Integrative transcriptomic analysis of the amyotrophic lateral sclerosis spinal cord implicates glial activation and suggests new risk genes

Integrative transcriptomic analysis of the amyotrophic lateral sclerosis spinal cord implicates glial activation and suggests new risk genes

Jack Humphrey[1,2,3,4*], Sanan Venkatesh[1,3,5], Rahat Hasan[1,2,3,4], Jake T. Herb[6], Katia de Paiva Lopes[1,2,3,4], Fahri Küçükali[7,8], Marta Byrska-Bishop[9], Uday S. Evani[9], Giuseppe Narzisi[9], Delphine Fagegaltier[9,10], NYGC ALS Consortium[#], Kristel Sleegers[7,8], Hemali Phatnani[9,10,11], David A. Knowles[9,12], Pietro Fratta[13], Towfique Raj[1,2,3,4*]

1. Nash Family Department of Neuroscience & Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA
2. Ronald M. Loeb Center for Alzheimer's disease, Icahn School of Medicine at Mount Sinai, New York, NY, USA
3. Department of Genetics and Genomic Sciences & Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, NY, USA
4. Estelle and Daniel Maggin Department of Neurology, Icahn School of Medicine at Mount Sinai, New York, NY, USA
5. Department of Psychiatry, Pamela Sklar Division of Psychiatric Genomics, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA
6. Graduate School of Biomedical Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA
7. Complex Genetics of Alzheimer's Disease Group, Center for Molecular Neurology, VIB, Antwerp, Belgium
8. Department of Biomedical Sciences, University of Antwerp, Antwerp, Belgium
9. New York Genome Center, New York, NY, USA
10. Center for Genomics of Neurodegenerative Disease, New York Genome Center, New York, NY, USA
11. Department of Neurology, Columbia University Irving Medical Center, Columbia University, New York, NY, USA
12. Department of Computer Science, Columbia University, New York, NY, USA
13. Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, London, UK

#. Full list of contributors in **Supplementary Acknowledgements**.

*. Corresponding authors: Jack Humphrey (jack.humphrey@mssm.edu), Towfique Raj (towfique.raj@mssm.edu)

Abstract

Amyotrophic lateral sclerosis (ALS) is a progressively fatal neurodegenerative disease affecting motor neurons in the brain and spinal cord. Here we investigated gene expression changes in ALS via RNA-seq in 380 post-mortem samples from cervical, thoracic, and lumbar spinal cord segements from 154 individuals with ALS and 49 control individuals. We observed an increase in microglia and astrocyte gene expression, accompanied by a decrease in oligodendrocyte gene expression. By creating a gene co-expression network in the ALS samples, we identify several activated microglia modules that negatively correlate with retrospective disease duration. We map molecular quantitative trait loci and find several potential ALS risk loci that may act through gene expression or splicing in the spinal cord and assign putative cell-types for *FNBP1*, *ACSL5*, *SH3RF1* and *NFASC*. Finally, we outline how common genetic variants associated with splicing of C9orf72 act as proxies for the well-known repeat expansion,, and use the same mechanism to suggest *ATXN3* as a putative risk gene.

Introduction

Amyotrophic lateral sclerosis (ALS) is a progressively fatal neurodegenerative disease characterized by degeneration of upper and lower motor neurons that control voluntary movement via the corticospinal tract. Most patients have a disease onset in middle age but there is a wide clinical variability in onset of symptoms and the pace of disease progression before death[1]. About 5-10% of ALS cases have a family history of disease[3], with the remaining patients deemed to be sporadic. The field has focused on rare mutations of large effect size, such as large repeat expansions in the gene *C9orf72*, found in 40% of familial ALS and also in 10% of sporadic ALS cases[4]. Other rare mutations, in genes such as *SOD1*, *TARDBP*, *FUS*, *NEK1, TBK1,* and *KIF5A* make up only a small fraction of the total familial ALS population[5–8] and the majority of non-familial ALS cases have no known causative mutation. Large-scale genome-wide association studies have repeatedly found common genetic variants associated with ALS risk[8–11]. Common polymorphic short-tandem repeats are a further contributor to genetic risk of ALS, including *ATXN2*[12], and other *ATXN family members*[13–15] where intermediate repeat lengths impart a small increase in ALS risk. The interplay between rare and common genetic variants in shaping ALS risk is still being explored. Crucially, there has been little progress in assigning risk genes to particular cell-types. One method to achieve this is the mapping of molecular quantitative trait loci (QTLs), the association between common genetic variants and a molecular phenotype such as gene expression. By performing this in a relevant tissue, QTL variants can be colocalized with GWAS risk variants to identify risk genes[16]. In Alzheimer's disease, multiple studies have applied this framework to identify multiple disease risk variants as acting through gene expression and/or splicing in genes specific to microglia and monocytes[17–19].

Although motor neurons are thought to be the predominantly affected cell type within the spinal cord, much research has focused on non-neuronal contributions to disease initiation and progression. Studies using mouse models of *SOD1* mutations have identified a non-neuronal contribution to disease initiation and length of survival[20,21]. These studies and many others identified both astrocytes and microglia as being able to modify disease duration[22–26]. As motor neurons degenerate during disease they release factors which cause microglia to assume an activated pro-inflammatory state[27,28], which can then induce an activated state in astrocytes[29]. Both activated microglia and astrocytes are toxic to motor neurons[30,31], and blocking this microglia-astrocyte crosstalk extends survival in a *SOD1* mouse model[32]. Several studies have profiled gene expression in human post-mortem ALS tissues, in spinal cord[33–35], frontal cortex[35], and motor cortex[36]. These studies have identified a broad upregulation of inflammatory and immune-related genes and a downregulation in oligodendrocyte and neuron genes. Further

investigation of glial activation and neuron-glia crosstalk in the context of ALS is therefore required. However due to small sample sizes, these studies have been unable to identify more subtle changes in gene expression, nor to compare these changes with clinically variable traits, or to leverage molecular QTLs.

Results

Cellular composition changes in the ALS spinal cord

We aligned and processed post-mortem RNA-seq data from three spinal cord regions (cervical, lumbar, and thoracic) from 154 subjects with ALS and 49 non-neurological controls from the New York Genome Center ALS Consortium, contributed by 8 different medical centres. All samples underwent extensive quality control (**Supplementary Fig. 1-3**). Demographic and technical information for the donors is summarised in **Table 1;** full details are in **Supplementary Table 1**.

We performed differential gene expression between all ALS cases and controls in each spinal cord section, controlling for sex, age at death, sequencing batch, submitting site, and technical factors including RNA integrity number (RIN). At a false discovery rate (FDR) < 0.05 we found large numbers of differentially expressed genes (DEGs) in the Cervical and Lumbar regions, with 7,349 and 4,694 respectively, and only 256 in the smaller Thoracic cohort (**Fig. 1a-b; Table 2; Supplementary Table 2**). Of the genes significant in both lumbar and cervical spinal cord, 238 were upregulated with LFC > 1 in at least one of two regions, with 109 in both, all of which were more strongly upregulated in the Cervical region (**Fig. 1c)**. Although highly concordant, only 12 of those 109 genes passed FDR < 0.05 in the thoracic spinal cord, demonstrating the added benefit of our increased sample size. A smaller number of DEGs were strongly upregulated ($\log_2$ fold change > 2, equivalent to a 4-fold increase in mean expression), including *CHIT1*, *CCL18, CHRNA1, GPNMB, and LYZ,* mostly genes encoding proteins secreted by activated macrophages/microglia. *CHIT1,* encoding the enzyme chitotriosidase, *is known* to be upregulated in the cerebrospinal fluid (CSF) and plasma of ALS patients[37]. *GPNMB,* encoding glycoprotein nonmetastatic melanoma B, is upregulated at the protein level in ALS patient spinal cord, CSF, and sera[38,39] and is expressed by activated microglia[40]. A common genetic variant in *GPNMB* is associated with Parkinson's disease[41,42]. *CCL18* is a cytokine released by myeloid cells. *LYZ* encodes human lysozyme, an antibacterial protein secreted by myeloid cells. Neither *CCL18* nor *LYZ* have been previously linked to ALS. *CHRNA1*, encoding the alpha subunit of the muscle acetylcholine receptor, is a known marker of denervation of muscles in *SOD1* mouse models[43]. A marker of astrocyte activation, *C3*[29,32], was also upregulated, albeit with a lower effect size.

Of the 67 genes downregulated with LFC < -1 in at least one of the two regions, only 13 were < -1 in both, with the majority (46) more strongly downregulated in the lumbar spinal cord. The downregulated genes include the small nucleolar RNA gene *SNORD3C* and *MOBP*, a marker of oligodendrocytes (**Fig. 1d**). The motor neuron marker genes *MNX1* and *ISL1* were both downregulated in the cervical and lumbar spinal cord with LFC > -1 (FDR < 0.05). The top 20 strongest (by effect size) upregulated and downregulated genes are presented (**Fig. 1e**). 37 of the 67 most downregulated genes (55%) were non-coding, including antisense transcripts and long intergenic non-coding RNAs, compared to only 38 of the 246 (15%) of the upregulated genes.

We performed Gene Set Enrichment Analysis (GSEA)[44] using both curated molecular pathways and sets of cell-type marker genes. Using MSigDB curated pathways[45], we identified 21 pathways positively enriched in both regions (normalised enrichment score (NES) > 1; adjusted P < 0.05), which mostly reflected different immune and inflammatory pathways (**Fig. 1f; Supplementary Fig. 8**). We next performed GSEA with lists of the 100 most specific human cell-type marker genes for six major brain cell-types[46]. We observed strong positive enrichment of microglia markers, whereas oligodendrocyte markers were negatively enriched (**Fig. 1g**). Repeating the analysis with several other marker gene sets resulted in concordant results and revealed positive enrichments in endothelial cells and pericytes, despite there being low overlap between genes used in each set (**Supplementary Figure 9a-b; Supplementary Table 3**).

