) were the most significantly dysregulated genes in all ALS vs. unaffected individuals (all downregulated). *SLC13A4* has not previously been implicated in neurological disorders, however, it is essential for neurodevelopment and its gene product transports sulfate from the blood to the CSF. Moreover, it is known to be expressed in the pia mater and choroid plexus.[22] SLC13A4 represents a promising novel target identified in this analysis and no previous study implicating *SLC13A4* in ALS was found. *In vitro* studies have shown that oxidative stress occurs in glial cells after exposure to the sera of individuals with ALS.[23] More recently, serum IGKC has been implicated as a protein-based biomarker for ALS.[24] *CXCL8* encodes for IL-8 which is a proinflammatory cytokine. CSF IL-8 has been shown to be negatively correlated with a metric of the functional status of ALS patients.[25]

Differential expression analysis in cortex across female patients and female controls Reveals SLC13A4, MYH11, and SFRP2 to be relevant for disease

There were 1,274 upregulated ERV loci and 668 downregulated ERV loci. Of these, there were 24 significantly dysregulated HERV-K features (10 upregulated and 14 downregulated) including 7 downregulated Env-coding loci (Figure 1B). In terms of ENSGs, there were 2,614 downregulated genes, and 1,109 upregulated. *SLC13A4* (Solute Carrier Family 13 Member 4), *MYH11* (Myosin-11), and *SFRP2* (Secreted Frizzled Related Protein 2) were the most significantly dysregulated (all downregulated) in female ALS patients vs. female unaffected control individuals. Myosins are known to be expressed in the brain and spinal cord,[26] and mutations in *MYH11* have been linked to cerebrovascular disease[27] but has not been studied in the context of ALS. *SFRP2* has been shown to be dysregulated at the level of astroglia in the cortical gray matter of a mouse model of ALS.[28]

Differential expression analysis in cortex across male patients and male controls Reveals HSPA6 and HML-2 locus 19q11 to be relevant for disease

There were 1,179 upregulated ERV loci and 653 downregulated ERV loci. Of these, there were 12 significantly dysregulated HERV-K features (3 upregulated, and 9 downregulated) including one upregulated Env-coding locus, 19q11 (centromere) (Figure 1C). In terms of ENSGs, there were 511 downregulated genes, and 546 upregulated. The upregulated Env-coding locus is significantly associated with biological sex and was therefore removed prior to the all ALS patient vs. unaffected control analysis. Fewer features met effect size and significance cut-offs in this analysis. *HSPA6* (Heat Shock Protein Family A [Hsp70] Member 6) was most significantly dysregulated (downregulated) in the male ALS vs. unaffected individuals. Hsp70 acts as a chaperone and its down regulation leads to increased protein misfolding in ALS.[29] *HSPA6* has also been implicated in neuronal stress response in a human neuronal cell-based model.[30]

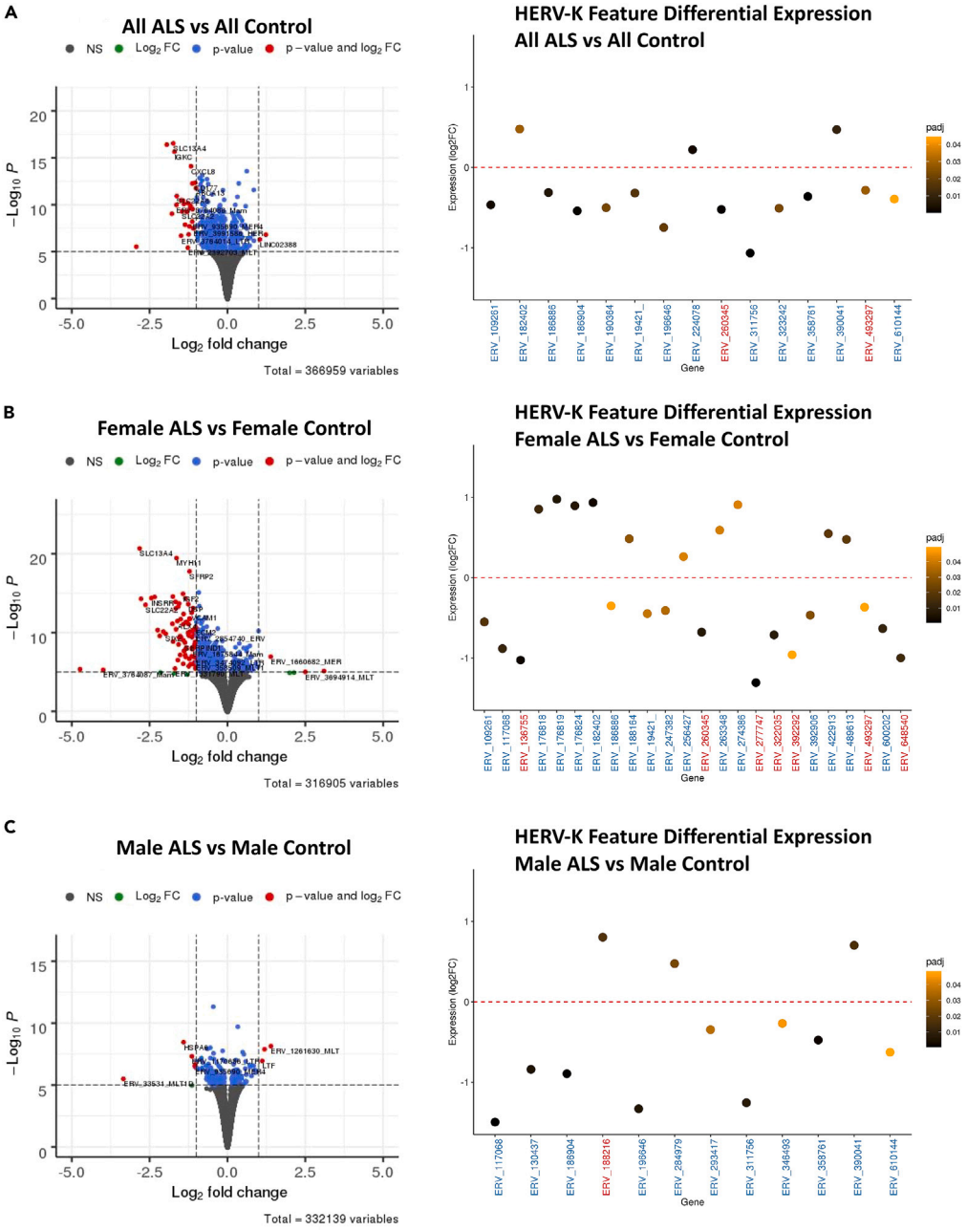


Figure 1. ALS patients vs. unaffected control differential expression analysis (DEA) volcano and scatterplots of HERV-K features in the cortex (CTX)

(A) Volcano plots (left) and scatterplots (right) for all samples, (B) female samples only, and (C) male samples only in the CTX. Features that were significantly associated with biological sex (A only) and did not have sufficient expression were removed prior to the analysis. For volcano plots, log2FC cut-off is >|2| and cut-off for *p* value is 10E-6. Feature plot of significantly differentially expressed (qval <0.05) HERV-K encoding features with partial or full-length Env-coding features are shown in red. Darker points indicate more significant differential expression.

Differential expression analysis in spinal cord across all ALS patients and controls Reveals CHIT1, GPNMB, and APOC1 to be relevant for disease

There were 3,435 upregulated ERV loci and 9,809 downregulated ERV loci. Of these, there were 65 significantly dysregulated HERV-K features (17 upregulated and 48 downregulated) including 9 downregulated Env-coding loci and four upregulated Env-coding loci (Figure 2A). There was a total of 16 HERV-K Env-coding features (Table S2) significantly dysregulated across all SC DEAs.

In terms of ENSGs, there were 5,212 downregulated, and 5,987 upregulated. The most significantly dysregulated (all upregulated) ensemble genes in all ALS vs. unaffected controls were *CHIT1* (chitinase 1), *GPNMB* (Glycoprotein Nmb), *APOC1* (Apolipoprotein C1),

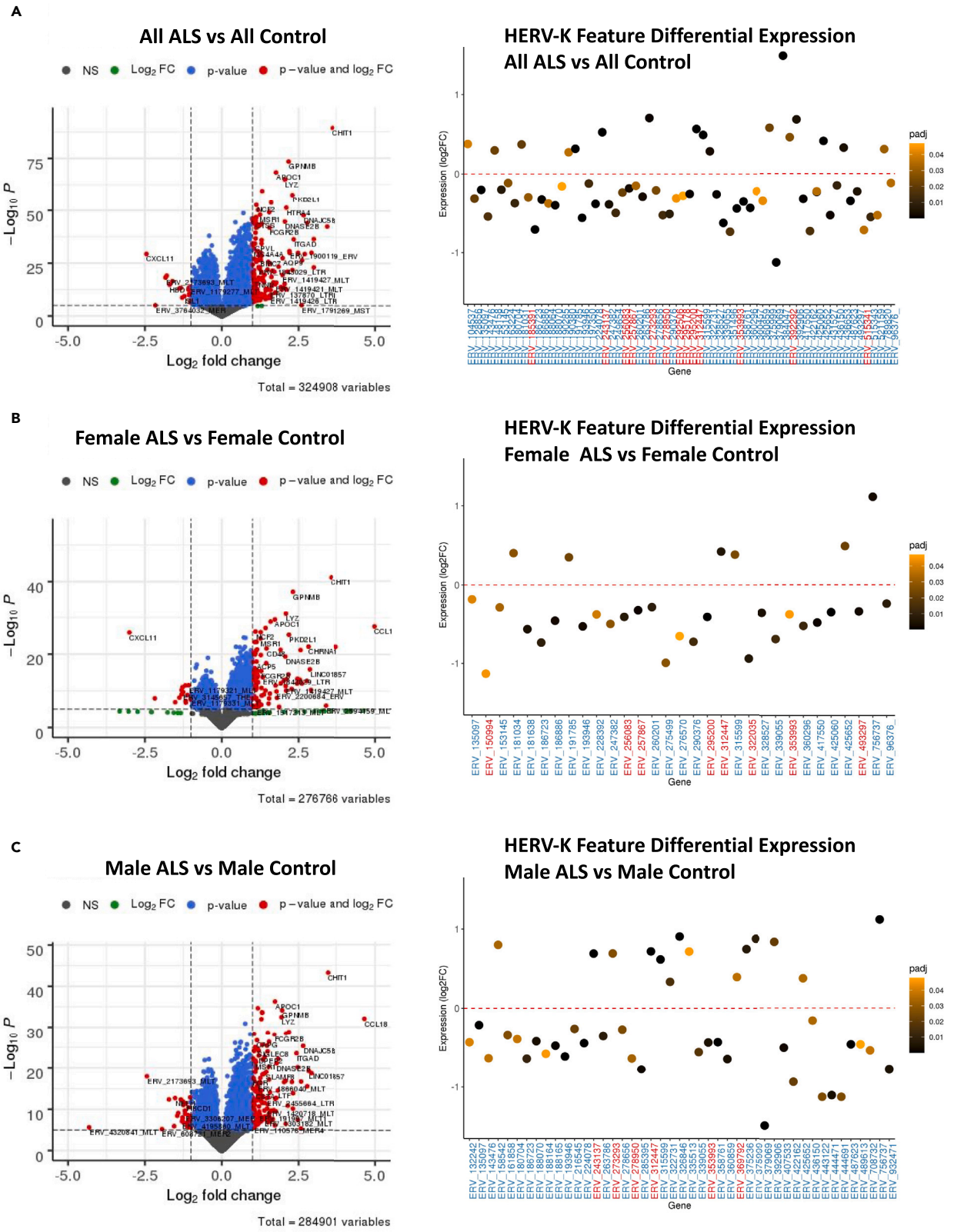

A
All ALS vs All Control
NS
Log2 FC
p-value
p – value and log2 FC
−Log10 P
Log2 fold change
Total = 324908 variables
HERV-K Feature Differential Expression
All ALS vs All Control
Expression (log2FC)
padj
Gene
B
Female ALS vs Female Control
Total = 276766 variables
HERV-K Feature Differential Expression
Female ALS vs Female Control
C
Male ALS vs Male Control
Total = 284901 variables
HERV-K Feature Differential Expression
Male ALS vs Male Control

Figure 2. ALS patient vs. unaffected control differential expression analysis (DEA) volcano and scatterplots of HERV-K features in the spinal cord (SC)
(A) Volcano plots (left) and scatterplots (right) for all samples, (B) female samples only, and (C) male samples only in the SC. Features that were significantly associated with biological sex (A only) and did not have sufficient expression were removed prior to the analysis. For volcano plots, log2FC cut-off is >|2| and cut-off for *p* value is 10E-6. Feature plot of significantly differentially expressed (qval <0.05) HERV-K encoding features with partial or full-length Env-coding features in red. Darker points indicate more significant differential expression.

LYZ (Lysozyme), and *CCL18* (Chemokine ligand 18). Chitinases, including CHIT1, and GPNMB have already been shown to be elevated in the CSF and spinal cord of ALS patients and may serve as potential biomarkers.[31,32] Chitinases regulate the metabolism of chitin, a polysaccharide, and CHIT1 is released from microglia and participates in reactive astrogliosis.[33] CHIT1 has been implicated in Alzheimer's disease, stroke, multiple sclerosis, and other neurological disorders in addition to ALS.[33] GPNMB is a transmembrane protein that regulates inflammation and has been found to be localized to microglia in neurological disorders.[34] Apolipoproteins play a role in lipid transport and metabolism in the CNS and, APOC1 is known to be expressed in the CNS.[35] A recent study shows that APOC1 is differentially expressed in both ALS patient motor neurons as well as induced pluripotent stem cell (iPSC)-derived motor neurons from ALS patients.[36] Lysozyme is known to interact with heat shock proteins to optimize their chaperone-like activity.[37] Dysregulation of this pathway can lead to increased protein aggregation in ALS. CCL18 is a chemotactic protein that recruits immune cells and regulates immune tolerance. CCL18 transcripts and protein elevations have been detected in the CNS tissues of patients with neurological diseases.[35] A slight increase in CCL18 in the CSF of ALS patients compared to controls has also been observed.[38]

Differential expression analysis in spinal cord across female patients and female controls reveals CHIT1, GPNMB, and PLA2G7 to be relevant for disease

There were 842 upregulated ERV loci and 4,877 downregulated ERV loci. Of these, there were 31 significantly dysregulated HERV-K features (6 upregulated and 25 downregulated) including 7 downregulated Env-coding loci and one upregulated Env-coding loci on 1q22 (Figure 2B). In terms of ENSGs, there were 2,847 downregulated, and 4,242 upregulated. Differential expression analysis of the female-only group similarly showed significant dysregulation of CHIT1 (chitinase 1), GPNMB (Glycoprotein Nmb), LYZ (Lysozyme) and APOC1 (Apolipoprotein C1). Additionally, there was an increase in PLA2G7 (Phospholipase A2 Group VII) (Figure 2B). CNS cellular membranes are comprised of polyunsaturated fatty acids, the metabolism of which is tightly controlled by phospholipase A_2 (PLA_2). The phospholipases that metabolize cellular membranes play important roles in apoptosis, inflammation, and oxidative stress.[39] Cytosolic PLA_2 (cPLA2) has been shown to be upregulated in motor neurons, astrocytes, and microglia in the spinal cord of ALS patients.[40] Moreover, in a SOD1 mouse model of ALS, elevated cPLA2 was detected prior to the onset of symptoms.[41]

Of the 8 Env-coding (1 upregulated and 7 downregulated) in the female sample only group (Figure 2B), the upregulated locus, ERV_312447 (1q22), was also upregulated in the all samples and male-only samples SC DEAs and is located on the negative strand of Ch1. This locus encodes for a full-length Gag protein, but only a truncated Env protein.

Differential expression analysis in spinal cord across male patients and male controls reveals CHIT1, MAFB, and DENN to be relevant for disease

There were 3,081 upregulated ERV loci and 6,840 downregulated ERV loci. Of these, there were 45 significantly dysregulated HERV-K features (14 upregulated and 31 downregulated) including 2 downregulated Env-coding loci and 4 upregulated Env-coding loci (Figure 2C). In terms of ENSGs, there were 4,534 downregulated, and 5,432 upregulated. Similarly, in the male-only group the most highly dysregulated ensemble genes were *CHIT1* (chitinase 1), *APOC1* (Apolipoprotein C1) and *GPNMB* (Glycoprotein Nmb). Additionally, *DENND2D* (DENN Domain Containing 2D), and *MAFB* (MAF BZIP (Basic Leucine Zipper Domain) Transcription Factor B) were dysregulated (Figure 2C). DENN (differentially expressed in normal versus neoplastic) is a tumor necrosis factor receptor-associated protein that plays a role in anti-apoptotic and cell survival processes and is known to be strongly expressed in CNS tissues.[42,43] Moreover, its downregulation was shown to be associated with neurodegeneration, possibly due to an oxidative stress- or cytokine-dependent mechanism.[44] MAFB is a leucine-zipper transcription factor that mediates the M2, anti-inflammatory, type of macrophages and can help restore T helper cell imbalances and alleviate inflammation at epithelium.[45]

Of the 6 Env-coding (4 upregulated and 2 downregulated) HERV-K features that were present in the male-only DEA (Figure 2C), two of these features (ERV_312447 and ERV_3697923) could encode for a full-length Gag, however, none could encode for a full-length Env. All but ERV_2431376 were LTR5_Hs and located on the reverse strand. One of the features that was significantly decreased in expression, ERV_2789502, could encode a full-length Env and is predicted to be under the transcriptional regulation of an enhancer.

Six HERV-K env-coding loci were robustly dysregulated across cortex and spinal cord

We observed 21 Env-coding loci that were significantly dysregulated in at least one of the six DEAs performed (Tables S1 and S2), male patient samples had a significantly higher odds of increased expression of a HERV-K Env-coding loci compared to females (OR = 23, Fisher's exact *p*-value = 0.007) and spinal cord samples had a trending higher odds of increased expression of a HERV-K Env-coding loci compared to cortex samples (OR = 3.8, Fisher's exact *p*-value = 0.35).

Differential transcription between biological sex groups were also examined more directly (Figure S16). There were no significant differences in terms of the GSEA, although there were trending increases in apoptotic gene sets across the CTX and SC. In terms of HERV-K

Env-coding features, there were relatively few significantly dysregulated features (three in the CTX and four in the SC). On the other hand, there were individual TE features that were highly downregulated (>16x more transcripts in males than females) in females compared to males.

Since significant differences between features in individual DEAs may lead to false positives, we also looked across all six DEAs in the CTX and SC. There were six most significantly dysregulated HERV-K Env-coding loci in terms of their significant differential expression across and between tissue types (Table 1). Most of these features were of the LTR5_Hs variety of HML-2 (almost 70%) and were downregulated in patients relative to controls (almost 70%). Only ERV_2603457 was not predicted to be under the transcriptional regulation of at least one modality. The other features were under the transcriptional regulation of at least CCCTC-binding factor (CTCF), a transcription factor, and an enhancer region. Only ERV_3922921 could encode for a full-length Env, however, it was upregulated across the all patients SC DEA and downregulated in the female-only CTX DEA. On the other hand, all but two of these loci (ERV_2603457 and ERV_3539935) could encode for a full-length gag and were located on the reverse strand. No chromosome was represented more than once. ERV_3220357, ERV_3922921, and ERV_493297 were dysregulated in both the CTX and SC: ERV_3220357 was female ALS patient-specific, ERV_3922921 was robustly decreased in expression across two DEAs in SC and CTX, and ERV_493297 had both biological sex- and CNS region-specific expression patterns.

In summary, male patients are more likely to have increased expression of an Env-coding feature, whereas female patients are more likely to have a HERV-K feature differentially expressed. The SC generally had more differentially expressed HERV-K Env-coding loci, and the loci that were robustly differentially expressed across CNS regions are more likely to be on the reverse strand, encode for full-length Gag, be under CTCF and enhancer transcriptional regulation, and be of the LTR5_Hs variety.

There is evidence of increased active inflammation in the SC compared to the CTX

To analyze the biological relevance of the most significantly differentially expressed individual genes by DEA, we utilized Qiagen Ingenuity Pathway Analysis (IPA).[46] More specifically, we performed IPA on the 1,550 ENSGs (since TE annotations are not supported) with the lowest DESeq2 *p*-value, using the DESeq2 $\log_2$ fold change for ranking. The heatmaps show the most significant pathways/regulators based on the degree of dysregulation or activation/inhibition across the analyses being compared (i.e., sum of |IPA *Z* score| across DEAs).

The most significantly dysregulated IPA canonical pathways had the opposite patterns of dysregulation in the CTX (Figure S1) and SC (Figure S2) (downregulated in CTX and upregulated in SC) (Figure 3A). Almost all the pathways relate to either inflammation/infection or wound healing. In these postmortem samples, ALS patients generally exhibited active inflammation in the SC and less inflammation in the CTX. This likely represents a differential time course in disease pathogenesis whereby the disease starts in the CTX and progresses toward the limbs. Analysis of upstream regulators showed that the same pattern was observed: all but SB203580 and ETV6-RUNX1 were predicted to be activated in SC and vice versa in CTX (Figure 3B). These two upstream regulators are involved with cancer and MAPK (mitogen-activated protein kinases) activity. ERV_3922921 (8p23.1) (Table 1) was upregulated in all patients in the SC and downregulated in female patients in the CTX. IPA pathways and regulators involved in the inflammatory response and infection were upregulated in SC and downregulated in CTX – the same pattern as ERV_3922921 (Figure 3A).

High HERV-K-expressing ALS patients and unaffected controls represent transcriptionally distinct subgroups

To determine the optimum proportion of samples to use for high HERV-K-expressing ALS patients based on frequency value (a threshold-based metric, specific to disease-status group and CNS region, for quantifying overall HERV-K expression) for both ALS patients and controls, differences in median frequency values and Mann-Whitney tests were used (Figure 4A). The frequency values for the top proportion of ALS patients or unaffected controls going from 0.05 to 0.5 (5%–50%) in increments of 0.05 (5%) was compared to the other group (unaffected controls when determining the ALS patient proportion and ALS patients when determining the unaffected control proportion). The proportion which gave the lowest *p*-value and highest difference in medians was used. These groupings were then validated using distinct methodologies including gene set enrichment analysis (GSEA) and DEA (Figures S3–S9; Tables S3 and S4). The distribution of frequency values was shown to be similar in the SC and CTX across all ALS patients and all controls (Figures S10 and S11).

In the CTX, the top 20% of ALS patients and top 30% of unaffected controls represent the optimal subgroups based on this approach. The optimum percentage of patients in the SC was also 20% and controls was 30% (Figure 4B). This result indicates that the optimum number of patients to select for the high HERV-K-expressing subgroup is CNS region-independent. Interestingly, the significance level and difference in medians was generally lower for SC than for CTX (Figures S3 and S9), despite there being slightly more super-threshold HERV-K features by frequency analysis in ALS patients relative to controls in the SC compared to CTX (Figures S10 and S11). This result indicates that the high HERV-K-expressing ALS patient group in the CTX likely has greater upregulation of HERV-K features relative to the corresponding unaffected control group than the SC high HERV-K-expressing ALS patient subpopulation. There were no significant differences in terms of clinical and phenotypic characteristics between high HERV-K ALS patients and other ALS patients in the CTX, SC, and both CTX and SC (Tables S6–S8). On the other hand, there was a trending association with high HERV-K ALS status and older age of onset, age at death, and dementia.

The high HERV-K ALS group was compared with the high HERV-K control group (Figure S15). There were minimal differences in the CTX and no significant difference in terms of HERV-K Env-coding features. In the SC, there were more individual differences in terms of transcription, including six Env-coding features as well as ENSGs. Like in the all ALS vs. all control analysis, CHIT1, LYZ, and APOC1 were the most dysregulated features.

Table 1. Significant HERV-K Env-coding features across spinal cord (SC) and cortex (CTX) differential expression analyses (DEAs)

HERVd ID	HERV-K protein coding information from HERVd	Chromosome: start-end	DNA Strand (from Hervd and RepeatMasker)	Chromosome Band Number	RepeatMasker repetitive features (from Ensembl)	Additional Notes (from Ensembl)	Regulatory Information (from Ensembl)	Relevant DEA
ERV_312447[a]	Full Gag; Partial env, pol, and Rec	1:155626666-155635845	reverse	1q22 (ERVK-7)	LTR5-Hs, HERVK-int, SVA_A	lncRNA and unprocessed pseudogene on reverse strand as well as Enhancer region with non-coding exon variant SNPs and promoter flank region with intron variant SNPs on the LTR5-Hs region	CTCF, enhancer, promoter, promoter flank	SC: All Three Analyses
ERV_2603457[b]	Partial env, pol, and Rec	3:130211772-130212782	forward	3q22.1	SVA_A, LTR5A, HERVK-int	N/A	N/A	CTX: All Patients and Females Only
ERV_3220357[b]	Full Gag; Partial env, pol, and Rec	5:30486653-30496098	reverse	5p13.3	LTR5-Hs, HERVK-int, SVA_A	LTR5-Hs in enhancer region with regulator variant SNPs	CTCF, enhancer	SC and CTX: Females Only
ERV_3539935[b]	Partial env, and Rec	6:150859613-150862438	forward	6q25.1	HERVK-int, SVA_A, LTR5B	Overlapping enhancer and CTCF region with regulatory region variant SNPs	CTCF, enhancer	SC: All Three Analyses
ERV_3922921[c]	Full env, Rec and Gag; Partial Pol	8:7497875-7507337	reverse	8p23.1 (ERVK-8)	LTR5-Hs, HERVK-int, SVA_A, MamRTE1	LTR5-Hs in enhancer region with intronic variant SNPs and DEFB107B protein coding transcripts in forward strand	CTCF, enhancer	SC: All Patients Only CTX: Females Only
ERV_493297[b]	Full Gag, Partial env, pol, and Rec	10:6824179-6833641	reverse	10p14 (ERVK-16)	LTR5-Hs, HERVK-int, SVA_A	lncRNAs on both forward and reverse strands as well as multiple enhancer regions on LTR5-Hs, some with non-coding exon variant SNPs	CTCF, enhancer	SC and CTX: Females and All Patients Only

[a]Indicates features that are upregulated across CTX and SC.
[b]Indicates features that are downregulated across CTX and SC, and.
[c]Indicates features that are upregulated in the SC and downregulated in the CTX.