We then prepared a panel of immune activation genes using four studies of microglia and/or astrocyte responses to pro-inflammatory stimuli in mice. These are disease-associated microglia (DAM)[47], disease-associated astrocytes (DAA)[48], reactive astrocytes (RA)[49], and plaque-associated genes (PIG)[50]. These gene lists only partially overlap (**Supplementary Fig. 10),** and represent signatures of microglia and astrocyte responses to a range of stimuli, including amyloid plaques, neurodegeneration, hypoxia (MCAO) and lipopolysaccharide (LPS). All glial activation sets were enriched in the upregulated genes in both regions (**Fig. 1h**).

We then estimated cell-type proportions in the bulk RNA-seq using both single-nucleus and single-cell RNA-seq from human cortical samples[46,51], using two different algorithms[52,53], producing four different predictions per sample. Predictions for each cell-type were highly correlated between tools and references **(Supplementary Fig. 11-15; Supplementary Table 4)**. We highlight the deconvolution estimates for the cervical spinal cord using single-nucleus RNA-seq reference data from human frontal cortex[46] and the MuSiC algorithm[52] (**Fig. 1i**). As a further analysis of cell-type changes we ran expression-weighted cell-type enrichment (EWCE)[54] using the differentially expressed genes and the

same single-nucleus RNA-seq data[46], which confirmed the observations from deconvolution (**Supplementary Fig. 17**).

We overlapped the 7,349 cervical spinal cord DEGs (FDR < 0.05) with a recently published mass spectrometry proteomic dataset (**Supplementary Table 5**)[39]. The study performed differential protein expression in an independent cohort of post-mortem spinal cord samples from 8 ALS cases and 7 controls, and cerebrospinal fluid (CSF) from 24 cases and 16 controls. Of the 287 differentially expressed proteins found in the spinal cord (FDR < 0.05), 153 were also DEGs in our dataset (OR = 2.8; P < 1e-16, Fisher's exact test), and 137 (90%) had the same direction of effect between mRNA and protein (**Fig. 1j**). The top two most upregulated genes were *GPNMB* and *IQGAP2*. *PEX5L*, found to be highly oligodendrocyte specific in single cell and single nucleus RNA-seq[46,51] was downregulated at both RNA and protein level. In the CSF, of the 30 genes significant at the protein level, 17 were DEGs (P = 0.001), with all but one upregulated (**Fig. 1k**). *GPNMB* and *CHIT1* were both upregulated in CSF, validating their associations with ALS. As well as *GPNMB, SERPINA3* was upregulated in both RNA and protein in spinal cord and CSF. Together, these results suggest that ALS spinal cord experiences a robust inflammatory reaction driven by microglia and astrocytes, with dysregulation of oligodendrocytes.

C9orf72-ALS transcriptomes indistinguishable from sporadic ALS

Analysis of frontal cortex and cerebellum has reported distinct sets of differentially expressed genes between *C9orf72* repeat expansion carriers and sporadic ALS and/or FTD patients[55,56]. We repeated the differential expression analysis but split patients by *C9orf72* repeat expansion status, as assessed by repeat-primed PCR or estimated through ExpansionHunter[57]. Comparing each disease set to controls, the directionality of expression changes in each comparison were highly concordant within each spinal cord section (**Supplementary Fig. 19**). Directly comparing *C9orf72* carriers to sporadic ALS cases, no differentially expressed genes were observed, with the exception of *C9orf72* itself, which was downregulated in C9orf72-ALS (cervical spinal cord: $\log_2$ fold change = -0.45; P = 1e-5). This has been previously observed due to hypermethylation of the *C9orf72* promoter in expansion carriers[58].

Gene co-expression networks associate with disease duration

We then created a weighted gene co-expression network using all 303 ALS samples, adjusting for spinal cord region, contributing site, and other technical factors. We identified 23 modules of co-expressed genes (**Fig. 2a; Supplementary Table 6**) and labelled them in ascending order of size from M1 (50

genes) to M23 (3,121). For each module we created a module eigengene (ME), equivalent to the first principal component of the expression of all genes within that module in each sample (**Supplementary Table 7**). Modules are presented clustered by eigengene correlation (**Fig. 2a**). Co-expression modules are known to identify cell-types[59], and 13 modules were significantly enriched with the top 100 cell-type marker genes for the six major cell types of the brain[46] (**Fig. 2b; Supplementary Table 8**). Similarly, we correlated each ME with cell-type proportion estimates in the ALS samples (created using the same Mathys reference and MuSiC algorithm) and found the same modules with marker gene enrichment were strongly positively correlated (Spearman's R = 0.46-0.82) with the respective cell-type proportion (**Supplementary Fig. 20).** Using the same panel of glial activation gene sets as before, we found 6 modules enriched with genes from the different sets (**Fig. 2c**). We observed that the module enriched with microglia marker genes (M17) was also enriched for disease-associated microglia and plaque-induced genes, whereas of the four astrocyte marker-enriched modules, one was enriched only with disease-associated astrocytes (M3). The two modules enriched with reactive astrocyte (RA) markers (M9, M18) were enriched with endothelial and/or endothelial cell markers, not astrocytes.

We next performed gene ontology (GO) enrichment on each module using the GO Biological Process gene sets. Overall, 22 of 23 modules had at least 1 significant GO term (**Supplementary Table 9**). We manually collapsed GO terms into broad sets (**Fig. 2d**). Some sets reflect potentially cell-type specific functions, such as myelination terms with oligodendrocytes, vasculature with endothelial cells/pericytes, and immune response with microglia, whereas modules enriched in terms relating to gene expression and translation were not enriched with cell-type specific or glial activation markers. To assess each module's relevance to ALS-specific changes, we performed enrichment tests using a consensus set of genes upregulated or downregulated in the ALS spinal cord versus controls (**Fig. 2e**). 3 modules were enriched in downregulated genes, two of which were also enriched in oligodendrocyte markers, whereas the six modules enriched with upregulated genes were also enriched with astrocyte, microglia, endothelial, and glial activation markers, confirming our previous cell-type proportion analyses.

We then used the modules to find associations with clinical variables (**Supplementary Table 10**). Correlating each ME with different clinical traits, we observed 1 module (M3) to correlate with age at death and age of onset (**Fig. 2g**), whereas 5 modules correlated with retrospective disease duration, defined as the length of time between the age at recorded disease onset and age at death. All 3 positively correlated modules were enriched with astrocyte marker genes (**Fig. 2h**), and of the two negatively correlated modules were enriched with microglia (**Fig. 2i**) and endothelial marker genes respectively. 2 modules associated with sex, an oligodendrocyte module (M16), and an astrocyte

module (M6), suggesting potential cell composition differences between males and females. Our previous study[60] used these same samples to estimate the abundance of truncated *STMN2* (tSTMN2), a novel cryptic exon transcript created by loss of nuclear TDP-43[61,62], which may be a biomarker of TDP-43 pathology[60]. 2 modules correlated with *tSTMN2* abundance. One module, M20, was positively correlated with *tSTMN2* (**Fig. 2j**). M20 is a large module containing 2048 genes and is enriched with neuronal marker genes, including the full-length *STMN2* gene, as well as motor neuron markers *MNX1* and *ISL1*, though as the sole neuronal module it is likely non-specific to motor neurons. The module negatively correlated with *tSTMN2 (*M4) is enriched with pericyte marker genes. No modules were significantly associated with the site of motor onset (limb vs bulbar).

Glial composition associates with disease duration

To further investigate the associations with disease duration, we performed a transcriptome-wide correlation analysis with disease duration as a continuous variable. 745 and 39 genes were significantly associated with disease duration at FDR < 0.05 in the cervical and lumbar spinal cord, respectively (**Supplementary Table 11**). Estimated fold-changes represent unit change in expression per month of disease. No effect size threshold was applied. Test statistics for each gene were highly concordant between the cervical and lumbar cords (Pearson R = 0.71, P < 1e-16; **Supplementary Fig 21**). Using GSEA, we found that negatively correlated genes were enriched with microglia markers and microglia activation genes, whereas positively correlated genes were enriched with astrocyte markers and pericyte markers but not astrocyte activation gene sets (**Fig. 3b-c**). Using cell-type proportion estimates from the cervical spinal cord, we observed the same negative correlation between duration and microglial proportion (R = -0.31; adjusted P = 0.002), (**Fig. 3d**), but not with astrocyte proportion (R = 0.15; adjusted P = 0.49). One of the strongest positive correlations with disease duration was found for the paraxaonase gene *PON3,* which has been previously linked to ALS through rare mutations[63]. The previously observed *CHIT1* was found to be the strongest negatively correlated gene with disease duration in both cervical and lumbar spinal cord. There is a non-linear relationship between age of onset and age at death in ALS, with shorter durations seen in both younger and older onset patients. We confirm that the association with *CHIT1* expression is strongest with disease duration, and not with age of onset or death (**Fig. 3e-f**).

Mapping spinal cord QTLs

We took common genetic variants (minor allele frequency > 1%) from the matched whole genome sequencing for all donors of European ancestry (**Supplementary Fig. 22; Supplementary Table 12**)

in the cohort, including cases of non-ALS neurodegeneration. We used this to map quantitative trait loci (QTLs) for gene expression and splicing, the latter using the intron-junction clustering method Leafcutter[64]. We identified 9,492 genes with an expression QTL (eQTL) and 5,627 with a splicing QTL (sQTL) in at least one region (**Fig. 4a; Supplementary Fig. 23**). As a comparison, we downloaded summary statistics for the only other available human spinal cord dataset, from GTEx (v8). We discovered substantially more genes with sQTLs than the 965 found by GTEx. Using Storey's π_1 we observed high sharing of QTLs between each region and with GTEx (**Fig. 4b-c**), although sharing was higher in sQTLs than eQTLs, as previously observed[65,66]. We used our previously generated cell-type proportion estimates to find cell-type interaction QTLs[67] but no tissue had sufficient power to detect any such associations.