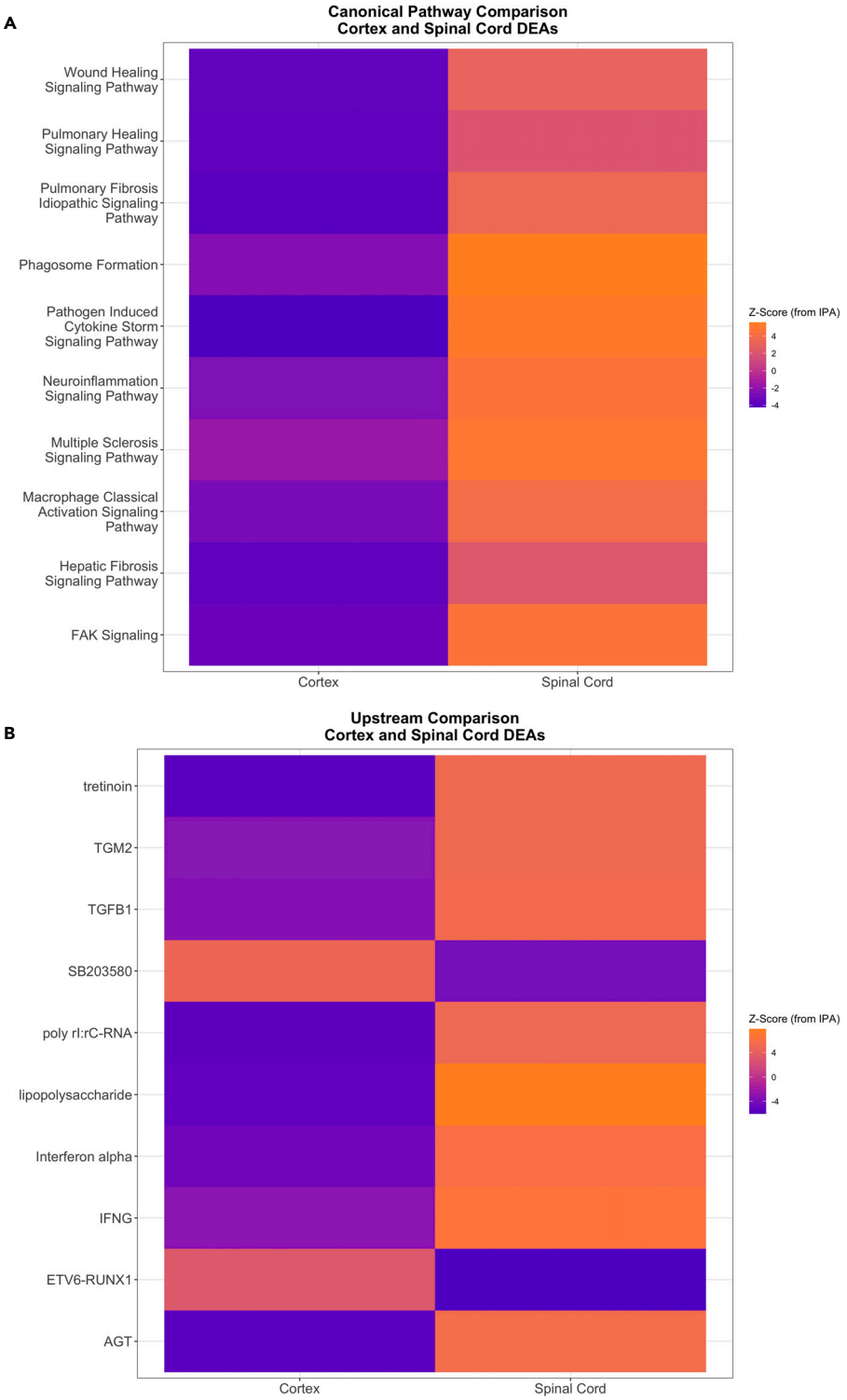


Figure 3. Ingenuity pathway analysis (IPA) comparison cortex (CTX) and spinal cord (SC) all patients vs. controls

(A) IPA results for canonical pathway analysis and (B) upstream regulator analysis using DEA results comparing all ALS patients vs. controls in the CTX (right) and SC (left). *Z* score values were exported from IPA and are represented from low (blue) to high (orange). Gray values indicate that there was no clear direction of change. A *Z* score over 2 represents significant upregulation and lower than −2 represents significant downregulation. In general, the CTX and SC were opposites in terms of patterns of regulation among canonical pathways and upstream regulators in ALS patients relative to controls.

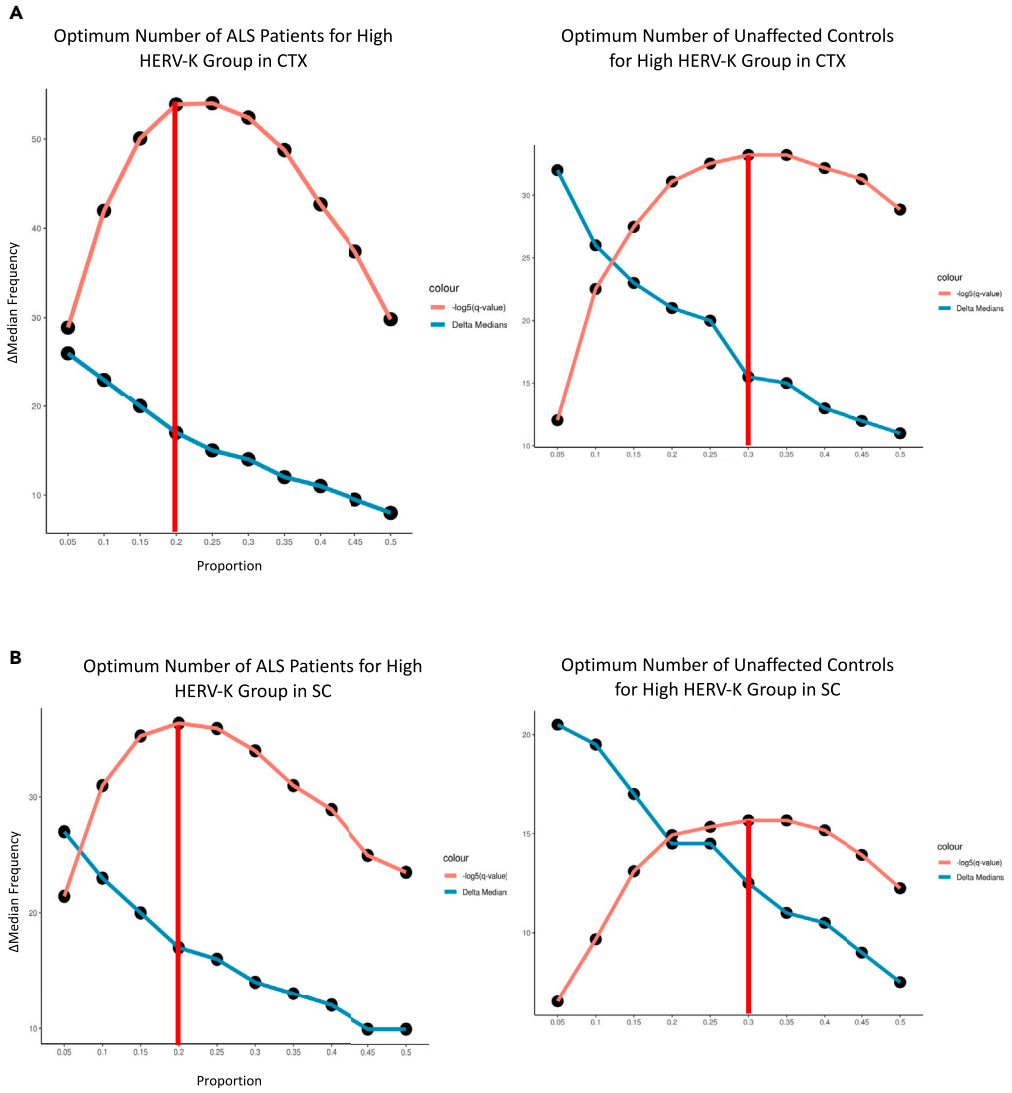


Figure 4. Optimum proportion of samples to include in high HERV-K group cortex (CTX) and spinal cord (SC)

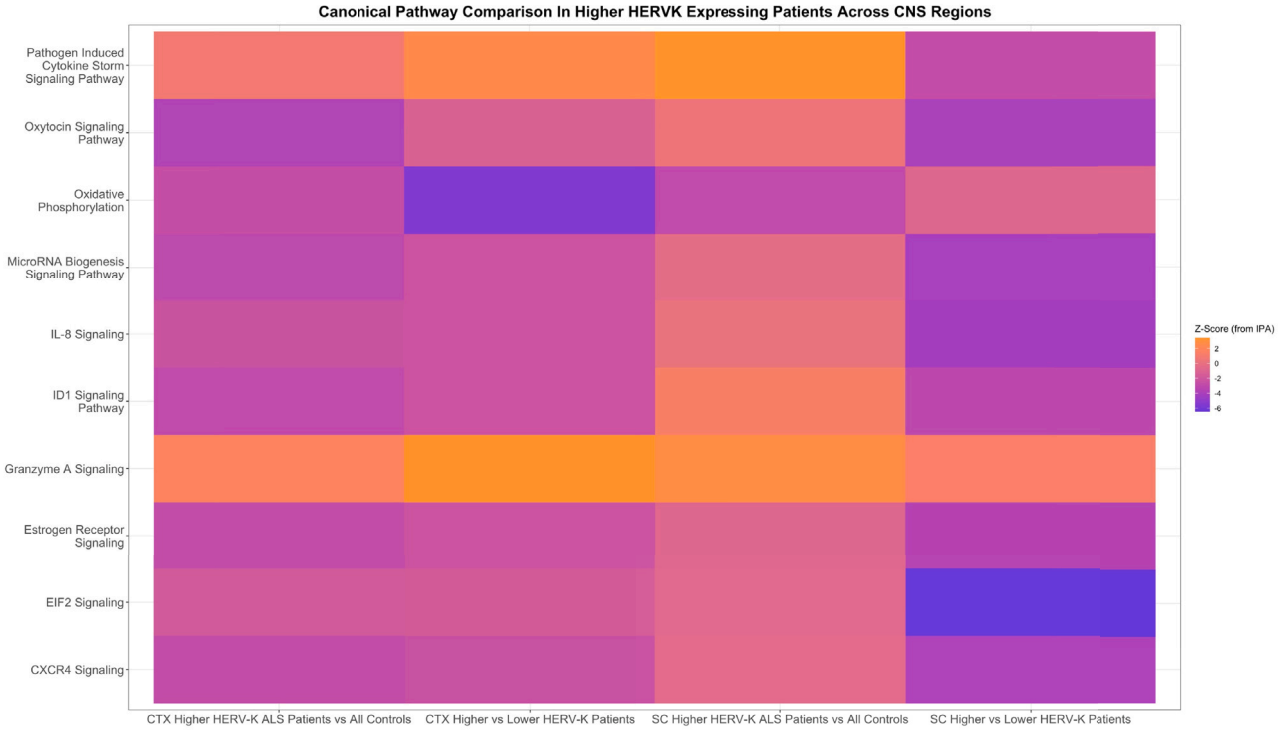
(A) Optimization was based on maximizing significance of q-value from Mann-Whitney test and effect size (difference in medians) in ALS patients (left) and unaffected controls (right) in CTX and (B) SC. All comparisons are based on the frequency metric of overall HERV-K expression. Orange line represents the Mann-Whitney significance level, while the blue line represents the difference in effect size of the corresponding ALS patient subgroup compared to all controls (or control subgroup compared to all ALS patients). The vertical red line indicates optimal proportion of samples to theoretically maximize differences in HERV-K expression. The optimum proportions were 0.2 in the ALS group and 0.3 in the control group.

There was a significant positive correlation (Spearman r > 0.2 and FDR adjusted *p*-value <0.05) between *APOC1*, *CHIT1*, and *GPNMB* and both ERV_3922921, 8p23.1, and ERV_1882163, 19q11, as well as *IGKC* and ERV_3922921 (Figure S17; Table S11). These significant correlations were only present across all tissue types in the NYGC dataset, there were no significant positive correlations in the CTX and SC specifically.

Granzyme A and demethylation-related features are robustly dysregulated in high HERV-K-expressing ALS patients

To determine which pathways and upstream regulators were most significantly, and specifically, dysregulated in high HERV-K-expressing ALS patients, a separate IPA comparison analysis was performed combining CTX and SC analyses and high HERV-K-expressing ALS patients vs. controls and the other ALS patients (low HERV-K-expressing ALS patient) DEAs (Figure 5). Granzyme A signaling represents the most robustly upregulated pathway across CNS regions and high HERV-K-expressing patient DEAs. Other robustly dysregulated pathways include other inflammatory-related pathways as well as miRNA biogenesis and oxidative phosphorylation, although they were generally less consistently and specifically upregulated (Figure 5A).

A


Canonical Pathway Comparison In Higher HERVK Expressing Patients Across CNS Regions
Pathogen Induced Cytokine Storm Signaling Pathway
Oxytocin Signaling Pathway
Oxidative Phosphorylation
MicroRNA Biogenesis Signaling Pathway
IL-8 Signaling
ID1 Signaling Pathway
Granzyme A Signaling
Estrogen Receptor Signaling
EIF2 Signaling
CXCR4 Signaling
CTX Higher HERV-K ALS Patients vs All Controls
CTX Higher vs Lower HERV-K Patients
SC Higher HERV-K ALS Patients vs All Controls
SC Higher vs Lower HERV-K Patients
Z-Score (from IPA)
2
0
-2
-4
-6

B


Canonical Pathway Comparison In Higher HERVK Expressing Patients Across CNS Regions
ST1926
RICTOR
NFE2L2
MYC
lipopolysaccharide
KDM5A
GSKJ4
GABA
CD 437
5-fluorouracil
CTX Higher HERV-K ALS Patients vs All Controls
CTX Higher vs Lower HERV-K Patients
SC Higher HERV-K ALS Patients vs All Controls
SC Higher vs Lower HERV-K Patients
Z-Score (from IPA)
4
0
-4
-8

Figure 5. Ingenuity Pathway Analysis (IPA) across cortex (CTX) and spinal cord (SC) High HERV-K-expressing ALS vs. rest of ALS patients and unaffected controls

(A) IPA results for canonical pathway analysis and (B) upstream regulator analysis using DEA results comparing high HERV-K-expressing ALS patients vs. all controls in the CTX (left column), high HERV-K-expressing ALS patients vs. low HERV-K-expressing ALS patients in the CTX (middle left column), high HERV-K-expressing ALS patients vs. all controls in the SC (middle right column), and high HERV-K-expressing ALS patients vs. low HERV-K-expressing ALS patients in the SC (right column). *Z* score values were exported from IPA and are represented from low (blue) to high (orange). Gray values indicate that there was no clear direction of change. A *Z* score over 2 represents significant upregulation and lower than −2 represents significant downregulation. In general, there were similar expression patterns for canonical pathways across comparisons.

Regarding upstream regulators (Figure 5B), none were specifically predicted to be inhibited in the high HERV-K-expressing group, however, the GSKJ4-, a histone demethylase inhibitor that can cross the blood brain barrier, mediated regulatory cascade was predicted to be activated. GSKJ4 has been implicated as a possible therapeutic target for Parkinson's disease and glioma.[47,48] GSKJ4 increases trimethylation at histone 3 lysine 27 (H3K27me3) via selective H3K27 demethylase inhibition.[49] H3K27me3 is associated with neuronal function and its enrichment has been shown in ALS and FTD patients.[50] MYC and NFE2L2 were downregulated across at least one of the two analyses in the CTX and SC. MYC is implicated in a variety of cancers, although, it is challenging to target directly.[51] Nuclear factor, erythroid 2 like 2 (NFE2L2, also known as NRF2) is a transcription factor that regulates oxidative stress and inflammatory response and has been implicated as a marker of immune infiltration in glioma.[52]

CA1, CXCL11, and ERVL-E are the most important differentially expressed genes by logistic regression

Logistic regression (LR) was used to further validate individual features identified by DEA. LR is a machine learning (ML) technique that uses an input set of features (DESeq2 normalized count values for 500 features in this case) to predict a binary outcome measure (ALS patient or control status in this case). Due to the large number of features used as predictors, we used a ridge regularization penalty (RLR) to avoid the potential of overfitting our model to the training data and improve generalizability. We split our data into a training dataset (70% of data) to train the model and a testing dataset (30% of data) to determine the performance of our model. We specifically examined whether the 500 most significantly differential expressed features in ALS patients compared to unaffected controls by DEA could predict patient status using LR.

LR models that were about 84% accurate at classifying ALS patients from controls utilizing differentially expressed genes (DEGs) from the DEA analyses were generated in both the CTX (Figure S12) and SC (Figure S13). There were 9 features that were in the final fitted model trained on the 500 most significant features by DEA analyses using the balanced models (Table 2). In other words, these features are among the 500 most differentially expressed features and are important to an LR model that is highly accurate at distinguishing patients from controls in both the CTX and SC. These features consisted of 4 TEs (2 LTRs, a MER61, and an ERVL-E) and 5 ENSGs. All but three (JAML, LILRA5, and S100A4) of these features had the same pattern of expression in both SC and CTX. This indicates that unlike the overall expression pattern of the most dysregulated individual features and IPA pathways, there is relatively high concordance in terms of expression patterns for these features between the CTX and SC. The ERVL-E feature was the only one that was increased in expression in both the CTX and SC. This ERVL-E feature was also the most important of these nine features for both the CTX and SC LR models. However, CA1, CXCL11, LTR17, LTR12C, and MER61 were decreased in expression in both regions.

DISCUSSION

ALS is a neurodegenerative disorder that involves the UMN in the cortex and the LMNs in the spinal cord. However, the pattern of degeneration and disease progression is variable. It affects males more commonly than females. Some genetic mutations conferring an increased risk of developing ALS have been identified although most cases are sporadic in nature. A previous study grouped patients based on transcriptional profiles in brain at autopsy.[53] They identified a group characterized by transposable element (TE) dysregulation; however, we specifically identified a high HERV-K-expressing subgroup in this study. Here we performed an extensive analysis to determine if there were specific ENSGs and HERVs dysregulated in the CNS of ALS patients compared to unaffected controls. Our initial analysis showed that several HERV-K transcripts were downregulated in ALS patients compared to controls. Female patients compared to female controls had more loci dysregulated than male patients compared to male controls. A recent study showed there was baseline expression of some HERV-K loci in normal CNS tissue and that the specific loci which are expressed varies depending on tissue type (Burn et al., Plos Biology 2022). The physiological role of these endogenous viral genes is unknown. However, several loci have been shown to be upregulated in ALS patients which may be of pathological significance.[13,16] In particular, the HERV-K subtype HML-2 Env protein has been shown to be neurotoxic.[54] Hence we explored if any of the upregulated loci could encode for the Env protein. In high HERV-K ALS patients in both the CTX and SC the specific HERV-K loci that were upregulated were 1q21.3 (ERV_127104), 5q15 (ERV_3255463), 8q24.3 (ERV_4007080), and 11q13.4 (ERV_617480) (Table S9). Moreover, all 8 HERV-K loci in the human genome known to encode a full-length Env were upregulated in at least one DEA (Table S10). The most highly expressed loci in order of increasing expression were ERV_3922921 (8p23.1), ERV_1882163 (19q11), and ERV_2599259 (3q21.2). 8p23.1 has both LTR's and sits in the opposite orientation to a Defensin beta protein coding transcript (DEFB107B).[55] 19q11 lacks a 5′ LTR; however, the 5′ end of the locus is in a region predicted to be under the transcriptional regulation of a promoter. Additionally, the nearest ERV LTR is over 10 kb's away.[55] 3q21.2 has both a 5′ and a 3′ LTR, but it has stop codons in-frame. However, it could make an entire signal peptide and about 40% of the surface subunit even with the stop codons.[56] Moreover, there is one LTR within 3 kb's upstream of the 5′ LTR and 2 LTR's within 3 kb's downstream of the 3′ LTR, so alternate reading frames and splicing events should be possible. The exact mechanism by which these loci are expressed at such high levels relative to other loci needs further research to clarify.

Table 2. Significantly dysregulated features implicated in final fitted logistic regression (LR) model in the cortex (CTX) and spinal cord (SC)

ENSG/ERV	Feature Importance CTX	Feature Importance SC	Feature Name	L2FC CTX	P-adj CTX	L2FC SC	P-adj SC
ENSG00000133742[b]	1.959E-02	1.103E-03	CA1	−0.872	2.960E-03	−1.717	4.753E-15
ENSG00000160593[c]	1.317E-04	9.296E-05	JAML	−0.745	6.908E-06	1.025	9.477E-21
ENSG00000169248[b]	6.628E-03	2.170E-04	CXCL11	−1.345	3.049E-07	−2.543	3.707E-27
ENSG00000187116[c]	4.891E-03	1.350E-04	LILRA5	−0.921	4.406E-04	1.270	5.937E-14
ENSG00000196154[c]	2.597E-05	1.452E-05	S100A4	−0.918	2.223E-07	0.957	2.080E-21
ERV_1378003[b]	5.358E-03	2.468E-04	LTR17	−0.676	1.547E-02	−0.750	6.855E-15
ERV_2200684[a]	1.163E-01	4.817E-03	ERVL-E	0.966	9.688E-05	1.211	2.958E-09
ERV_2964235[b]	5.052E-05	2.019E-04	LTR12C	−0.912	2.878E-03	−0.841	2.293E-11
ERV_909983[b]	8.662E-04	7.517E-04	MER61	−0.942	9.740E-03	−2.038	5.506E-04

DESeq2 $\log_2$fold change (L2FC) and FDR-adjusted *p*-values as well as LR feature importance values are included for both CTX and SC. For ENSGs, the feature name is the HGNC symbol, and for TEs, the feature name is the LTR or ERV associated with that locus.

[a]indicates features that are upregulated across CTX and SC.

[b]indicates features that are downregulated across CTX and SC, and.

[c]indicates features that are upregulated in the SC and downregulated in the CTX.

Interestingly, the spinal cord had greater HERV-K dysregulation with more loci being upregulated. Of these, ERV3922921 on 8p23.1, also termed HERV-K115, which encodes full length *env*, *gag* and *rec* sequences[57] was the only locus significantly upregulated across all ALS patients compared to controls in the CTX and SC that was not associated with biological sex. HERV-K115 has a 1 bp deletion 92 bases upstream from the stop codon of *gag*. This mutation alters the carboxyl terminus of the encoded Gag precursor protein and results in a ribosomal frameshift such that *pro* and *pol* cannot be translated. However, it encodes for a full length Env protein.[58] The 8p23.1 inversion (*8p23-inv*) is one of the largest polymorphic inversions found in humans, encompassing ~4.5 Mb.[59,60] In the small samples studied to date, the inversion has estimated frequencies of 59% in the Yoruba, 20%–50% in European, and 12%–27% in Asian ancestry populations.[61] This region encompasses loci associated with autoimmune- and cardiovascular-related diseases. Therefore, the role of HERV-K115 in the pathogenesis of ALS should be explored in future studies.

A possible pathophysiological mechanism for the differences in transcriptional activity between the two CNS regions could be the human silencing hub (HUSH) complex,[62] which is able to transcriptionally repress repetitive genomic elements via methylation of a specific histone. In this study, we suggest dysregulation at the epigenetic level based on IPA canonical pathway and upstream regulator analysis, including H3K27, which may be involved in these transcriptional differences.

Despite the heterogeneity of expression in males and females, or CTX versus SC, ERV312447 on Chr 1q22 also termed HERV-K102 or ERVK-7, was upregulated in ALS patients in the SC specifically. This is a relatively intact provirus with a full-length Gag and Pol. Regarding Env, it can encode a full ERVK-7 Env which contains surface and transmembrane domains only and is in an intergenic region of locus 1q22. It constitutes most of the HML-2 derived transcripts in response to pro-inflammatory activation of macrophages, especially in the context of gamma interferon. HERV-K102 mediates the switch from IFN-γ signaling to the activation of type I interferon expression, thus potentially enhancing pro-inflammatory signaling.[63] Future studies should explore its role in the innate immune activation in ALS.

In male ALS patients in the CTX, ERV_1882163 (19q11), which can encode a full-length Env, Gag, Rec, Protease, and Reverse transcriptase (RT) was upregulated. Further, there were more HERV-K loci expressed in the spinal cord compared to the cortex, suggesting the possibility that HERV-K might be a more significant contributor to spinal cord pathology in these patients. Another interesting observation was that many of the HERV-K loci expressed in ALS patients are in regions predicted to be regulated by a zinc finger transcription factor, CTCF. Even though many of the dysregulated loci in this study cannot encode a full-length Env, they may still be relevant to ALS pathophysiology via mediating neuroinflammation.

In addition to *env*, *gag* is upregulated in several loci. Gag protein plays a critical role in viral propagation and hence may participate in ALS pathophysiology and viral spread. The *pol* gene encodes for reverse transcriptase, integrase and protease. These enzymes have been the target of clinical trials in ALS patients.[64] Like Env, there is evidence of both up and downregulation, depending on the locus, of Pol-encoding HERV-K loci. While not a focus of this study, future studies should determine the relevance of RT, integrase, and protease encoding loci.

Additionally, we directly predicted high HERV-K ALS status using ENSGs and ERVs that explain the most variation between samples. HERVs are known to cause activation of innate immune responses through a variety of mechanisms. Consistent with this observation, we found several inflammatory pathways were activated. This included the chemokine-, cytokine-, and immunoglobulin-related pathways. The activation of inflammatory pathways was more prominent in the SC, like HERV-K expression. We also found activation of genes involved in oxidative stress. Oxidative stress pathways have been well studied in the pathophysiology of ALS[65–67] and may constitute another subset of patients with ALS.[53,68] Patients with high HERV-K expression also showed evidence of transcriptional dysregulation in biological pathways and transcriptional regulators involved with DNA methylation, MYC, NFE2L2 and Granzyme A.

The subgroup of patients with high HERV-K expression was characterized to determine if they had any unique clinical phenotype. About 20% of ALS patients and 30% of unaffected controls formed this subgroup. The higher percentage in controls may reflect the more heterogeneous nature of this group as it comprises healthy controls as well as those with neurological conditions other than ALS and other non-neurological diseases. There were no significant differences in terms of clinical attributes that could be used to identify the high HERV-K-expressing ALS subgroup of patients. However, ML models based on transcript abundance metrics were highly accurate at identifying high HERV-K-expressing ALS patients from other samples.

This study facilitated the nomination of 12 ENSGs that could be considered, and indeed some already are, possible druggable or diagnostic targets for ALS (Table S5). Only those features that were dysregulated across at least four analyses were considered. By performing separate sex-specific and combined DEAs in the CTX and SC, the robustness of the results was improved beyond traditional DEA in which only one analysis is performed, often without subsetting samples based on biological sex. Moreover, many of these features were important to the LR model (Table 2), further supporting their potential utility. Based on these results, the ENSGs that should be prioritized for future study are *CA1* and *CXCL11*. This is because these were the only two genes with the same pattern of differential expression across CNS regions, as well as within regions, and were significant across all samples in the CTX and SC. While both genes have relatively low average transcript abundances across all samples (Table S5, base means column), they are highly significantly increased in expression in unaffected controls relative to patients in all samples.

In terms of ENSGs that were significant in the DEA and important to the LR model, carbonic anhydrase I (CA1) is known to be expressed in human spinal motor neurons and has been suggested to be involved in ALS.[69] Additionally, CAs catalyze the hydration of CO_2; are important in pH homeostasis; regulate CSF and blood flow to the CNS; and have been implicated as potential therapeutic targets for stroke and Alzheimer's disease.[70] Junctional adhesion molecule-like protein (JAML) is a recently discovered junctional adhesion molecule (JAM) in the immunoglobulin superfamily primarily implicated in cancer. JAMs are also known to have essential roles in maintaining BBB integrity.[71,72] CXCL11 (C-X-C motif chemokine ligand 11) has been shown to be expressed in ALS patients.[73] S100 Calcium Binding Protein A4, S100A4, is a Ca^{2+}-binding protein with inflammation- and fibrosis-related mechanisms of action and has been implicated in the impaired autophagy and fibrosis pathologies seen in ALS. It has also been suggested to be a potential marker of reactive microglia in *SOD1*-mediated ALS.[74,75] Leukocyte immunoglobulin-like receptors are a family of proteins that regulate inflammatory states, plasticity in the CNS, and differential expression and polymorphisms of these genes have been associated with both autoimmune and infectious diseases.[76] However, the relevance of LILRA5 (Leukocyte Immunoglobulin Like Receptor A5) to ALS seems to have not been explored previously.

Of the transposable elements, LTR17 is a solo-LTR associated with HERV17 from the HERVW9 group of the ERV1 superfamily (Gammaretroviruses and Epsilonretroviruses). Meanwhile, LTR12C is a solo-LTR associated with HERV9 from the same group and superfamily. ERVL-E is in the HERVL group and ERV3 superfamily. MER61 is associated with the HEPSI (human epsilon) group of the ERV1 superfamily.[77,78] A previous study had identified a HML6 locus on chromosome 3 as being upregulated in the motor cortex of ALS patients.[79] The corresponding locus in the GTF file used for this analysis is ERV_2532252 and was not significantly dysregulated across any of the ALS patient vs. control DEAs in the CTX and SC or in the high HERV-K patient DEAs. DEA *p*-values were positively correlated with LR feature importance score in both CTX and SC (Figure S14).