Putative ALS risk variants colocalise with spinal cord QTLs

We then used our QTLs, in combination with GTEx, to prioritise common genetic risk loci using the latest available ALS GWAS[8] (**Fig. 4d**). Taking a relaxed approach, we extended our search from the 10 genome-wide significant loci (P < 5e-8) to 64 nominally significant subthreshold loci (P < 1e-5) (**Supplementary Table 13**). Among genome-wide significant loci, we identified strong colocalization with QTLs at a posterior probability of colocalization hypothesis 4 (PP4) > 0.8, only in *C9orf72*. In the *UNC13A* locus we observed a potentially spurious colocalization with *MVB12A* only in GTEx (PP4 = 0.5). Among the subthreshold loci, we observed colocalization in 16 loci, with the strongest colocalizing genes (PP4 > 0.8) across our tissues and GTEx seen for *ATXN3*, *GGNBP2, ACSL5,* and *FNBP1* (**Supplementary Table 14**).

We then ran transcriptome-wide association study (TWAS), an orthogonal method that uses common variants, gene expression, and splicing ratios to predict cis-regulated expression and splicing. TWAS then imputes those models to GWAS summary statistics to identify genes that are associated with disease risk. We generated TWAS models for each spinal cord section and used summary statistics from the latest available ALS GWAS[8]. In both cervical and lumbar spinal cord, splicing in *C9orf72* and *ATXN3* were significantly associated with ALS (FDR < 0.05) (**Supplementary Fig. 24; Supplementary Table 15**). The lumbar spinal cord TWAS models also identified an association with expression of *MAPT-AS1* and splicing of *LINC02210* and *LINC02210-CRHR1.* These three genes are within the contentious *MAPT* H1/H2 haplotype region, which has a complex linkage disequilibrium structure, and so are potential false positives. As a comparison, we downloaded pre-computed expression and splicing weights for the dorsolateral frontal cortex (n = 453;[68]), which found associations with *C9orf72* in both splicing and expression. In addition, the cortex TWAS models identified *SLC9A8*, *G2E3*, *SCFD1*, and *GPX3 (***Supplementary Fig. 24**).

Prioritised genes annotate to cell-types

We took each colocalised protein-coding gene (PP4 > 0.7) in any of the three spinal cord datasets and annotated them to a cell-type, and to understand how these genes might be involved in ALS. We first took cell-type fidelity ratings[69], expressed as a fidelity score from 0-100, with high scores suggesting greater cell-type specificity. Although most genes showed no preference towards any cell type, *FNBP1* (fidelity = 92) showed high specificity to oligodendrocytes (**Fig. 5a**; **Supplementary Fig. 25**). We then used the ALS co-expression network modules generated earlier to infer roles for the genes specifically in ALS. Using guilt-by-association, if a gene belongs to a module enriched in a particular cell-type or marker list, it may also be involved in that cell-type. Both *FNBP1* and *SH3RF1* were placed in module M16, highly enriched for oligodendrocytes (**Fig. 5b**). *NFASC* was placed within M6, a module enriched in both astrocyte marker genes and in disease-associated astrocytes, whereas *ACSL5* was located in M14, a module enriched in disease-associated microglia genes but not microglia markers. We then correlated each prioritised gene with estimated cell-type proportions for six cortical cell types[46]. A positive correlation with a particular cell-type proportion is suggestive evidence for specificity. *FNBP1*, *SH3RF1*, and *NFASC* all positively correlated with oligodendrocyte proportions (**Fig. 5c**). *ACSL5* positively correlated with microglia, endothelial and pericyte proportions, with the strongest correlation seen with endothelial cells. Repeating the analysis in just the control samples replicated the correlations between *FNBP1* and oligodendrocytes and *ACSL5* with endothelial cells (**Supplementary Fig. 26**).

Finally, both *FNBP1* and *SH3RF1* are downregulated in ALS cases, whereas *NFASC* expression is positively associated with disease duration in the cervical spinal cord, the only colocalised gene to do so (**Fig. 5d**). *GGNBP2* was upregulated in ALS patients but did not show a clear cell-type specificity. Despite *C9orf72* being highly expressed in mouse microglia[70], we observed no associations between *C9orf72* and any cell-type or module.

Splicing QTLs implicate repeat expansions in ALS risk

The *C9orf72* gene produces transcripts from two alternative promoters, exon 1a and exon 1b. The ALS-associated G_4C_2 hexanucleotide repeat expansion (HRE) is located between the two exons (**Fig. 6a**), with more than 30 copies of the HRE considered to be pathogenic[71]. The *C9orf72* GWAS locus colocalizes with a splicing QTL in the *C9orf72* transcript in the NYGC lumbar spinal cord, as well as an eQTL in GTEx (**Fig. 4a**). The sQTL increases the usage of the intron J1 connecting exon 1a with exon 2, which spans the HRE (**Fig. 6a**). The lead GWAS SNP rs8349943 and the lead sQTL SNP rs1537712 are in strong LD in Europeans ($R^2 = 0.75$) and we show that the GWAS SNP is also associated with J1

intron usage (**Fig. 6b; Supplementary Table 15**). The lead GWAS SNP rs8349943 is known to tag a founder haplotype which is more susceptible to the HRE[72]. Using ExpansionHunter to estimate the length of the HRE in our cohort, we replicate this finding, as carriers of rs8349943 are also enriched for the HRE (**Fig. 6c**). The usage of the J1 intron is correlated with repeat length (**Fig. 6d**). Therefore, the sQTL colocalization result is likely being driven by the effect of the tagged repeat expansion on the splicing of intron J1.

We propose a similar mechanism for the colocalization of a subthreshold GWAS locus (P = 3.2e-7) with the splicing of *ATXN3*, a promising potential ALS risk gene. The lead SNP rs10143310 was below genome-wide significance in the European ALS GWAS[8] but crossed the threshold in a multi-ethnic meta-analysis[73]. A CAG repeat in exon 10 of *ATXN3* is highly polymorphic, and expansions greater than 45 copies cause spinocerebellar ataxia type 3 (SCA3), also known as Machado-Josephs disease[74]. SCA3 patients have lower motor neuron loss and have detectable TDP-43 protein inclusions[75]. Intermediate length expansions, not sufficient to cause ataxia, have been shown to increase ALS risk in several other ataxin family genes, most notably *ATXN2*[12], but also *ATXN1*[13] and *ATXN8OS*[15]. Tagging repeat expansions in *ATXN3* with common genetic variants has been previously explored in SCA3 patients[76]. In both lumbar and cervical spinal cord samples, and in GTEx, the lead QTL SNP rs200388434 is associated with splicing with a cluster of introns at the 3' end of the *ATXN3* gene, just downstream of the site of the repeat expansion in exon 10 (**Fig. 5e**). The lead QTL SNP rs200388434 is in high linkage disequilibrium (R^2 = 0.93) with the lead GWAS SNP rs10143310 in Europeans, and rs10143310 also associated with intron splicing (**Fig. 5f**). We hypothesise that the GWAS association is tagging an intermediate length CAG repeat, and this may be the underlying causal genetic factor. We were able to genotype the CAG repeat in 304 individuals in the cohort using ExpansionHunter, observing that the lead QTL SNP associated with a narrow range of repeat lengths >= 16 (**Fig. 5g**). CAG repeat length also correlated with splicing in the lumbar spinal cord (**Fig. 5h**).

Discussion

In this study we assembled the largest ever cohort of post-mortem ALS spinal cords. This has allowed us not only to identify differentially expressed genes when compared to controls, but to identify genes associated with clinical characteristics within the ALS patient cohort. By integrating common genetic variants we prioritise several new candidate ALS genes that may have cell-type-specific functions. In this way, we investigated both the cause (genetic risk) and likely consequence (post-mortem gene expression changes) of disease.

Comparing ALS cases to controls we identified robust shifts in cell-type in the spinal cord, primarily comprised of a downregulation of oligodendrocytes and motor neurons, and an upregulation in astrocytes and microglia, as well as smaller upward shifts in endothelial cells and pericytes. We observe this across the three spinal cord regions with multiple orthogonal techniques (GSEA, deconvolution, EWCE). However, the interpretation of our results is constrained by the relatively low number of control samples in the cohort, as well as the inherent limitations in the use of bulk tissue sections. The reduction in oligodendrocyte gene expression may reflect genuine cell loss due to secondary demyelination accompanying axonal loss[77], but this may also reflect a relative shift in proportion compared to increased astrocytes and microglia. For both microglia and astrocytes, although we saw overall upregulation of multiple microglia and astrocyte activation gene lists, it is currently intractable to separate changes in cell-type proportion from changes in cell state in bulk tissue RNA-seq. We also cannot rule out that the increased microglia and activated microglia gene expression signatures may be driven by peripheral monocytes and/or T-cells, which are known to migrate into the spinal cord[78]. We also observed small increases in endothelial cells and pericytes. Alterations to the choroid plexus, including reductions in pericytes, have been observed in ALS[79]. Increases in the recently identified perivascular fibroblast cell-type have been observed in ALS spinal cord RNA-seq as well as ALS mouse models[80], although we did not explicitly look for this cell type. Crucially, we observed high concordance between our data and a published proteomic dataset from an independent ALS spinal cord and cerebrospinal fluid cohort, suggesting that the gene expression changes we identify are maintained at the protein level, increasing their utility as potential biomarkers.

Using co-expression networks built in ALS samples only, we observed a series of associations with disease duration and co-expression modules enriched in microglia and astrocyte genes, in opposing directions. Increased numbers of activated microglia, as measured by CD68 staining in the spinal cord, have been observed in faster progressing ALS patients[85]. However, it is unclear whether microglia activation accelerates neuronal death, or whether microglia activation is an attempted compensatory process, with disease duration driven by some other factor. The negative correlation observed between *CHIT1* expression and disease duration replicates previous findings at the protein level[37,39,86,87], but we also find hundreds of new associations, including in the ALS-linked gene *PON3*.