The Human Protein Atlas[80] was referenced to determine the general expression patterns of CA1 and CXCL11. CA1 is mostly expressed (at both the RNA and protein levels) in the gastrointestinal, bone marrow, and lymphoid tissues. CXCL11 has higher RNA expression in brain than CA1, particularly in the cerebral cortex and brainstem, however it is more widely expressed in peripheral tissues. Like CA1, it is highly expressed at the RNA level in the gastrointestinal tract and bone marrow. However, it is also highly expressed in the pancreas and respiratory system tissues, which CA1 is not. This suggests that CA1 and CXCL11 could serve as peripheral biomarkers for ALS that can be detected in specific tissue types.

This study, the largest study of its kind to date, utilized a combination of standard DEA as well as novel ML methods to arrive at the most robustly differentially regulated ENSG and TE features in ALS patients while accounting for CNS region- and biological sex-specific effects. Across all ALS patients, *CA1*, *CXCL11* and the TE *ERVLE* are highly differentially expressed across CNS-regions and are important features for predicting ALS status. Using a thresholding-based method, we identified a subset of ALS patients with upregulated HERV-K expression. These patients also have a unique transcriptional signature of ENSGs. We found the expression of HERV-K proviruses implicated in our study correlated with ENSGs that were significantly dysregulated in the SC. These findings could lead to the development of new diagnostic or therapeutic targets in ALS and highlights the significance of the high HERV-K-expressing subgroup of ALS patients.

Limitations of the study

The main limitations of this study are that all samples are postmortem and were taken from patients with severe disease, so everything in this study relates to severe, late-stage ALS. Future studies should aim to characterize changes over the course of disease, particularly at pre-symptomatic stages, by utilizing model systems such as patient-derived cerebral organoid models,[81] or by studying potential serum-based markers throughout the disease course. Additionally, it is not known whether these findings are a consequence or cause of disease. Further research should be directed toward determining the mechanistic relationship between the features nominated in this study and ALS. Moreover, the quantification of repetitive elements using short-read RNA-seq technologies represents a challenge due to ambiguously mapped reads resulting in relatively low transcript abundances for many TEs.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Patients and samples
  - Sample preparation
- METHOD DETAILS
  - Differential expression analysis
  - Frequency metric
  - Pathway and gene set analysis
  - Genome browser
  - Protein-level analysis
  - Logistic regression
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110147.

ACKNOWLEDGMENTS

We would like to thank the patients and families of those who participated in the participating sites that donated tissue samples for this project as part of the NYGC's ALS Consortium. We would also like to thank Dr's Molly Hammell, Oliver Tam, and Bianca Dumitrascu for their assistance with designing analyses. Finally, we would like to thank the NIH high performance computing (Biowulf) team for supporting our analyses. The study was supported by intramural funds (NS 03130) from the National Institute of Neurological Disorders and Stroke at the National Institutes of Health.

AUTHOR CONTRIBUTIONS

N.P., T.D.-O., O.P., and A.N. conceived and designed the study. N.P. designed and implemented the statistical methods and performed the computational analyses. T.D.-O. and K.J. helped prepare the raw sequencing data for analysis. N.P. interpreted analytical results. N.P. wrote the manuscript and all authors edited the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 15, 2024
Revised: April 25, 2024
Accepted: May 27, 2024
Published: May 28, 2024

REFERENCES

1. Ryan, M., Heverin, M., McLaughlin, R.L., and Hardiman, O. (2019). Lifetime Risk and Heritability of Amyotrophic Lateral Sclerosis. JAMA Neurol. *76*, 1367–1374. https://doi.org/10.1001/jamaneurol.2019.2044.
2. Gibson, S.B., Downie, J.M., Tsetsou, S., Feusier, J.E., Figueroa, K.P., Bromberg, M.B., Jorde, L.B., and Pulst, S.M. (2017). The evolving genetic risk for sporadic ALS. Neurology *89*, 226–233. https://doi.org/10.1212/WNL.0000000000004109.
3. Keller, M.F., Ferrucci, L., Singleton, A.B., Tienari, P.J., Laaksovirta, H., Restagno, G., Chiò, A., Traynor, B.J., and Nalls, M.A. (2014). Genome-Wide Analysis of the Heritability of Amyotrophic Lateral Sclerosis. JAMA Neurol. *71*, 1123–1134. https://doi.org/10.1001/jamaneurol.2014.1184.
4. Mejzini, R., Flynn, L.L., Pitout, I.L., Fletcher, S., Wilton, S.D., and Akkari, P.A. (2019). ALS Genetics, Mechanisms, and Therapeutics: Where Are We Now? Front. Neurosci. *13*, 1310. https://doi.org/10.3389/fnins.2019.01310.
5. Cacabelos, D., Ramírez-Núñez, O., Granado-Serrano, A.B., Torres, P., Ayala, V., Moiseeva, V., Povedano, M., Ferrer, I., Pamplona, R., Portero-Otin, M., and Boada, J. (2016). Early and gender-specific differences in spinal cord mitochondrial function and oxidative stress markers in a mouse model of ALS. Acta Neuropathol. Commun. *4*, 3. https://doi.org/10.1186/s40478-015-0271-6.
6. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921. https://doi.org/10.1038/35087627.
7. Bowles, H., Kabiljo, R., Al Khleifat, A., Jones, A., Quinn, J.P., Dobson, R.J.B., Swanson, C.M., Al-Chalabi, A., and Iacoangeli, A.

(2022). An assessment of bioinformatics tools for the detection of human endogenous retroviral insertions in short-read genome sequencing data. Front. Bioinform. *2*, 1062328. https://doi.org/10.3389/fbinf.2022.1062328.

8. He, J., Babarinde, I.A., Sun, L., Xu, S., Chen, R., Shi, J., Wei, Y., Li, Y., Ma, G., Zhuang, Q., et al. (2021). Identifying transposable element expression dynamics and heterogeneity during development at the single-cell level with a processing pipeline scTE. Nat. Commun. *12*, 1456. https://doi.org/10.1038/s41467-021-21808-x.
9. Jin, Y., Tam, O.H., Paniagua, E., and Hammell, M. (2015). TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics *31*, 3593–3599. https://doi.org/10.1093/bioinformatics/btv422.
10. Li, W., Lin, L., Malhotra, R., Yang, L., Acharya, R., and Poss, M. (2019). A computational framework to assess genome-wide distribution of polymorphic human endogenous retrovirus-K In human populations. PLoS Comput. Biol. *15*, e1006564. https://doi.org/10.1371/journal.pcbi.1006564.
11. Subramanian, R.P., Wildschutte, J.H., Russo, C., and Coffin, J.M. (2011). Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. Retrovirology *8*, 90. https://doi.org/10.1186/1742-4690-8-90.
12. Wildschutte, J.H., Williams, Z.H., Montesion, M., Subramanian, R.P., Kidd, J.M., and Coffin, J.M. (2016). Discovery of unfixed endogenous retrovirus insertions in diverse human populations. Proc. Natl. Acad. Sci. USA *113*, E2326–E2334. https://doi.org/10.1073/pnas.1602336113.
13. Douville, R., Liu, J., Rothstein, J., and Nath, A. (2011). Identification of active loci of a human endogenous retrovirus in neurons of patients with amyotrophic lateral sclerosis. Ann. Neurol. *69*, 141–151. https://doi.org/10.1002/ana.22149.
14. Dhillon, P., Mulholland, K.A., Hu, H., Park, J., Sheng, X., Abedini, A., Liu, H., Vassalotti, A., Wu, J., and Susztak, K. (2023). Increased levels of endogenous retroviruses trigger fibroinflammation and play a role in kidney disease development. Nat. Commun. *14*, 559. https://doi.org/10.1038/s41467-023-36212-w.
15. Bhetariya, P.J., Kriesel, J.D., and Fischer, K.F. (2017). Analysis of human endogenous retrovirus expression in multiple sclerosis plaques. J. Emerg. Dis. Virol. *3*. https://doi.org/10.16966/2473-1846.133.
16. Li, W., Lee, M.H., Henderson, L., Tyagi, R., Bachani, M., Steiner, J., Campanac, E., Hoffman, D.A., von Geldern, G., Johnson, K., et al. (2015). Human endogenous retrovirus-K contributes to motor neuron disease. Sci. Transl. Med. *7*, 307ra153. https://doi.org/10.1126/scitranslmed.aac8201.
17. Chang, Y.-H., and Dubnau, J. (2023). Endogenous retroviruses and TDP-43 proteinopathy form a sustaining feedback driving intercellular spread of Drosophila neurodegeneration. Nat. Commun. *14*, 966. https://doi.org/10.1038/s41467-023-36649-z.
18. Mayer, J., Harz, C., Sanchez, L., Pereira, G.C., Maldener, E., Heras, S.R., Ostrow, L.W., Ravits, J., Batra, R., Meese, E., et al. (2018). Transcriptional profiling of HERV-K(HML-2) in amyotrophic lateral sclerosis and potential implications for expression of HML-2 proteins. Mol. Neurodegener. *13*, 39. https://doi.org/10.1186/s13024-018-0275-3.
19. Garson, J.A., Usher, L., Al-Chalabi, A., Huggett, J., Day, E.F., and McCormick, A.L. (2019). Quantitative analysis of human endogenous retrovirus-K transcripts in postmortem premotor cortex fails to confirm elevated expression of HERV-K RNA in amyotrophic lateral sclerosis. Acta Neuropathol. Commun. *7*, 45. https://doi.org/10.1186/s40478-019-0698-2.
20. Ishihara, T., Koyama, A., Hatano, Y., Takeuchi, R., Koike, Y., Kato, T., Tada, M., Kakita, A., and Onodera, O. (2022). Endogenous human retrovirus-K is not increased in the affected tissues of Japanese ALS patients. Neurosci. Res. *178*, 78–82. https://doi.org/10.1016/j.neures.2022.01.009.
21. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.
22. Harvey, T.J., Davila, R.A., Vidovic, D., Sharmin, S., Piper, M., and Simmons, D.G. (2021). Genome-wide transcriptomic analysis of the forebrain of postnatal Slc13a4+/− mice. BMC Res. Notes *14*, 269. https://doi.org/10.1186/s13104-021-05687-5.
23. Milošević, M., Milićević, K., Božić, I., Lavrnja, I., Stevanović, I., Bijelić, D., Dubaić, M., Živković, I., Stević, Z., Giniatullin, R., et al. (2017). Immunoglobulins G from Sera of Amyotrophic Lateral Sclerosis Patients Induce Oxidative Stress and Upregulation of Antioxidative System in BV-2 Microglial Cell Line. Front. Immunol. *8*, 1619. https://doi.org/10.3389/fimmu.2017.01619.
24. Katzeff, J.S., Bright, F., Lo, K., Kril, J.J., Connolly, A., Crossett, B., Ittner, L.M., Kassiou, M., Loy, C.T., Hodges, J.R., et al. (2020). Altered serum protein levels in frontotemporal dementia and amyotrophic lateral sclerosis indicate calcium and immunity dysregulation. Sci. Rep. *10*, 13741. https://doi.org/10.1038/s41598-020-70687-7.
25. Tateishi, T., Yamasaki, R., Tanaka, M., Matsushita, T., Kikuchi, H., Isobe, N., Ohyagi, Y., and Kira, J.i. (2010). CSF chemokine alterations related to the clinical course of amyotrophic lateral sclerosis. J. Neuroimmunol. *222*, 76–81. https://doi.org/10.1016/j.jneuroim.2010.03.004.
26. Javier-Torrent, M., and Saura, C.A. (2020). Conventional and Non-Conventional Roles of Non-Muscle Myosin II-Actin in Neuronal Development and Degeneration. Cells *9*, 1926. https://doi.org/10.3390/cells9091926.
27. Keylock, A., Hong, Y., Saunders, D., Omoyinmi, E., Mulhern, C., Roebuck, D., Brogan, P., Ganesan, V., and Eleftheriou, D. (2018). Moyamoya-like cerebrovascular disease in a child with a novel mutation in myosin heavy chain 11. Neurology *90*, 136–138. https://doi.org/10.1212/WNL.0000000000004828.
28. Miller, S.J., Glatzer, J.C., Hsieh, Y.C., and Rothstein, J.D. (2018). Cortical astroglia undergo transcriptomic dysregulation in the G93A SOD1 ALS mouse model. J. Neurogenet. *32*, 322–335. https://doi.org/10.1080/01677063.2018.1513508.
29. Kinger, S., Dubey, A.R., Kumar, P., Jagtap, Y.A., Choudhary, A., Kumar, A., Prajapati, V.K., Dhiman, R., and Mishra, A. (2023). Molecular chaperones' potential against defective proteostasis of amyotrophic lateral sclerosis. Cells *12*, 1302. https://doi.org/10.3390/cells12091302.
30. Deane, C.A.S., and Brown, I.R. (2017). Differential Targeting of Hsp70 Heat Shock Proteins HSPA6 and HSPA1A with Components of a Protein Disaggregation/Refolding Machine in Differentiated Human Neuronal Cells following Thermal Stress. Front. Neurosci. *11*, 227. https://doi.org/10.3389/fnins.2017.00227.
31. Gaur, N., Perner, C., Witte, O.W., and Grosskreutz, J. (2020). The Chitinases as Biomarkers for Amyotrophic Lateral Sclerosis: Signals From the CNS and Beyond. Front. Neurol. *11*, 377. https://doi.org/10.3389/fneur.2020.00377.
32. Oeckl, P., Weydt, P., Thal, D.R., Weishaupt, J.H., Ludolph, A.C., and Otto, M. (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 Neuropathol. *139*, 119–134. https://doi.org/10.1007/s00401-019-02093-x.
33. Pinteac, R., Montalban, X., and Comabella, M. (2021). Chitinases and chitinase-like proteins as biomarkers in neurologic disorders. Neurol. Neuroimmunol. Neuroinflamm. *8*, e921. https://doi.org/10.1212/NXI.0000000000000921.
34. Aichholzer, F., Klafki, H.-W., Ogorek, I., Vogelgsang, J., Wiltfang, J., Scherbaum, N., Weggen, S., and Wirths, O. (2021). Evaluation of cerebrospinal fluid glycoprotein NMB (GPNMB) as a potential biomarker for Alzheimer's disease. Alzheimer's Res. Ther. *13*, 94. https://doi.org/10.1186/s13195-021-00828-1.
35. Elliott, D.A., Weickert, C.S., and Garner, B. (2010). Apolipoproteins in the brain: implications for neurological and psychiatric disorders. Clin. Lipidol. *51*, 555–573. https://doi.org/10.2217/clp.10.37.
36. Mamoor, S. (2022). Differential expression of APOC1 in amyotrophic lateral sclerosis. Preprint at OSF.
37. Muranova, L.K., Weeks, S.D., Strelkov, S.V., and Gusev, N.B. (2015). Characterization of mutants of human small heat shock protein HspB1 carrying replacements in the N-terminal domain and associated with hereditary motor neuron diseases. PLoS One *10*, e0126248. https://doi.org/10.1371/journal.pone.0126248.
38. Martinez-Merino, L., Iridoy, M., Galbete, A., Roldán, M., Rivero, A., Acha, B., Irún, P., Canosa, C., Pocoví, M., Mendioroz, M., and Jericó, I. (2018). Evaluation of Chitotriosidase and CC-Chemokine Ligand 18 as Biomarkers of Microglia Activation in Amyotrophic Lateral Sclerosis. Neurodegener. Dis. *18*, 208–215. https://doi.org/10.1159/000490920.
39. Sun, G.Y., Xu, J., Jensen, M.D., and Simonyi, A. (2004). Phospholipase A2 in the central nervous system: implications for neurodegenerative diseases. J. Lipid Res. *45*, 205–213. https://doi.org/10.1194/jlr.R300016-JLR200.
40. Shibata, N., Kakita, A., Takahashi, H., Ihara, Y., Nobukuni, K., Fujimura, H., Sakoda, S., and Kobayashi, M. (2010). Increased expression and activation of cytosolic

phospholipase A2 in the spinal cord of patients with sporadic amyotrophic lateral sclerosis. Acta Neuropathol. *119*, 345–354. https://doi.org/10.1007/s00401-009-0636-7.

41. Malada Edelstein, Y.F., Solomonov, Y., Hadad, N., Alfahel, L., Israelson, A., and Levy, R. (2021). Early upregulation of cytosolic phospholipase A2α in motor neurons is induced by misfolded SOD1 in a mouse model of amyotrophic lateral sclerosis. J. Neuroinflammation *18*, 274. https://doi.org/10.1186/s12974-021-02326-5.
42. Chow, V.T., and Lee, S.S. (1996). DENN, a novel human gene differentially expressed in normal and neoplastic cells. DNA Sequence *6*, 263–273. https://doi.org/10.3109/10425179609020873.
43. Lim, K.M., and Chow, V.T.K. (2002). Induction of marked apoptosis in mammalian cancer cell lines by antisense DNA treatment to abolish expression of DENN (differentially expressed in normal and neoplastic cells). Mol. Carcinog. *35*, 110–126. https://doi.org/10.1002/mc.10082.
44. Del Villar, K., and Miller, C.A. (2004). Down-regulation of DENN/MADD, a TNF receptor binding protein, correlates with neuronal cell death in Alzheimer's disease brain and hippocampal neurons. Proc. Natl. Acad. Sci. USA *101*, 4210–4215. https://doi.org/10.1073/pnas.0307349101.
45. Sun, Y., Liu, T., and Bai, W. (2022). MAF bZIP Transcription Factor B (MAFB) Protected Against Ovalbumin-Induced Allergic Rhinitis via the Alleviation of Inflammation by Restoring the T Helper (Th) 1/Th2/Th17 Imbalance and Epithelial Barrier Dysfunction. J. Asthma Allergy *15*, 267–280. https://doi.org/10.2147/JAA.S335560.
46. Krämer, A., Green, J., Pollard, J., and Tugendreich, S. (2014). Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics *30*, 523–530. https://doi.org/10.1093/bioinformatics/btt703.
47. Mu, M.-D., Qian, Z.-M., Yang, S.-X., Rong, K.-L., Yung, W.-H., and Ke, Y. (2020). Therapeutic effect of a histone demethylase inhibitor in Parkinson's disease. Cell Death Dis. *11*, 927. https://doi.org/10.1038/s41419-020-03105-5.
48. Sui, A., Xu, Y., Li, Y., Hu, Q., Wang, Z., Zhang, H., Yang, J., Guo, X., and Zhao, W. (2017). The pharmacological role of histone demethylase JMJD3 inhibitor GSK-J4 on glioma cells. Oncotarget *8*, 68591–68598. https://doi.org/10.18632/oncotarget.19793.
49. Sakaki, H., Okada, M., Kuramoto, K., Takeda, H., Watarai, H., Suzuki, S., Seino, S., Seino, M., Ohta, T., Nagase, S., et al. (2015). GSKJ4, A Selective Jumonji H3K27 Demethylase Inhibitor, Effectively Targets Ovarian Cancer Stem Cells. Anticancer Res. *35*, 6607–6614.
50. Bennett, S.A., Tanaz, R., Cobos, S.N., and Torrente, M.P. (2019). Epigenetics in amyotrophic lateral sclerosis: a role for histone post-translational modifications in neurodegenerative disease. Transl. Res. *204*, 19–30. https://doi.org/10.1016/j.trsl.2018.10.002.
51. Llombart, V., and Mansour, M.R. (2022). Therapeutic targeting of "undruggable" MYC. EBioMedicine *75*, 103756. https://doi.org/10.1016/j.ebiom.2021.103756.
52. Ju, Q., Li, X., Zhang, H., Yan, S., Li, Y., and Zhao, Y. (2020). NFE2L2 Is a Potential Prognostic Biomarker and Is Correlated with Immune Infiltration in Brain Lower Grade Glioma: A Pan-Cancer Analysis. Oxid. Med. Cell. Longev. *2020*, 3580719. https://doi.org/10.1155/2020/3580719.
53. Tam, O.H., Rozhkov, N.V., Shaw, R., Kim, D., Hubbard, I., Fennessey, S., Propp, N., NYGC ALS Consortium, Fagegaltier, D., Harris, B.T., et al. (2019). Postmortem cortex samples identify distinct molecular subtypes of ALS: retrotransposon activation, oxidative stress, and activated glia. Cell Rep. *29*, 1164–1177.e5. https://doi.org/10.1016/j.celrep.2019.09.066.
54. Steiner, J.P., Bachani, M., Malik, N., DeMarino, C., Li, W., Sampson, K., Lee, M.H., Kowalak, J., Bhaskar, M., Doucet-O'Hare, T., et al. (2022). Human endogenous retrovirus K envelope in spinal fluid of amyotrophic lateral sclerosis is toxic. Ann. Neurol. *92*, 545–561. https://doi.org/10.1002/ana.26452.
55. Howe, K.L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., Bhai, J., et al. (2021). Ensembl 2021. Nucleic Acids Res. *49*, D884–D891. https://doi.org/10.1093/nar/gkaa942.
56. Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. *16*, 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2.
57. Ruggieri, A., Maldener, E., Sauter, M., Mueller-Lantzsch, N., Meese, E., Fackler, O.T., and Mayer, J. (2009). Human endogenous retrovirus HERV-K(HML-2) encodes a stable signal peptide with biological properties distinct from Rec. Retrovirology *6*, 17. https://doi.org/10.1186/1742-4690-6-17.
58. Turner, G., Barbulescu, M., Su, M., Jensen-Seaman, M.I., Kidd, K.K., and Lenz, J. (2001). Insertional polymorphisms of full-length endogenous retroviruses in humans. Curr. Biol. *11*, 1531–1535. https://doi.org/10.1016/S0960-9822(01)00455-9.
59. Giglio, S., Broman, K.W., Matsumoto, N., Calvari, V., Gimelli, G., Neumann, T., Ohashi, H., Voullaire, L., Larizza, D., Giorda, R., et al. (2001). Olfactory receptor–gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. Am. J. Hum. Genet. *68*, 874–883. https://doi.org/10.1086/319506.
60. Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Federhen, S., et al. (2012). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. *40*, D13–D25. https://doi.org/10.1093/nar/gkr1184.
61. Salm, M.P.A., Horswell, S.D., Hutchison, C.E., Speedy, H.E., Yang, X., Liang, L., Schadt, E.E., Cookson, W.O., Wierzbicki, A.S., Naoumova, R.P., and Shoulders, C.C. (2012). The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. Genome Res. *22*, 1144–1153. https://doi.org/10.1101/gr.126037.111.
62. Seczynska, M., Bloor, S., Cuesta, S.M., and Lehner, P.J. (2022). Genome surveillance by HUSH-mediated silencing of intronless mobile elements. Nature *601*, 440–445. https://doi.org/10.1038/s41586-021-04228-1.
63. Russ, E., Mikhalkevich, N., and Iordanskiy, S. (2023). Expression of human endogenous retrovirus group K (HERV-K) HML-2 correlates with immune activation of macrophages and type I interferon response. Microbiol. Spectr. *11*, e0443822. https://doi.org/10.1128/spectrum.04438-22.
64. Gold, J., Rowe, D.B., Kiernan, M.C., Vucic, S., Mathers, S., van Eijk, R.P.A., Nath, A., Garcia Montojo, M., Norato, G., Santamaria, U.A., et al. (2019). Safety and tolerability of Triumeq in amyotrophic lateral sclerosis: the Lighthouse trial. Amyotroph. Lateral Scler. Frontotemporal Degener. *20*, 595–604. https://doi.org/10.1080/21678421.2019.1632899.
65. Shaw, P.J., Ince, P.G., Falkous, G., and Mantle, D. (1995). Oxidative damage to protein in sporadic motor neuron disease spinal cord. Ann. Neurol. *38*, 691–695. https://doi.org/10.1002/ana.410380424.
66. Tohgi, H., Abe, T., Yamazaki, K., Murata, T., Ishizaki, E., and Isobe, C. (1999). Remarkable increase in cerebrospinal fluid 3-nitrotyrosine in patients with sporadic amyotrophic lateral sclerosis. Ann. Neurol. *46*, 129–131. https://doi.org/10.1002/1531-8249(199907)46:1<129::aid-ana21>3.0.co;2-y.
67. Simpson, E.P., Henry, Y.K., Henkel, J.S., Smith, R.G., and Appel, S.H. (2004). Increased lipid peroxidation in sera of ALS patients: a potential biomarker of disease burden. Neurology *62*, 1758–1765. https://doi.org/10.1212/WNL.62.10.1758.
68. Eshima, J., O'Connor, S.A., Marschall, E., NYGC ALS Consortium, Bowser, R., Plaisier, C.L., and Smith, B.S. (2023). Molecular subtypes of ALS are associated with differences in patient prognosis. Nat. Commun. *14*, 95. https://doi.org/10.1038/s41467-022-35494-w.
69. Liu, X., Lu, D., Bowser, R., and Liu, J. (2016). Expression of Carbonic Anhydrase I in Motor Neurons and Alterations in ALS. Int. J. Mol. Sci. *17*, 1820. https://doi.org/10.3390/ijms17111820.
70. Lemon, N., Canepa, E., Ilies, M.A., and Fossati, S. (2021). Carbonic Anhydrases as Potential Targets Against Neurovascular Unit Dysfunction in Alzheimer's Disease and Stroke. Front. Aging Neurosci. *13*, 772278. https://doi.org/10.3389/fnagi.2021.772278.
71. Fang, Y., Yang, J., Zu, G., Cong, C., Liu, S., Xue, F., Ma, S., Liu, J., Sun, Y., and Sun, M. (2021). Junctional Adhesion Molecule-Like Protein Promotes Tumor Progression and Metastasis via p38 Signaling Pathway in Gastric Cancer. Front. Oncol. *11*, 565676. https://doi.org/10.3389/fonc.2021.565676.
72. Jia, W., Martin, T.A., Zhang, G., and Jiang, W.G. (2013). Junctional adhesion molecules in cerebral endothelial tight junction and brain metastasis. Anticancer Res. *33*, 2353–2359.
73. Mizwicki, M.T., Fiala, M., Magpantay, L., Aziz, N., Sayre, J., Liu, G., Siani, A., Chan, D., Martinez-Maza, O., Chattopadhyay, M., and La Cava, A. (2012). Tocilizumab attenuates inflammation in ALS patients through inhibition of IL6 receptor signaling. Am. J. Neurodegener. Dis. *1*, 305–315.
74. Serrano, A., Apolloni, S., Rossi, S., Lattante, S., Sabatelli, M., Peric, M., Andjus, P., Michetti, F., Carrì, M.T., Cozzolino, M., and D'Ambrosi, N. (2019). The S100A4 transcriptional inhibitor niclosamide reduces pro-inflammatory and migratory phenotypes of microglia: implications for amyotrophic lateral sclerosis. Cells *8*, 1261. https://doi.org/10.3390/cells8101261.