By mapping QTLs we provide a genetic resource for the ALS and wider neuroscience community to understand common genetic drivers of gene expression and splicing in the spinal cord.

Colocalization has allowed us to prioritise new ALS risk genes, but we must stress that the bulk of our findings rely on nominally significant genetic loci. We are also mindful of the potential for false positive

associations due to gene co-expression and LD contamination, which affect both colocalization and TWAS[89].

Taken together, our analyses of the ALS spinal cord point to non-neuronal cells as firmly in the heart of disease in the spinal cord in responding to, and potentially driving, progression of the disease. Our genetic analyses highlight potential new genes that may act on ALS through specific glial cell types. Future genome-wide survival studies may highlight more glial genes in also driving ALS progression. We hope our data are a useful resource for the design of future experiments.

Acknowledgements

We thank all members of the Raj lab for their feedback on the manuscript. This work was supported by grants from NIH NIA R56-AG055824 and U01-AG068880 (J.H. and T.R.), NIH NINDS U54NS123743 (J.H., T.R., P.F.), NIH NIH Medical Scientist Training Program grant T3GM007280 (J.T.H.), P.F. is supported by a UK Medical Research Council Senior Clinical Fellowship and Lady Edith Wolfson Fellowship (MR/M008606/1 and MR/S006508/1). F.K. is supported by a BOF DOCPRO fellowship of the University of Antwerp Research Fund. P.F. is supported by the UK Motor Neurone Disease Association, Rosetrees Trust, and the UCLH NIHR Biomedical Research Centre. This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai. Research reported in this paper was supported by the Office of Research Infrastructure of the National Institutes of Health under award number S10OD018522 and S10OD026880. All NYGC ALS Consortium activities are supported by the ALS Association (ALSA, 19-SI-459) and the Tow Foundation. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

JH and TR conceived and designed the project. JH led the main analysis, with SV, RH, JTH, KPL, FK, KS, MBB, GN, USE contributing code and performing additional data analyses. JH and TR oversaw all aspects of the study, with input from DAK, HP and PF. DF and HP designed the sample collection methodology, reviewed sample and data quality, and coordinated NYGC ALS Consortium postmortem RNA research activity. The NYGC ALS Consortium and the Target ALS Human Postmortem Tissue Core provided human tissue samples as well as pathological, genetic, and clinical information. JH wrote the manuscript with input from all co-authors.

Competing interest declaration

The authors have no competing interests.

Tables

Table 1 - Clinical and technical characteristics of the differential gene expression cohort

	Control	ALS	ALS-*C9orf72*	P-value
Donors	49	125	29	-
% Female	53.1%	43.2%	58.6%	0.29
% Bulbar onset	-	24.2%	25.1%	1
Disease duration, months	-	35 (6-156)	31 (6-90)	0.12
Age at death	66 (16-89)	66 (32-85)	64 (50-78)	0.63
Sequencing platform (NovaSeq / HiSeq 2500)	66.7%	59.3%	78.8%	0.104
Tissues				
Cervical Spinal Cord	35	111	28	-
RIN	6.4 (5.1-8.1)	7 (5-9)	6.5 (5.1-8.6)	**0.0077**
Lumbar Spinal Cord	32	101	21	-
RIN	5.8 (5-7.8)	6.9 (5.1-8.7)	6.3 (5.1-8)	**3e-05**
Thoracic Spinal Cord	10	37	5	-
RIN	6.35 (5.6-8.1)	6.5 (5-8)	7.5 (5.5-8)	0.79

ALS-C9orf72: ALS with confirmed *C9orf72* hexanucleotide expansion. RIN: RNA integrity number. Continuous variables presented as median and range. Categorical variables compared with Fisher's exact test, continuous variables with Kruskal-Wallis or Wilcoxon rank sum tests. P-values shown are uncorrected for multiple testing, P-values < 0.05 are bolded

Table 2 - Differentially expressed genes (DEGs) found in each spinal cord region

Region	Control	ALS	Genes tested	All DEGs (FDR < 0.05)	DEGs |LFC] > 1	DEGs |LFC| > 2
Cervical	35	139	25,389	7,349	377	29
Thoracic	10	42	19,367	256	65	9
Lumbar	32	122	25,601	4,694	233	7

Figure Legends

Fig. 1 | **Differential gene expression in the ALS spinal cord is driven by cell-type composition. a-b.** Volcano plots comparing ALS patients to controls in each spinal cord section. P-values for each gene generated from empirical Bayes moderated t-statistics (limma-voom), followed by Benjamini-Hochberg multiple testing adjustment. Genes coloured by whether not differentially expressed (FDR < 0.05; grey), differentially expressed but with modest effects ($|\log_2$ fold change (LFC)| < 1; orange) and with stronger effects (|LFC| > 1; red). Numbers of genes in each category above the plot. **c-d.** Comparing LFC effect sizes between the two regions for the most upregulated (**c**) or downregulated (**d**) genes. **e.** The 20 most upregulated (left) and downregulated (right) genes, ordered by LFC. Asterisks represent Benjamini-Hochberg adjusted P < 0.05 across the 25,389 and 25,601 respective genes from differential expression. (**f-h**) Gene Set Enrichment Analysis results.Normalised enrichment score (NES) is a measure of enrichment of a gene set within a ranked list of genes compared to a permuted background. All pathways are enriched in upregulated genes. Significance derived from empirical P-values from a one-sided permutation test followed by Benjamini-Hochberg correction. (f) GSEA results for the 50 molecular signature hallmark pathways genes sets. 100 tests performed. (g GSEA results for the cell-type signature gene sets. 12 tests performed. **h.** GSEA results for the glial activation gene sets. 10 tests performed. DAA: disease-associated astrocytes; DAM: disease-associated microglia; PIG: plaque-induced genes; RA-LPS: reactive astrocytes in response to lipopolysaccharide; RA-MCAO: reactive astrocytes in reponse to hypoxia. **i.** Estimated cell-type proportions in the cervical spinal cord, between 139 ALS patients and 35 controls. n=174 biologically independent samples. P-values from a two-sided Wilcoxon non-parametric test comparing residuals after regressing technical covariates, followed by Bonferroni correction. 6 tests performed. Boxplots show the median, first and third quartile of the distribution with whiskers extending to 1.5 times the interquartile range. **j-k.** Correlating differentially expressed genes (FDR < 0.05) in the Cervical spinal cord with differentially expressed proteins (FDR < 0.05) from post-mortem spinal cord (j) and cerebrospinal fluid (k). Asterisks reflect magnitude of adjusted P-values: *** q < 1e-4; ** q < 1e-3; * q < 0.05; . q > 0.05.

Fig. 2 | **Gene co-expression network in the ALS spinal cord. (a-g)** Weighted gene co-expression network analysis of 303 ALS spinal cord samples identifies 23 gene modules. **a.** Modules are presented as hierarchical clustering based on module eigengene (ME) correlation. **b-e:** Enrichment results between each module and **b**) cell-type marker genes from Mathys et al, **c**) glial activation genes, **d**) gene ontology (biological process) enrichment, manually collapsed, **e**) differentially expressed genes (FDR < 0.05, no fold change cutoff) between ALS and controls, across all spinal cord regions, **f)** Spearman correlation with disease traits. g-j). MEs for each ALS patient. M3 correlates with age of symptom onset, M8 and M17 with duration of disease in months, and M20 with *tSTMN2* expression. R refers to Spearman correlation. * refers to Bonferroni adjusted P < 0.05, adjusted for the number of cells in each panel separately. Tests performed: 138 (b), 115 (c), 3,326 (d), 46 (e), 161 (f). *tSTMN2* - truncated STMN2. TPM - transcripts per million. P-values for b,c,e from one-sided Fisher's exact test followed by Bonferroni adjustment, d from one-sided hypergeometric test followed by g:SCS adjustment, f from two-sided Spearman correlation test followed by Bonferroni adjustment.

Fig. 3 | **Gene expression correlations with duration of disease. a.** Volcano plots for correlation in each tissue. Log_2 fold-changes represent unit change in expression per month of disease duration. P-values for each gene generated from empirical Bayes moderated t-statistics (limma-voom), followed by Benjamini-Hochberg multiple testing adjustment **b.** GSEA with cell-type marker genes. 8 tests performed. **c.** GSEA with glia activation gene lists. 10 tests performed. GSEA P-values generated from a one-sided permutation test followed by Benjamini-Hochberg correction **d.** Cell-type proportions in the cervical spinal cord estimated with deconvolution plotted against disease duration. **e**. *CHIT1* is strongly upregulated in ALS in all three tissues. Sample numbers in Table 1. **f.** *CHIT1* expression negatively correlates with disease duration, but not with age of onset, and only weakly with age at death. All correlations are Spearman rank correlations. Two-sided P-values in panels b and c are Bonferroni corrected for 12 and 10 tests respectively. P-values in d. are Bonferroni-corrected for 6 tests. Asterisks reflect magnitude of adjusted P-values: *** q < 1e-4; ** q < 1e-3; * q < 0.05; . q > 0.05. Boxplots show the median, first and third quartile of the distribution with whiskers extending to 1.5 times the interquartile range.

Fig. 4 | **Quantitative trait loci (QTL) colocalize with putative ALS risk variants. a.** QTL discovery in the three spinal cord tissues and compared with GTEx (v8). Numbers refer to genes with an expression QTL (eGenes) or a splicing QTL (sGenes) at qvalue < 0.05. **b-c.** Sharing of QTLs between tissues using Storey's π_1 metric. Values are not symmetric. **d.** Colocalization of subthreshold ALS GWAS loci with spinal cord QTLs. Loci are named for their nearest protein-coding gene. P-values refer to the association of the lead variant in the locus with ALS risk from the GWAS (logistic regression). Numbers refer to the probability of a single shared variant in both GWAS and QTL (PP4). All genes and loci shown with PP4 > 0.5 in at least one QTL dataset. Genes taken for further analysis are in bold font. Circles refer to eQTLs, triangles to sQTLs. PP4: posterior probability of colocalization hypothesis 4.

Fig. 5 | **Annotating colocalised genes with cell-type information. a-d.** Each protein-coding gene with PP4 > 0.7 in at least one spinal cord QTL dataset. **a**. Cell-type fidelity scores from Kelley et al., higher scores imply higher cell-type specificity. **b.** The cell-type and activation marker enrichment p-values (one-sided Fisher's exact test) from Fig. 4 for the modules containing each gene. **c.** Each gene correlated with estimated cell-type proportions in cervical spinal cord in the ALS samples only. Two-sided Pearson correlation test. **d.** Log_2 fold changes from differential expression in ALS vs Control (upper panel) and ALS disease duration (lower panel) in cervical and lumbar spinal cord. P-values in b. and c. Bonferroni adjusted for 138 (b upper panel), 115 (b lower panel), and 72 (c) tests. P-values in d. from limma-voom adjusted by Benjamini-Hochberg method for all genes tested in each cohort. Asterisk denotes adjusted p-value < 0.05.

Fig. 6 | **Splicing QTLs illuminate genetic associations with repeat expansions in *C9orf72* and *ATXN3*. a.** The ALS-causing GGGGCC repeat expansion lies in between the two first exons, 1a and 1b. The intron connecting the exon 1a with exon 2 (J1) has an sQTL in the lumbar spinal cord that colocalises with ALS risk (PP4 = 0.78). **b**. The lead GWAS SNP rs8349943 is associated with J1 intron splicing in the lumbar spinal cord (P = 1.3e-9, linear regression, n=197 independent samples). **c**. The GGGGCC expansion is only observed in carriers of rs8349943. 30 copies of the repeat is considered the threshold for disease initiation. (P=1.6e-4, linear regression, n=139 independent samples) [71]. **d**. The GGGGCC repeat expansion is associated with J1 intron splicing (R=0.33, P = 8.8e-5, two-sided Pearson correlation test). **e.** The *ATXN3* gene produces multiple transcripts, including several short transcripts at the 3' end of the gene. Three introns have sQTLs that colocalise with a subthreshold ALS risk GWAS locus with high PP4. The introns are all immediately downstream of a CAG repeat within exon 10. **f.** The lead GWAS SNP rs10143310 is associated with usage of the J1 intron, (P=3.8e-13, linear regression, n=196 independent samples). **g.** Carriers of rs10143310 have a CAG repeat length > 16 copies. (P=1.1e-7, linear regression, n=130 independent samples). **h.** The length of the CAG repeat correlates with J1 splicing (p = 0.05, two-sided Pearson correlation test). Boxplots show the median, first and third quartile of the distribution with whiskers extending to 1.5 times the interquartile range.

References

1. Ravits, J. M. & La Spada, A. R. ALS motor phenotype heterogeneity, focality, and spread: deconstructing motor neuron degeneration. *Neurology* **73**, 805–811 (2009).

2. Neumann, M. *et al.* Ubiquitinated TDP-43 in Frontotemporal Lobar Degeneration and Amyotrophic Lateral Sclerosis. *Science* **314**, 130–133 (2006).
3. Byrne, S. *et al.* Rate of familial amyotrophic lateral sclerosis: a systematic review and meta-analysis. *J. Neurol. Neurosurg. Psychiatry* **82**, 623–627 (2011).
4. Majounie, E. *et al.* Frequency of the C9orf72 hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: A cross-sectional study. *Lancet Neurol.* **11**, 323–330 (2012).
5. Renton, A. E., Chiò, A. & Traynor, B. J. State of play in amyotrophic lateral sclerosis genetics. *Nat. Neurosci.* **17**, 17–23 (2014).
6. Cirulli, E. T. *et al.* Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Nat. Methods* **347**, 1436–1441 (2016).
7. Kenna, K. P. *et al.* NEK1 variants confer susceptibility to amyotrophic lateral sclerosis. *Nat. Genet.* (2016) doi:10.1038/ng.3626.
8. Nicolas, A. *et al.* Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. *Neuron* **97**, 1268–1283.e6 (2018).
9. van Es, M. A. *et al.* Genome-wide association study identifies 19p13.3 (UNC13A) and 9p21.2 as susceptibility loci for sporadic amyotrophic lateral sclerosis. *Nat. Genet.* **41**, 1083–1087 (2009).
10. Van Rheenen, W. *et al.* Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat. Genet.* **48**, 1043–1048 (2016).
11. van Rheenen, W. *et al.* Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *Nat. Genet.* **53**, 1636–1648 (2021).
12. Elden, A. C. *et al.* Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. *Nature* **466**, 1069–1075 (2010).
13. Tazelaar, G. H. P. *et al.* repeat expansions confer risk for amyotrophic lateral sclerosis and contribute to TDP-43 mislocalization. *Brain Commun* **2**, fcaa064 (2020).
14. Lattante, S. *et al.* ATXN1 intermediate-length polyglutamine expansions are associated with amyotrophic lateral sclerosis. *Neurobiol. Aging* **64**, 157.e1–157.e5 (2018).
15. Hirano, M. *et al.* Noncoding repeat expansions for ALS in Japan are associated with the ATXN8OS gene. *Neurology Genetics* vol. 4 e252 Preprint at https://doi.org/10.1212/nxg.0000000000000252 (2018).
16. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
17. Young, A. M. H. *et al.* A map of transcriptional heterogeneity and regulatory variation in human microglia. *Nat. Genet.* **53**, 861–868 (2021).
18. Novikova, G. *et al.* Integration of Alzheimer's disease genetics and myeloid genomics identifies disease risk regulatory elements and genes. *Nat. Commun.* **12**, 1610 (2021).
19. Lopes, K. de P. *et al.* Genetic analysis of the human microglial transcriptome across brain regions, aging and disease pathologies. *Nat. Genet.* **54**, 4–17 (2022).
20. Pramatarova, A., Laganière, J., Roussel, J., Brisebois, K. & Rouleau, G. A. Neuron-specific expression of mutant superoxide dismutase 1 in transgenic mice does not lead to motor impairment. *J. Neurosci.* **21**, 3369–3374 (2001).

21. Jaarsma, D., Teuling, E., Haasdijk, E. D., De Zeeuw, C. I. & Hoogenraad, C. C. Neuron-specific expression of mutant superoxide dismutase is sufficient to induce amyotrophic lateral sclerosis in transgenic mice. *J. Neurosci.* **28**, 2075–2088 (2008).
22. Yamanaka, K. *et al.* Astrocytes as determinants of disease progression in inherited amyotrophic lateral sclerosis. *Nat. Neurosci.* **11**, 251–253 (2008).
23. Lepore, A. C. *et al.* Focal transplantation–based astrocyte replacement is neuroprotective in a model of motor neuron disease. *Nat. Neurosci.* **11**, 1294–1301 (2008).
24. Boillée, S. *et al.* Onset and progression in inherited ALS determined by motor neurons and microglia. *Science* **312**, 1389–1392 (2006).
25. Wang, L., Sharma, K., Grisotti, G. & Roos, R. P. The effect of mutant SOD1 dismutase activity on non-cell autonomous degeneration in familial amyotrophic lateral sclerosis. *Neurobiol. Dis.* **35**, 234–240 (2009).
26. Phatnani, H. P. *et al.* Intricate interplay between astrocytes and motor neurons in ALS. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E756–65 (2013).
27. Town, T., Nikolic, V. & Tan, J. The microglial 'activation' continuum: from innate to adaptive responses. *J. Neuroinflammation* **2**, 24 (2005).
28. Chiu, I. M. *et al.* A neurodegeneration-specific gene-expression signature of acutely isolated microglia from an amyotrophic lateral sclerosis mouse model. *Cell Rep.* **4**, 385–401 (2013).
29. Liddelow, S. A. *et al.* Neurotoxic reactive astrocytes are induced by activated microglia. *Nature* **541**, 481–487 (2017).
30. Zhao, W. *et al.* Activated microglia initiate motor neuron injury by a nitric oxide and glutamate-mediated mechanism. *J. Neuropathol. Exp. Neurol.* **63**, 964–977 (2004).
31. Haidet-Phillips, A. M. *et al.* Astrocytes from familial and sporadic ALS patients are toxic to motor neurons. *Nat. Biotechnol.* **29**, 824–828 (2011).
32. Guttenplan, K. A. *et al.* Knockout of reactive astrocyte activating factors slows disease progression in an ALS mouse model. *Nat. Commun.* **11**, 3753 (2020).
33. D'Erchia, A. M. *et al.* Massive transcriptome sequencing of human spinal cord tissues provides new insights into motor neuron degeneration in ALS. *Scientific Reports* vol. 7 Preprint at https://doi.org/10.1038/s41598-017-10488-7 (2017).
34. Brohawn, D. G., O'Brien, L. C. & Bennett, J. P., Jr. RNAseq Analyses Identify Tumor Necrosis Factor-Mediated Inflammation as a Major Abnormality in ALS Spinal Cord. *PLoS One* **11**, e0160520 (2016).
35. Andrés-Benito, P., Moreno, J., Aso, E., Povedano, M. & Ferrer, I. Amyotrophic lateral sclerosis, gene deregulation in the anterior horn of the spinal cord and frontal cortex area 8: implications in frontotemporal lobar degeneration. *Aging* **9**, 823–851 (2017).
36. Dols-Icardo, O. *et al.* Motor cortex transcriptome reveals microglial key events in amyotrophic lateral sclerosis. *Neurol Neuroimmunol Neuroinflamm* **7**, (2020).
37. Thompson, A. G. *et al.* Cerebrospinal fluid macrophage biomarkers in amyotrophic lateral sclerosis. *Ann. Neurol.* **83**, 258–268 (2018).
38. Tanaka, H. *et al.* The potential of GPNMB as novel neuroprotective factor in amyotrophic lateral sclerosis. *Sci. Rep.* **2**, 573 (2012).
39. Oeckl, P. *et al.* 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 (2020).
40. Hüttenrauch, M. *et al.* Glycoprotein NMB: a novel Alzheimer's disease associated marker expressed in a subset of activated microglia.