75. Milani, M., Mammarella, E., Rossi, S., Miele, C., Lattante, S., Sabatelli, M., Cozzolino, M., D'Ambrosi, N., and Apolloni, S. (2021). Targeting S100A4 with niclosamide attenuates inflammatory and profibrotic pathways in models of amyotrophic lateral sclerosis. J. Neuroinflammation *18*, 132. https://doi.org/10.1186/s12974-021-02184-1.
76. Hirayasu, K., and Arase, H. (2015). Functional and genetic diversity of leukocyte immunoglobulin-like receptor and implication for disease associations. J. Hum. Genet. *60*, 703–708. https://doi.org/10.1038/jhg.2015.64.
77. Kojima, K.K. (2018). Human transposable elements in Repbase: genomic footprints from fish to humans. Mobile DNA *9*, 2. https://doi.org/10.1186/s13100-017-0107-y.
78. Vargiu, L., Rodriguez-Tomé, P., Sperber, G.O., Cadeddu, M., Grandi, N., Blikstad, V., Tramontano, E., and Blomberg, J. (2016). Classification and characterization of human endogenous retroviruses; mosaic forms are common. Retrovirology *13*, 7. https://doi.org/10.1186/s12977-015-0232-y.
79. Jones, A.R., Iacoangeli, A., Adey, B.N., Bowles, H., Shatunov, A., Troakes, C., Garson, J.A., McCormick, A.L., and Al-Chalabi, A. (2021). A HML6 endogenous retrovirus on chromosome 3 is upregulated in amyotrophic lateral sclerosis motor cortex. Sci. Rep. *11*, 14283. https://doi.org/10.1038/s41598-021-93742-3.
80. Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Tissue-based map of the human proteome. Science *347*, 1260419. https://doi.org/10.1126/science.1260419.
81. Szebényi, K., Wenger, L.M.D., Sun, Y., Dunn, A.W.E., Limegrover, C.A., Gibbons, G.M., Conci, E., Paulsen, O., Mierau, S.B., Balmus, G., and Lakatos, A. (2021). Human ALS/FTD brain organoid slice cultures display distinct early astrocyte and targetable neuronal pathology. Nat. Neurosci. *24*, 1542–1554. https://doi.org/10.1038/s41593-021-00923-4.
82. Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics *21*, 3940–3941. https://doi.org/10.1093/bioinformatics/bti623.
83. Kuhn, M. (2008). Building predictive models in R using the caret package. J. Stat. Software *28*, 1–26. https://doi.org/10.18637/jss.v028.i05.
84. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. J. Stat. Software *33*, 1–22. https://doi.org/10.18637/jss.v033.i01.
85. Harrell, F., and Dupont, C. (2024). Hmisc: Harrell Miscellaneous (CRAN).
86. Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191. https://doi.org/10.1038/nprot.2009.97.
87. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the Tidyverse. J. Open Source Softw. *4*, 1686. https://doi.org/10.21105/joss.01686.
88. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47. https://doi.org/10.1093/nar/gkv007.
89. Kassambara, A. (2023). Ggpubr: "Ggplot2" Based Publication Ready Plots (CRAN).
90. Meyer, D., Zeileis, A., and Hornik, K. (2006). The Strucplot framework: visualizing multi-way contingency tables with vcd. J. Stat. Software *17*, 1–48. https://doi.org/10.18637/jss.v017.i03.
91. Aragon, T., Fay, M., Wollschlaeger, D., and Omidpanah, A. (2020). Epitools: Epidemiology Tools (CRAN).
92. Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M., and Sergushichev, A. (2016). Fast gene set enrichment analysis. Preprint at bioRxiv. https://doi.org/10.1101/060012.17.
93. Humphrey, J., Venkatesh, S., Hasan, R., Herb, J.T., de Paiva Lopes, K., Küçükali, F., Byrska-Bishop, M., Evani, U.S., Narzisi, G., Fagegaltier, D., et al. (2023). Integrative transcriptomic analysis of the amyotrophic lateral sclerosis spinal cord implicates glial activation and suggests new risk genes. Nat. Neurosci. *26*, 150–162. https://doi.org/10.1038/s41593-022-01205-3.
94. Tam, O.H., Rozhkov, N.V., Shaw, R., Kim, D., Hubbard, I., Fennessey, S., Propp, N., NYGC ALS Consortium, Fagegaltier, D., Harris, B.T., et al. (2019). Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia. Cell Rep. *29*, 1164–1177.e5. https://doi.org/10.1016/j.celrep.2019.09.066.
95. Chiò, A., Logroscino, G., Hardiman, O., Swingler, R., Mitchell, D., Beghi, E., and Traynor, B.G.; Eurals Consortium (2009). Prognostic factors in ALS: A critical review. Amyotroph Lateral Scler. *10*, 310–323. https://doi.org/10.3109/17482960802566824.
96. Gregorich, M., Strohmaier, S., Dunkler, D., and Heinze, G. (2021). Regression with Highly Correlated Predictors: Variable Omission Is Not the Solution. Int. J. Environ. Res. Publ. Health *18*, 4259. https://doi.org/10.3390/ijerph18084259.
97. Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B., and Lewis, S.; AmiGO Hub; Web Presence Working Group (2009). AmiGO: online access to ontology and annotation data. Bioinformatics *25*, 288–289. https://doi.org/10.1093/bioinformatics/btn615.
98. Cunningham, F., Allen, J.E., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Austine-Orimoloye, O., Azov, A.G., Barnes, I., Bennett, R., et al. (2022). Ensembl 2022. Nucleic Acids Res. *50*, D988–D995. https://doi.org/10.1093/nar/gkab1049.
99. Madeira, F., Pearce, M., Tivey, A.R.N., Basutkar, P., Lee, J., Edbali, O., Madhusoodanan, N., Kolesnikov, A., and Lopez, R. (2022). Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Res. *50*, W276–W279. https://doi.org/10.1093/nar/gkac240.
100. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.
101. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. *50*, D439–D444. https://doi.org/10.1093/nar/gkab1061.
102. Pettersen, E.F., Goddard, T.D., Huang, C.C., Meng, E.C., Couch, G.S., Croll, T.I., Morris, J.H., and Ferrin, T.E. (2021). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci. *30*, 70–82. https://doi.org/10.1002/pro.3943.
103. Beimforde, N., Hanke, K., Ammar, I., Kurth, R., and Bannert, N. (2008). Molecular cloning and functional characterization of the human endogenous retrovirus K113. Virology *371*, 216–225. https://doi.org/10.1016/j.virol.2007.09.036.
104. Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., et al. (2021). Highly accurate protein structure prediction for the human proteome. Nature *596*, 590–596. https://doi.org/10.1038/s41586-021-03828-1.
105. Rose, J., and Eisenmenger, F. (1991). A fast unbiased comparison of protein structures by means of the Needleman-Wunsch algorithm. J. Mol. Evol. *32*, 340–354. https://doi.org/10.1007/BF02102193.
106. Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E.H., Britto, R., Bye-A-Jee, H., Cukura, A., et al. (2023). UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. *51*, D523–D531. https://doi.org/10.1093/nar/gkac1052.
107. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Software *33*, 1–22. https://doi.org/10.18637/jss.v033.i01.
108. Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. J. Stat. Software *39*, 1–13. https://doi.org/10.18637/jss.v039.i05.
109. Tay, J.K., Narasimhan, B., and Hastie, T. (2023). Elastic Net Regularization Paths for All Generalized Linear Models. J. Stat. Software *106*, 1. https://doi.org/10.18637/jss.v106.i01.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Postmortem cortex and spinal cord samples	NYGC ALS Consortium	N/A
Deposited data		
NYGC ALS Consortium bulk RNA sequencing files	GEO	GEO: GSE137810
Software and algorithms		
ROCR v1.0-11	Sing et al.[82]	http://ipa-tys.github.io/ROCR/
Caret v6.0-94	Kuhn et al.[83]	https://github.com/topepo/caret/
glmnet v4.1-7	Friedman et al.[84]	https://glmnet.stanford.edu
TEtranscripts	Jin et al.[9]	https://github.com/mhammell-laboratory/TEtranscripts
Hmisc	Harrell et al.[85]	https://hbiostat.org/r/hmisc/
DESeq2 v1.38.3	Love et al.[21]	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
biomaRt v2.54.1	Durinck et al.[86]	https://www.bioconductor.org/packages/release/bioc/html/biomaRt.html
tidyverse v2.0.0	Wickham et al.[87]	https://github.com/tidyverse/tidyverse
LIMMA v3.54.2	Ritchie et al.[88]	https://bioinf.wehi.edu.au/limma/
ggpubr v0.6.0	Kassambara et al.[89]	https://rpkgs.datanovia.com/ggpubr/
vcd v1.4-11	Meyer et al.[90]	https://cran.r-project.org/web/packages/vcd/index.html
epitools v0.5–10.1	Aragon et al.[91]	https://cran.r-project.org/web/packages/epitools/index.html
fgsea v1.24.0	Korotkevich et al.[92]	https://bioconductor.org/packages/release/bioc/html/fgsea.html
DEA and LR Scripts	This Paper	https://github.com/nickap5/Endogenous-Retroviruses-are-Dysregulated-in-ALS-

RESOURCE AVAILABILITY

Lead contact

Further information regarding resources related to this manuscript should be directed to the corresponding author, Dr Avi Nath (avindra.nath@nih.gov)

Materials availability

This study did not generate any new materials or reagents. Additional information and requests for resources related to this mansucript should be directed to and will be fulfilled by the lead contact, Dr Avi Nath (avindra.nath@nih.gov).

Data and code availability

- RNA-seq data used in this manuscript is publicly available as of the date of publication. The GEO ID is listed in the key resources table.
- This paper does not report the generation of new code libraries, although general R scripts used for this analysis are publicly accessible on GitHub at: https://github.com/nickap5/Endogenous-Retroviruses-are-Dysregulated-in-ALS.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Patients and samples

Bulk RNA-sequencing from postmortem spinal cord and cortical tissues of ALS patients and unaffected controls was provided by the New York Genome Center (NYGC) ALS Consortium. All samples were obtained via the appropriate IRB protocols for each participating institution and sent to the NYGC according to the governing laws and regulations. The demographics, including age and ancestry, of the ALS patients and unaffected controls as well as spinal cord and cortical samples used in this study are summarized in Table S11 and Figure S17. This table was generated based on information found in the meta data spreadsheet provided to our group by the NYGC. One control individual had an "unknown" biological sex; however, this individual's sex genotype was male (XY), so they are considered male for the purposes of this study. Overall, there was a lot of variability for age of symptom onset and disease course between samples. However, between cortical and spinal cord samples, there were fewer differences with respect to these attributes and the percentage of patients with potential genetic predispositions to ALS as well as site of onset. Of note, there were CNS regions other than SC and CTX (i.e., hippocampus, cerebellum, and choroid) that are included in the all individuals column. Unless otherwise noted, all descriptive statistics are based on comparisons at the patient-level, as opposed to the sample-level, to avoid biasing the descriptive statistics toward patients that contributed more samples from multiple tissue types. Individuals contributed between 2 and 3 cortical and spinal cord samples on average. The meta data for comorbidities in the provided meta data file was not in a format that was conducive for analysis: the amount of detail and way in which disease were described varied considerably between participating sites. Therefore, it was challenging to draw useful general conclusions.

Sample preparation

The NYGC provides high quality data even though it represents a multi-center consortium by utilizing comprehensive protocols for tissue collection, sample and library processing, and quality control. There are published papers detailing the methods used by different consortium members.[93,94] Briefly, RNA was extracted from frozen, postmortem samples using the Thermo Fisher Scientific (MA, USA) TRIzol-based Qiagen RNeasy Kit (Hilden, Germany). The sequencing libraries were made using 500 ng of RNA that was depleted for ribosomal RNA using the KAPA Stranded RNA-Seq Kit with RiboErase (Roche, Basel, Switzerland). Libraries were pooled with an average insert size of 375 bp's, standardized quality control guidelines were met, and were sequenced on Illumina HiSeq 2500 and Illumina NovaSeq platforms (CA, USA). The average sequencing depth of the samples was 42 million read pairs and only tissues that passed selection based on 250 markers for tissue, CNS region, and biological sex, were included.

Unless otherwise noted, all correlations were performed using the Spearman method in base R and plotted using the ggpubr and ggplot2 libraries in R. The default functionality of the package is to use all possible datapoints with valid values for the two variables being plotted, so unless otherwise noted, sample sizes should not be presumed to be equivalent. A linear regression line (blue) with 95% confidence interval (CI) in gray is also plotted.

METHOD DETAILS

Differential expression analysis

First, samples in the count matrix file that were collected from the CNS region of interest (either SC or CTX) were selected. Next, genes with a low number of transcript counts (less than 6) were filtered out as well as those features that were significantly associated with biological sex based on a DESeq2 FDR adjusted p value <0.05. Analyses were performed comparing all ALS patient samples to controls as well as male and female ALS patients vs. male and female controls separately. For the DEAs combining male and female samples, filtering for sex-associated features was required as biological sex explains most of the transcriptional variation between samples. Figure S23 shows the first two principal components from a principal component analysis (PCA) of the top 500 features by highest variance in the CTX and SC prior to filtering out sex-associated features. Figure S24 shows the same plots after filtering out sex-associated features. The clear separation of male and female samples which accounted for 25–32% of variance across samples (Figure S23) was adequately accounted for by filtering (Figure S24). PCA's were plotted using DESeq2's plotPCA function which selects the top 500 features based on row variance.

DEA was performed on the resulting non-normalized count data matrix using the DESeq2 library in R[21] to compare the expression profiles of ALS patients and controls. DEGs were counted as genes with an FDR adjusted p-value (qval) < 0.05. Unless indicated otherwise, all adjusted p-values were corrected using FDR. The generalized linear model for the DEA included terms for participating institutional site, tissue subregion, *C9orf72* status, age at death (as factor), RNA integrity number (RIN), and ALS or control status. Sex-associated features needed to be filtered out as opposed to being included in the generalized linear model (GLM) since sex-associated features accounted for a lot of variation between samples even if included in the GLM. Other potential sources of variability were not included in the GLM due to their association with variables that were included in the model (e.g., postmortem interval, PMI, and age of symptom onset) or limited information in meta data.

To determine which continuous variables to include in the model, a correlation matrix of spearman's rho statistics was computed using the "rcorr" function of the "Hmisc" library in R on all individual patients and control samples available in the provided NYGC meta data (N = 476). This represents a slightly larger sample size than the actual RNA-seq data analyzed in this study due to the specific CNS regions we studied and our not incorporating all possible samples from NYGC. Significant correlations were observed for RIN and PMI (R = −0.25, qval = 1.4E-4) and pH (R = −0.39, qval = 0.0017); age at symptom onset and age at death (R = 0.97, qval ≈0) and disease duration (R = −0.3, qval = 6.7E-6); and age at death and disease duration (R = −0.14, qval = 0.495) (Table S12; Figure S25). These results suggest that ALS patients that develop the disease earlier in life have a longer disease duration, an association supported by previous studies.[95] These patterns of

association indicate that there are two distinct types of variables in this meta data: RNA quality (i.e., RIN, PMI, and pH) and disease course (i.e., disease duration, age at symptom onset, and age at death). While there are significant associations between variables within a group, there were no significant associations between variables from different groups (e.g., pH and age at death).