Acta Neuropathol Commun **6**, 108 (2018).

41. Murthy, M. N. *et al.* Increased brain expression of GPNMB is associated with genome wide significant risk for Parkinson's disease on chromosome 7p15.3. *Neurogenetics* **18**, 121–133 (2017).
42. Nalls, M. A. *et al.* Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **18**, 1091–1102 (2019).
43. Halter, B. *et al.* Oxidative stress in skeletal muscle stimulates early expression of Rad in a mouse model of amyotrophic lateral sclerosis. *Free Radic. Biol. Med.* **48**, 915–923 (2010).
44. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
45. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425 (2015).
46. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).
47. Keren-Shaul, H. *et al.* A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease. *Cell* **169**, 1276–1290.e17 (2017).
48. Habib, N. *et al.* Disease-associated astrocytes in Alzheimer's disease and aging. *Nat. Neurosci.* **23**, 701–706 (2020).
49. Zamanian, J. L. *et al.* Genomic analysis of reactive astrogliosis. *J. Neurosci.* **32**, 6391–6410 (2012).
50. Chen, W.-T. *et al.* Spatial Transcriptomics and In Situ Sequencing to Study Alzheimer's Disease. *Cell* **182**, 976–991.e19 (2020).
51. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences* vol. 112 7285–7290 Preprint at https://doi.org/10.1073/pnas.1507125112 (2015).
52. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**, 380 (2019).
53. Hunt, G. J., Freytag, S., Bahlo, M. & Gagnon-Bartsch, J. A. dtangle: accurate and robust cell type deconvolution. *Bioinformatics* **35**, 2093–2099 (2019).
54. Skene, N. G. & Grant, S. G. N. Identification of Vulnerable Cell Types in Major Brain Disorders Using Single Cell Transcriptomes and Expression Weighted Cell Type Enrichment. *Front. Neurosci.* **10**, 16 (2016).
55. Prudencio, M. *et al.* Distinct brain transcriptome profiles in C9orf72-associated and sporadic ALS. *Nat. Neurosci.* **18**, 1175–1182 (2015).
56. Dickson, D. W. *et al.* Extensive transcriptomic study emphasizes importance of vesicular transport in C9orf72 expansion carriers. *Acta Neuropathol Commun* **7**, 150 (2019).
57. Dolzhenko, E. *et al.* ExpansionHunter: A sequence-graph based tool to analyze variation in short tandem repeat regions. Preprint at https://doi.org/10.1101/572545.
58. Jackson, J. L. *et al.* Elevated methylation levels, reduced expression levels, and frequent contractions in a clinical cohort of C9orf72 expansion carriers. *Molecular Neurodegeneration* vol. 15 Preprint at https://doi.org/10.1186/s13024-020-0359-8 (2020).
59. Oldham, M. C. *et al.* Functional organization of the transcriptome in human brain. *Nat. Neurosci.* **11**, 1271–1282 (2008).
60. Prudencio, M. *et al.* Truncated stathmin-2 is a marker of TDP-43 pathology in frontotemporal dementia. *J. Clin. Invest.* (2020) doi:10.1172/JCI139741.
61. Klim, J. R. *et al.* ALS-implicated protein TDP-43 sustains levels of STMN2, a mediator of motor neuron growth and repair. *Nat. Neurosci.*

22, 167–179 (2019).

62. Melamed, Z. *et al.* Premature polyadenylation-mediated loss of stathmin-2 is a hallmark of TDP-43-dependent neurodegeneration. *Nat. Neurosci.* **22**, 180–190 (2019).
63. Ticozzi, N. *et al.* Paraoxonase gene mutations in amyotrophic lateral sclerosis. *Ann. Neurol.* **68**, 102–107 (2010).
64. Li, Y. I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).
65. Lopes, K. de P. *et al.* Atlas of genetic effects in human microglia transcriptome across brain regions, aging and disease pathologies. Preprint at https://doi.org/10.1101/2020.10.27.356113.
66. Consortium, T. G. & The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* vol. 369 1318–1330 Preprint at https://doi.org/10.1126/science.aaz1776 (2020).
67. Kim-Hellmuth, S. *et al.* Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**, (2020).
68. Li, Y. I., Wong, G., Humphrey, J. & Raj, T. Prioritizing Parkinson's disease genes using population-scale transcriptomic data. *Nat. Commun.* **10**, 994 (2019).
69. Kelley, K. W., Nakao-Inoue, H., Molofsky, A. V. & Oldham, M. C. Variation among intact tissue samples reveals the core transcriptional features of human CNS cell classes. *Nat. Neurosci.* **21**, 1171–1184 (2018).
70. O'Rourke, J. G. *et al.* C9orf72 is required for proper macrophage and microglial function in mice. *Science* **351**, 1324–1329 (2016).
71. Renton, A. E. *et al.* A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72**, 257–268 (2011).
72. DeJesus-Hernandez, M. *et al.* Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* **72**, 245–256 (2011).
73. Nakamura, R. *et al.* A multi-ethnic meta-analysis identifies novel genes, including ACSL5, associated with amyotrophic lateral sclerosis. *Commun Biol* **3**, 526 (2020).
74. Paulson, H. Machado-Joseph Disease/Spinocerebellar Ataxia Type 3. *Genetic Instabilities and Neurological Diseases* 363–377 Preprint at https://doi.org/10.1016/b978-012369462-1/50025-9 (2006).
75. Seidel, K. *et al.* Axonal inclusions in spinocerebellar ataxia type 3. *Acta Neuropathol.* **120**, 449–460 (2010).
76. Prudencio, M. *et al.* Toward allele-specific targeting therapy and pharmacodynamic marker for spinocerebellar ataxia type 3. *Sci. Transl. Med.* **12**, (2020).
77. Kang, S. H. *et al.* Degeneration and impaired regeneration of gray matter oligodendrocytes in amyotrophic lateral sclerosis. *Nat. Neurosci.* **16**, 571–579 (2013).
78. Zondler, L. *et al.* Peripheral monocytes are functionally altered and invade the CNS in ALS patients. *Acta Neuropathol.* **132**, 391–411 (2016).
79. Saul, J. *et al.* Global alterations to the choroid plexus blood-CSF barrier in amyotrophic lateral sclerosis. *Acta Neuropathol Commun* **8**, 92 (2020).
80. Månberg, A. *et al.* Publisher Correction: Altered perivascular fibroblast activity precedes ALS disease onset. *Nat. Med.* **27**, 1308 (2021).
81. Lake, B. B. *et al.* Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586–1590 (2016).

82. Maniatis, S. *et al.* Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science* **364**, 89–93 (2019).
83. Blum, J. A. *et al.* Single-cell transcriptomic analysis of the adult mouse spinal cord reveals molecular diversity of autonomic and skeletal motor neurons. *Nat. Neurosci.* **24**, 572–583 (2021).
84. Ho, R. *et al.* Cross-Comparison of Human iPSC Motor Neuron Models of Familial and Sporadic ALS Reveals Early and Convergent Transcriptomic Disease Signatures. *Cell Syst* **12**, 159–175.e9 (2021).
85. Brettschneider, J. *et al.* Microglial activation correlates with disease progression and upper motor neuron clinical symptoms in amyotrophic lateral sclerosis. *PLoS One* **7**, e39216 (2012).
86. Varghese, A. M. *et al.* Chitotriosidase, a biomarker of amyotrophic lateral sclerosis, accentuates neurodegeneration in spinal motor neurons through neuroinflammation. *J. Neuroinflammation* **17**, 232 (2020).
87. Pagliardini, V. *et al.* Chitotriosidase and lysosomal enzymes as potential biomarkers of disease progression in amyotrophic lateral sclerosis: a survey clinic-based study. *J. Neurol. Sci.* **348**, 245–250 (2015).
88. Zeng, B. *et al.* Trans-ethnic eQTL meta-analysis of human brain reveals regulatory architecture and candidate causal variants for brain-related traits. *medRxiv* (2021).
89. Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).
90. van Rheenen, W. *et al.* Author Correction: Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *Nat. Genet.* **54**, 361 (2022).
91. Aspenström, P. Formin-binding proteins: modulators of formin-dependent actin polymerization. *Biochim. Biophys. Acta* **1803**, 174–182 (2010).
92. Wu, C.-H. *et al.* Mutations in the profilin 1 gene cause familial amyotrophic lateral sclerosis. *Nature* **488**, 499–503 (2012).
93. Nelson, A. D. & Jenkins, P. M. Axonal Membranes and Their Domains: Assembly and Function of the Axon Initial Segment and Node of Ranvier. *Front. Cell. Neurosci.* **11**, 136 (2017).
94. Efthymiou, S. *et al.* Biallelic mutations in neurofascin cause neurodevelopmental impairment and peripheral demyelination. *Brain* **142**, 2948–2964 (2019).
95. West, R. J. H., Ugbode, C., Gao, F.-B. & Sweeney, S. T. The pro-apoptotic JNK scaffold POSH/SH3RF1 mediates CHMP2BIntron5-associated toxicity in animal models of frontotemporal dementia. *Hum. Mol. Genet.* **27**, 1382–1395 (2018).
96. Saez-Atienzar, S. *et al.* Genetic analysis of amyotrophic lateral sclerosis identifies contributing pathways and cell types. *Cold Spring Harbor Laboratory* 2020.07.20.211276 (2020) doi:10.1101/2020.07.20.211276.
97. Klemens, J. *et al.* Neurotoxic potential of reactive astrocytes in canine distemper demyelinating leukoencephalitis. *Sci. Rep.* **9**, 11689 (2019).

Methods

NYGC ALS Consortium cohort

The 1,917 RNA-seq samples from the January 2020 freeze of the New York Genome Center (NYGC) ALS Consortium were downloaded, comprising of samples from cortical regions, cerebellum and spinal cord. This study used only the spinal cord samples. Diagnosis was determined by each contributing site. Donors include non-neurological disease controls (hereafter controls), those with classical ALS (hereafter ALS), frontotemporal dementia (FTD), mixed pathologies (ALS-FTD, ALS-Alzheimer's), and a small number of other diseases including Primary Lateral Sclerosis, Kennedy's Disease and Parkinson's Disease. *C9orf72* and *ATXN3* repeat expansion lengths were estimated by the Consortium using ExpansionHunter (v2.5.5)[1] on samples that had PCR-free whole genome sequencing available. Patients with greater than 30 repeats were defined as *C9orf72*-ALS. For ALS patients, age of symptom onset and age at death was reported by each contributing site. Disease duration was defined as the difference between age at death and symptom onset, in months. The NYGC ALS Consortium samples presented in this work were acquired through various institutional review board (IRB) protocols from member sites and the Target ALS postmortem tissue core and transferred to the NYGC in accordance with all applicable foreign, domestic, federal, state, and local laws and regulations for processing, sequencing, and analysis. The Biomedical Research Alliance of New York (BRANY) IRB serves as the central ethics oversight body for NYGC ALS Consortium. Ethical approval was given and is effective through 08/22/2022.

RNA-seq processing and quality control

The Consortium's RNA-seq sample processing has been, in part, previously described[2,3]. In brief, RNA was extracted from flash-frozen postmortem tissue using TRIzol (Thermo Fisher Scientific) chloroform, followed by column purification (RNeasy Minikit, QIAGEN). RNA integrity number (RIN)[4] was assessed on a Bioanalyzer (Agilent Technologies). RNA-Seq libraries were prepared from 500 ng total RNA using the KAPA Stranded RNA-Seq Kit with RiboErase (KAPA Biosystems) for rRNA depletion and Illumina-compatible indexes (NEXTflex RNA-Seq Barcodes, NOVA-512915, PerkinElmer, and IDT for Illumina TruSeq UD Indexes, 20022370). Pooled libraries (average insert size: 375 bp) passing the quality criteria were sequenced either on an Illumina HiSeq 2500 (125 bp paired end) or an Illumina NovaSeq (100 bp paired-end). Samples were subjected to extensive sequencing and RNA-Seq quality control metrics at the NYGC that are described below. Notably, a set of more than 250 markers was used to confirm tissue, neuroanatomical regions, and sex in the RNA-Seq data. Only samples passing these

metrics are available for distribution. The samples had a median sequencing depth of 42 million read pairs, with a range between 16 and 167 million read pairs.

Samples were uniformly processed using RAPiD-nf, an efficient RNA-Seq processing pipeline implemented in the NextFlow framework[5]. Following adapter trimming with Trimmomatic (version 0.36) [6], all samples were aligned to the hg38 build (GRCh38.primary_assembly) of the human reference genome using STAR (2.7.2a)[7], with indexes created from GENCODE, version 30 [8]. Gene expression was quantified using RSEM (1.3.1)[9]. Quality control was performed using SAMtools (v1.9) [10] and Picard (v2.22.3), and the results were collated using MultiQC (v1.8)[11].

Aligned RNA-seq samples were subjected to quality control modelled on the criteria of the Genotype Tissue Expression Consortium[12]. Any sample failing 1 of the following sequencing metric thresholds was removed: a unique alignment rate of less than 90%, ribosomal bases of greater than 10%, a mismatch rate of greater than 1%, a duplication rate of greater than 0.5%, intergenic bases of less than 10.5%, and ribosomal bases of greater than 0.1%. For tissue identity, both principal components analysis and UMAP were performed on the TMM-normalised gene expression matrix, followed by k-means clustering. This identified three clusters of samples, grouped by cerebellum, cortical regions, and spinal cord. Samples that clustered with a non-matching tissue type were flagged and tissue identity was re-confirmed using the expression of the cerebellar marker *CBNL1*, the cortical marker *NRGN* and the oligodendrocyte marker *MOBP*. 19 samples were removed for having ambiguous tissue identity. For duplicate samples, where samples of the same tissue from the same donor were sequenced, the sample with the highest RIN was retained, this removed 15 duplicate samples. Sex was confirmed using *XIST* and *UTY* expression. 11 samples with missing sex information were confirmed as males. Due to the large impact of RNA integrity number (RIN) on expression, only samples with RIN >= 5 were included in the differential expression analysis, totalling 380 spinal cord samples from 203 donors. For the QTL analyses (see below), no RIN threshold was applied.

Covariate selection and modelling for differential expression

The following was run for each tissue separately: Clinical variables (disease status, age at death, sex, contributing site) were combined with sequencing variables (RIN, sequencing preparation method, sequencing platform), technical metrics of the RNA-seq libraries from Picard (% mRNA bases, 3' bias, etc), and genotype principal components (see below). Using voom-normalised gene expression removing lowly expressed genes, principal components analysis was performed. The top 10 principal components were then associated with each potential confounding variable using a linear model,

estimating the variance explained (r^2) of the confounder on each principal component (**Supplementary Fig. 3a**). Using an orthogonal approach, variancePartition (v1.21.6)[13] was run on a reduced set of confounding variables, taking only the nominally independent sequencing metrics (**Supplementary Fig. 3b)**.

For performing differential gene expression between ALS and control samples, multiple model designs were fitted to account for differences in sequencing batch and contributing site, both of which are correlated with disease status. To account for potentially non-linear dependence of RIN and age at death, squared terms were included. To account for potential confounding differences due to genetic background, the first 5 genotype principal components (gPCs) from smartpca (v6.0.1)[14] were included. For filtering lowly expressed genes, a permission threshold of median TPM > 0 was applied, resulting in 24-25,000 genes being kept for each analysis. For Cervical and Lumbar spinal cord, the following model was fitted: *expression ~ disease + sex + library preparation method + contributing site + age + age^2 + RIN + RIN^2 + % mRNA bases + gPC1 + gPC2 + gPC3 + gPC4 + gPC5.* For the smaller set of Thoracic spinal cord samples, a reduced model was fitted as it maximised the gene-gene correlation of differential expression effect sizes with the other two regions: *expression ~ disease + sex + RIN + RIN^2 + age + age^2 + library preparation method + gPC1 + gPC2 + gPC3 + gPC4 + gPC5.* Differential gene expression was fitted using limma voom (v3.46.0)[15] on TMM-normalized[16] read counts. P-values were adjusted for multiple testing using FDR correction, with genes were considered differentially expressed at FDR < 0.05. A gene was considered to have a moderate effect size at $|\log_2$ fold change$| > 1$.
For transcriptome-wide correlations with disease duration, the same models as before were used in the ALS samples only, with disease duration (years) used as continuous variables. Downsampling was performed by taking random subsets of either the Cervical or Lumbar samples, without replacement.

Gene set enrichment analysis

Sets of genes were collected from multiple sources and compared to the full differential expression results for each tissue using Gene Set Enrichment Analysis (GSEA)[17], as implemented in the Clusterprofiler R package (v3.18.1)[18]. As input we included all tested genes from the differential expression or disease duration analysis for each tissue at nominal (unadjusted) P-value < 0.05, ranked by $\log_2$ fold change. For each gene set, a running cumulative tally is made of whether genes in a set are present or absent during a walk down the list. The maximal score during the walk is the enrichment score (ES), which reflects the degree of which a gene set is enriched at either the top or bottom of a list. Labels are then randomly permuted to generate an empirical null ES distribution and a P-value is calculated. To aid comparison between sets, each ES is then divided by the mean null ES to create a

normalised enrichment score (NES). Hallmark pathway gene sets (h.all.v7.2.symbols.gmt) were downloaded from the molecular signatures database[19]. Cell-type marker genes were created using single cell RNA-seq[20] and single nucleus RNA-seq[21] from human cortex. For each dataset, the top 100 cell-type specific genes were calculated by comparing gene expression of each cell-type group against the mean of all cells in Limma Voom. Marker genes for astrocytes, microglia, neurons, oligodendrocytes, and pericytes were downloaded from the Kelley et al[22], PanglaoDB[23] and Neuroexpresso[24] websites (**see URLs**). Disease-associated Microglia (DAM) signature genes[25], Disease-associated astrocytes[26], Plaque-associated genes[27], and LPS and MCAO-activated astrocyte genes[28] were downloaded from their respective supplementary materials. Mouse genes were lifted over to their human homologues using Homologene[29]. Any duplicate gene name, or gene name without a matching Ensembl ID in GENCODE v30 was removed.

Re-analysis of proteomics data

Summary statistics from a published study[30] applying isobaric tags for relative and absolute quantification (iTRAQ) proteomics for cerebrospinal fluid (26 ALS, 16 Control) and label-free proteomics to human spinal cord (8 ALS, 7 control) were downloaded from the study's supplementary data files. A total of 1,929 proteins were tested in the cerebrospinal fluid, of which 32 were called significant at FDR < 0.05. 5,115 peptides were tested in spinal cord samples, of which 292 were called significant at FDR < 0.05. Peptides assigned to multiple genes were discarded, resulting in 287 genes in the spinal cord and 30 in CSF.

Cell-type deconvolution

Filtered counts and cell-type labels for single nucleus RNA-seq from 80,660 cells from 48 human dorsolateral prefrontal cortex samples[21] were downloaded from Synapse (syn18681734). Only cells from the 14 donors without dementia were kept. Single cell RNA-seq data of 466 cells from 12 donors[20] was downloaded from Gene Expression Omnibus (GSE67835) using the count matrices and cell-type labels provided. Bulk spinal cord RNA-seq data was voom-normalized before deconvolution was estimated using MuSiC (v0.1.1)[31], a method which incorporates the variance between multiple donors from single cell/nucleus RNA-seq. In addition, we ran dtangle (v2.0.9)[32], which requires marker genes to be generated for each cell-type. Markers were created using Voom to compare each gene in purified cell-type to the mean of all cell-types. The top 100 genes ranked by effect size were used as cell-type markers. Estimated proportions of each cell-type were compared between ALS and control using non-parametric Wilcoxon tests after regressing the same technical covariates above. P-values were

corrected for multiple testing using Bonferroni correction. For comparing duration of onset, estimated cell-type proportions were correlated using a Spearman correlation.