Since the sample size for the various pairwise Spearman correlations varied, the total number of patients and controls (N = 476) was split into a smaller subset so only those subjects with complete information for all six variables (N = 63) were considered and the same pairwise analysis was performed (Table S13). Only associations between age at symptom onset and age at death (R = 0.98, qval ≈0) and pH and PMI (R = 0.395, qval = 0.01) were significant in this analysis.

The reduced significance levels could be due to the lower sample size as well as the fact that reducing the number of subjects to only those with complete information led to biasing of the data by the participating site that provided the samples (Figure S24). For example, Academic Medical Center (AMC) and University College London (UCL) were the two participating sites that provided the most samples, whereas Johns Hopkins University (JHU) and University of California San Diego (UoCSD) provided the most samples only for the subgroup of individuals that had complete information for the meta data variables of interest.

While the inclusion of highly correlated variables in linear models is still a topic of debate, whether to incorporate correlated variables or not seems to depend greatly on the type of data and question.[96] For the purpose of this study, it was decided to include only one variable from each of the two variable types (i.e., RNA quality and disease course) due to the high degree of correlation between variables, general consensus from conversations with others in the field, previously published articles which use fewer variables in the GLM, and lack of robust meta data for each of the six variables for all patients (only 63 out of 476 patients had complete meta data for all six variables).

Frequency metric

After the DEA, a HERV-K feature set consisting of all 1,060 possible HERV-K genes from the GTF annotation file was selected. The threshold value, theta, for each possible feature is defined below where mu is the base mean of the feature, FC is the Fold Change, and σFC is the standard error of the Fold Change:

$$\theta = |\mu| * |FC| * (4 * \sigma FC)$$

The threshold value for each feature was compared to the median of ratios normalized count value calculated as part of the DESeq2 pipeline. The frequency value represents the number of super-threshold HERV-K features, out of all 1,060 possible HERV-K features, for a particular individual. This metric is agnostic to protein-coding potential of features, sample type, and the degree to which a feature is over-expressed.

To determine whether there were any differences in terms of this general metric of HERV-K feature expression between ALS patients and controls, the frequency values from 200 ALS patients and 200 unaffected control samples were randomly sampled. The median of the unaffected samples was subtracted from the median of the ALS samples and the two sample groups were compared via a Mann-Whitney U test. Mann-Whitney test p-values were adjusted using the false discovery rate (FDR) method for multiple comparison. This process was repeated 100 times.

It was clear from examining the distribution of frequency values that there were unimodal distributions of this value for both ALS patients and unaffected controls. Therefore, differences in HERV-K expression would likely be more subtle. To identify the most appropriate subpopulation of patients that would be distinct in terms of their HERV-K expression based on the frequency metric, the proportion of samples in each group (ALS and controls) was optimized using the difference in median frequency value between a number of patient subsets and a comparison group as well as the Mann-Whitney U test significance level between the two groups. The effect size monotonically decreased with increasing proportion of samples, as expected from the unimodal distribution and minimal overall differences between ALS patients and controls. Therefore, the smallest proportion of samples that produced the most significant p-value was used as the high HERV-K-expressing group.

Pathway and gene set analysis

Pathway analysis was performed using QIAGEN IPA (QIAGEN Inc., https://digitalinsights.qiagen.com/IPA). IPA's canonical pathways and causal network analyses were used to draw inferences about the biological pathway dysregulation and potential druggable or diagnostic targets.[46] The top 1,550 genes based on p-value were used for the analysis. This number was determined by selecting a DEA result table and trying several different numbers of features toward the middle of the feature range suggested by QIAGEN (200–3000) until one was found to provide the most significantly up-or down-regulated pathways. This optimization was only done for one preliminary analysis and this number of features was then used for all subsequent analyses. Core analysis was performed using DESeq2 $\log_2$ fold change for the calculation of Z-scores. IPA upstream regulators analysis was also performed to identify drugs, transcription factors, and other molecules that could be affecting transcriptional states in patient populations of interest. IPA predicts the state of these elements by examining the pattern of expression of multiple genes in the RNA-seq data that are known to be downstream of the regulator. This analysis was used to nominate possible druggable and/or diagnostic targets. The top 10 most dysregulated pathways/regulators, based on the sum of the absolute value of the Z score across groups, are displayed in a heatmap. Pathways/regulators are arranged from top to bottom in alphabetical order.

The fast gene set enrichment analysis (FGSEA) library in R[92] was used to assess the degree and significance of gene sets not adequately captured by IPA, such as those involving HERV-K loci. This analysis does not rely on a cut-off for significance. A combination of curated and custom gene sets derived from the total HERV-K universe were used. This included all HERV-K genes, HERV-K genes that encode for partial or full envelope, and HML-2 full-length envelope-coding loci. The HML-2-specific gene set had too few genes to be robust (N = 12). The curated gene sets were derived from the gene ontology (GO) tool AmiGO[97] and were accessed using the biomaRt package in R.[86] When searching for

GO's on AmiGO, multiple interchangeable GOs were often found. In these cases, the GO with the greatest number of ENSGs that matched our annotation file was used. The ENSG gene sets included synaptic vesicle (GO:0008021), neuronal apoptosis (GO:0051402), neuronal death (GO:1901214), motor neuronal death (GO:0097049), presynaptic active zone (GO:0048786), postsynaptic membrane (GO:0045211), gliogeneisis (GO:0060252), oxidative stress (GO: 0006979), transcriptional regulation (GO:0140110), and inflammation (GO:0006954) GOs.

The order of the genes used for GSEA was determined by the rank value based on DEA statistics. The rank formula, where R is the rank, FC is the fold change from DESeq2 and pval is the unadjusted *p*-value from the DESeq2 was:

$$R = |log_2\ FC| * -log_{10}(pval)$$

Genome browser

The Ensembl project[98] was used to further analyze loci implicated in the DEA using the hg38 human reference genome. RepeatMasker (for repetitive elements), Regulatory Build (for regions with possible epigenomic and regulatory effects), Genome Aggregate Database (for SNP, single nucleotide polymorphism, variants), and GENCODE 42 (additional annotation) tracks were added and used to analyze significantly dysregulated loci at the genomic level.

Protein-level analysis

The European Molecular Biology Open Software Suite (EMBOSS) was used to convert DNA sequences of interest to RNA.[56,99] RNA sequences were then translated to amino acid sequences using EMBOSS Transeq. These AA sequences were then matched with the AlphaFold2 protein database using EBI FastA.[100,101]

Since none of the amino acid sequences of interest had an over 90% percent identity score with existing proteins in the database, the macromolecular structure was modeled with AlphaFold2 and visualized using ChimeraX.[102] The default monomer model for AlphaFold2 was used with a max template date of "2021-05-14". To compare the truncated env structures with a known reference structure, HERVK-113 (19p13.11), which is a well-studied HERV with ORFs for all viral proteins and is present in 30% of the population.[103] AlphaFold2 also reports the per-residue confidence in its predicted structure via the predicted local distance difference test (pLDDT). This is a number between 0 (lowest confidence) to 100 (highest confidence) that was validated and generally corresponds to how well the predicted structure would agree with an experimentally produced structure of the same molecule.[104] The structure of HERVK-113 was overlayed with the truncated env of interest using Matchmaker in ChimeraX, which aligns sequences by pairwise sequence alignment.

To determine the location and domains of the proteins, a pairwise sequence alignment between the HERV-K protein identified by FastA and the AA sequence generated as previously described was performed using EMBOSS Needle, which uses the Needleman-Wunsch algorithm for sequence alignment.[105] To determine which specific Env domain was relevant to a locus, the location of the best alignment to a reference Env protein was then compared to the domain location information of that Env protein using UniProt.[106]

Logistic regression

TE, ENSG and combined (TE and ENSG) count matrices were created and filtered for low abundance features (<6 reads across samples) and for features that were significantly associated with biological sex (DESeq2 FDR adjusted p-val <0.05). The counts were then converted into median of ratios (MoR) normalized counts using DESeq2. The logistic regression model was trained using 15 different input conditions: four DEA based (top 500 based on DEA *p*-value significance), and 11 random models. The DEA models consisted of the overall 500 best features (Sig), the top 250 ENSG and 250 TE features (Balanced), and the top 500 ENSG only (ENSG) and TE only (ERV) features. For the random models, one was 500 features selected at random (Random), the other ten were 250 random ENSGs and 250 random TEs (Balanced Random) with the seed of the R random number generator specified by the trial number (e.g., trial 1 had a seed value of 1). For non-balanced random models, the random seed was set to 1. This was performed so that the random trials could be reproduced and so the unbalanced nature of the feature set would not negatively affect the performance of the random model.

Features incorporating DEA-based information (i.e., Sig) were selected using the Rank value described in the GSEA section. In brief, for non-random models, the corresponding all ALS patient vs. control DEA values were ranked according to the product of the -$\log_2$*p*-value and the absolute value of the $\log_2$FC.

Prior to training the LR model, the data was split 70% in training and 30% in testing groups. The glmnet package in R[107] was used for regularized logistic regression. A ridge penalty ($\alpha = 0$) was used to improve generalizability of the model. The ridge penalty was selected to allow the greatest number of features to be used in the final fitted model. Briefly, the library uses a "proximal Newton" algorithm with an inner loop (weighted coordinate descent) and an outer loop (quadratic approximation) to resolve the appropriate penalty (iteratively reweighted penalized least squares). The specific objective function used by glmnet on the log-odds transformed logistic regression function is explained by the authors of the library in their previously published works.[84,108,109] In general, the elastic net penalty can be represented as below where P is the penalty, λ is the regularization tuning parameter, β_1 is the Lasso (L1) penalty, β_2^2 is the Ridge (L2) penalty, and α is the elastic net mixing parameter:

$$P_\alpha(\beta) = \lambda\left[\frac{1-\alpha}{2}\left(\beta_2^2\right) + \alpha(\beta_1)\right]$$

The probability cut-off of the regression model used for selecting a patient was set to 0.7. This was because the random models in the spinal cord were heavily skewed toward selecting patients. To keep all models comparable, this same cut-off value was used for other models,

although this decision probably led to a model that was less accurate for other conditions. 5-fold cross-validated logistic regression was performed with median of ratios counts as predictors and ALS patient or control status as outcomes. The value of λ used for the fitted model was the value of λ with the minimum mean cross-validated error. The validation loss metric used was mean squared error (MSE). The efficacy of the regularized logistic regression model for each condition was assessed via the following R functions and libraries: confusion matrix (caret library), 4-fold plot, and ROC (receiver operating characteristic) curve (ROCR library).

For HERV-K-based models, the number of DEGs that met the filtering cut-offs for DESeq2 were too few (<100), so random HERV-K features were also added. In the CTX, there were 15 HERV-K features in the DESeq2 results and 485 random HERV-K features. In the SC, there were 65 HERV-K features from DESeq2 and 435 random features.

The number of features to use as input for the LR models was determined based on characteristics of the distribution itself to preserve as much data as possible for the model. We determined the optimum number to be 500 by rounding up the number of features that were outliers. Outliers (O) were features with a rank value greater than the mean (μ) plus two standard deviations (σ) of all rank values:

$$O > \mu + 2(\sigma)$$

Since the goal was to use models with as similar parameters as possible, the same number of features was also used for the cortex. Figure S25 shows the distribution of the $\log_{10}$ of the rank values (429 outliers in SC and 141 in CTX). As with the LR probability cutoff, accuracy would probably be improved with separate optimization strategies, but this would have resulted in less comparable models between the CTX and SC.

To assess the performance metrics of the models, several metrics are reported which are shown below where TP = true positives, TN = true negatives, FP = false positives, FN = false negative, NPV = negative predictive value, PPV = Positive Predictive value or precision, and recall = sensitivity:

$$Sensitivty = \frac{TP}{(TP+FN)}$$

$$Specificity = \frac{TN}{(TN+FP)}$$

$$PPV = \frac{TP}{(TP+FP)}$$

$$NPV = \frac{TN}{(TN+FN)}$$

$$F1 = 2\left(\frac{Precision * Recall}{Precision+Recall}\right)$$

$$Prevalence = \frac{(TP+FN)}{(TP+TN+FP+FN)}$$

$$Detection\ Rate = \frac{TP}{(TP+TN+FP+FN)}$$

$$Detection\ Prevalence = \frac{(TP+FP)}{(TP+TN+FP+FN)}$$

$$Balanced\ Accuracy = \frac{(Sensitivity+Specificity)}{2}$$

QUANTIFICATION AND STATISTICAL ANALYSIS

All data storage and analysis related to this project was performed on the NIH high performance computing cluster (HPC) Biowulf (http://hpc.nih.gov). All code was run using R version 3.6.1 and 4.2.2 (the change in version number was due to a Biowulf upgrade midway through the project that resulted in a lack of support for R version 3) on Biowulf. Necessary R libraries that are not part of the base distribution were downloaded using the instructions for each package on the Comprehensive RArchive Network (https://cran.r-project.org) and include (versions based on R version 4.2.2) ROCR 1.0–11,[82] caret 6.0–94,[83] glmnet 4.1–7,[84] DESeq2 1.38.3,[21] Hmisc,[85] biomaRt 2.54.1,[86] tidyverse 2.0.0,[87] LIMMA 3.54.2,[88] ggpubr 0.6.0,[89] vcd 1.4–11,[90] epitools 0.5–10.1,[91] and fgsea 1.24.0.[92]

Statistical significance was defined as comparisons with a false detection rate (FDR) adjusted *p*-value less than 0.05, unless noted otherwise. Details for each analysis type are outlined above and referenced in the appropriate sections of the main manuscript.