Expression-weighted Cell-type Enrichment

Expression-weighted cell-type enrichment analysis was performed using the EWCE package[33]. Cell-type specificity scores for each gene were created using human frontal cortex single-nucleus RNA-seq[21]. Cell-type enrichment results were generated using the top 250 upregulated and downregulated genes, ordered by t-statistic, for the differential expression results for each segment. Specificity scores for each set were then compared to the mean of the empirical null distribution from 10,000 random gene sets. Enrichment was expressed as the number of standard deviations from the mean. P-values were Bonferroni corrected for multiple testing. Significance was set at adjusted $P < 0.05$.

Gene co-expression Networks

Gene expression from all 303 ALS samples from the three spinal cord regions was combined into a single matrix. Genes annotated as protein-coding by Ensembl were kept, and only then if each gene had at least 1 read count per million in at least 50% of samples, resulting in 16,992 genes. Gene counts were then transformed using Voom and TMM normalization. The following covariates were then regressed out using removeBatchEffect(): library preparation, contributing site, spinal cord section, RIN, % mRNA bases, and genomic PCs 1-5. Co-expression network analysis was performed using Weighted Gene Correlation Network Analysis (WGCNA; v1.70-3) following a standard pipeline. Scale-free topology ($R^2 > 0.8$) was achieved by applying a soft threshold power of 8 into a signed network model. The adjacency matrices were constructed using the average linkage hierarchical clustering of the topological overlap dissimilarity matrix (1-TOM). Co-expression modules were defined using a dynamic tree cut method with minimum module size of 50 genes and deep split parameter of 4. Modules highly correlated with each other, corresponding to a module eigengene (ME) correlation > 0.75, were merged, resulting in a total of 23 modules. Modules were labelled according to their size.

We calculated the Spearman correlation between each module eigengene and the following clinical variables: age of disease onset, age at death, disease duration (years), site of disease onset (bulbar or limb), *C9orf72* status, sex, and *tSTMN2* abundance. *tSTMN2* abundance in TPM for the matching samples was extracted from the supplementary data from[3].

Cell-type and glial activation genes were tested for enrichment within each module using Fisher's exact test using a background set of 16,922 genes, followed by Bonferroni correction for the number of tests performed. Gene ontology biological process terms were tested for enrichment using the gProfiler2

package (v0.2.0)[34]. Terms with less than 10 genes were removed before correction for multiple testing. Enriched terms were then manually grouped into sets for presentation. Full module assignments, eigengenes, and enrichment results are shared as **Supplementary Tables 6-10.**

Quantitative Trait Loci mapping

To perform expression QTL (eQTL) mapping, we created a pipeline based on the one created by the GTEX consortium. We completed a separate normalization and filtering method to previous analyses. Gene expression matrices were created from the RSEM output using tximport[47]. Matrices were then converted to GCT format, TMM normalized, filtered for lowly expressed genes, removing any gene with less than 0.1 TPM in 20% of samples and at least 6 counts in 20% of samples. Each gene was then inverse-normal transformed across samples. PEER[48] factors were calculated to estimate hidden confounders within our expression data. We created a combined covariate matrix that included the PEER factors and the first 5 genotyping principal component values as input to the analysis. We tested numbers of PEER factors from 0 to 30 and found that between 10 and 30 factors produced the largest number of eGenes in each region (**Supplementary Fig. 23**).

To test for cis-eQTLs, linear regression was performed using the tensorQTL (v1.0.5)[49] *cis_nominal* mode for each SNP-gene pair using a 1 megabase window within the transcription start site (TSS) of a gene. To test for association between gene expression and the top variant in cis we used tensorQTL cis permutation pass per gene with 1000 permutations. To identify eGenes, we performed q-value correction of the permutation P-values for the top association per gene at a threshold of 0.05.

We performed splicing quantitative trait loci (sQTL) analysis using the splice junction read counts generated by regtools (v0.5.1)[50]. Junctions were clustered using Leafcutter (psi_2019 branch)[51], specifying for each junction in a cluster a maximum length of 100kb. Following the GTEx pipeline, introns without read counts in at least 50% of samples or with fewer than 10 read counts in at least 10% of samples were removed. Introns with insufficient variability across samples were removed. Filtered counts were then normalized using *prepare_phenotype_table.py* from Leafcutter, merged, and converted to BED format, using the coordinates from the middle of the intron cluster. We created a combined covariate matrix that included the PEER factors and the first 5 genotype principal components as input to the analysis. We mapped sQTLs with between 0 and 30 PEER factors as covariates in our QTL model and determined 5 and 15 factors produce the largest number of sGenes (**Supplementary Fig. 23**).

To test for cis sQTLs, linear regression was performed using the tensorQTL nominal pass for each SNP-junction pair using a 100kb window from the center of each intron cluster. To test for association between

intronic ratio and the top variant in cis we used tensorQTL permutation pass, grouping junctions by their cluster using --grp option. To identify significant clusters, we performed q-value[52] correction using a threshold of 0.05.

We estimated pairwise replication (π1) of eQTLs and sQTLs using the q-value R package. This involves taking the SNP-gene pairs that are significant at q-value < 0.05 in the discovery dataset and extracting the unadjusted P-values for the matched SNP-gene pairs in the replication dataset.

GTEX Spinal Cord QTL summary statistics

Full summary statistics for the cervical spinal cord expression QTLs (v8) were downloaded from the eQTL catalogue (**see URLs**). The splicing QTLs were downloaded from the Google Cloud portal. Top associations for each gene were downloaded from the GTEx portal.

Genome-wide association study summary statistics

Full summary statistics for the 2018 ALS GWAS[53] were downloaded from the EBI GWAS Catalogue, which have lifted over the variants to the hg38 build. Genome-wide significant loci were taken to be the most significant variants within 1 megabase at a threshold of P < 5e-8. Subthreshold loci were defined at a relaxed threshold of P < 1e-5. Loci were named by their nearest protein-coding gene using SNPnexus (v4)[54].

Colocalization analysis

We used coloc (v3.2-1)[55] to test whether SNPs from different loci in the ALS GWAS colocalized with expression and splicing QTLs from the spinal cord. For each genome-wide and subthreshold locus in the ALS GWAS we extracted the nominal summary statistics of association for all SNPs within 1 megabase either upstream/downstream of the top lead SNP (2Mb-wide region total). In each QTL dataset we then extracted all nominal associations for all SNP-gene pairs within that range and tested for colocalization between the GWAS locus and each gene. To avoid spurious colocalization caused by long range linkage disequilibrium, we restricted our colocalizations to GWAS SNP - eQTL SNP pairs where the distance between their respective top SNPs was ≤ 500kb or the two lead SNPs were in moderate linkage disequilibrium (r^2 > 0.1), taken from the 1000 Genomes (Phase 3) European populations using the

LDLinkR package (v1.1.2)[56]. For splicing QTLs we followed the same approach but collapsed junctions to return only the highest PP4 value for each gene in each locus. Due to the smaller window of association (100kb from the center of the intron excision cluster) we restricted reported colocalizations to cases where the GWAS SNP and the top sQTL SNP were either within 100kb of each other or in moderate linkage disequilibrium ($r^2 > 0.1$).

All plots were created using ggplot2 (v3.3.3)[61] in R (version 4.0.4), with ggrepel (v0.9.1)[62], ggfortify (v0.4.11)[63], patchwork (v1.1.1)[64], ggbreak (v0.0.9)[65], and ggbio (v1.38.0)[66] for additional layers of visualization.

Data availability

All raw RNA-seq data can be accessed via the NCBI's GEO database (GEO GSE137810, GSE124439, GSE116622, and GSE153960). Processed gene expression count matrices with de-identified metadata have been deposited on Zenodo (10.5281/zenodo.6385747) and we provide an RMarkdown vignette on downloading them and performing differential expression (**see URLs**). In addition, we provide an interactive R Shiny app to visualise the gene expression and other clinical variable associations (**see URLs**). Full summary statistics for expression and splicing QTLs have been deposited on Zenodo (10.5281/zenodo.5248758). All TWAS weight files have been deposited on Zenodo (10.5281/zenodo.5256613). All RNA-seq and whole genome sequencing data generated by the NYGC ALS Consortium are made immediately available to all members of the Consortium and with other consortia with whom we have a reciprocal sharing arrangement. To request immediate access to new and ongoing data generated by the NYGC ALS Consortium and for samples provided through the Target ALS Postmortem Core, complete a genetic data request form at CGND_help@nygenome.org. All whole genome sequencing data will be deposited on dbGaP at the conclusion of the project in late 2023.

Code availability

All analysis code written in R is available in Rmarkdown workbooks in a Github repository, and specific data processing pipelines are in separate repositories (**see URLs**).

URLs

Website associated with this manuscript, including all code notebooks written for this project: https://jackhump.github.io/ALS_SpinalCord_QTLs/

Gene expression counts and TPMs with de-identified metadata:
https://zenodo.org/record/6385747
Code vignette demonstrating how to download data and perform differential expression with R:
https://jackhump.github.io/ALS_SpinalCord_QTLs/html/DE_Vignette.html
R Shiny app for visualisation:
https://jackhumphrey.shinyapps.io/als_spinal_cord_browser/
Full QTL summary statistics:
https://zenodo.org/record/5248758
Full TWAS weights:
https://doi.org/10.5281/zenodo.5256613
Molecular Signatures Database (MSigDb):
http://www.gsea-msigdb.org/gsea/msigdb/index.jsp
Kelley et. al. gene fidelity marker genes:
http://oldhamlab.ctec.ucsf.edu/data-download/
Neuroexpresso marker genes:
http://neuroexpresso.org/
PanglaoDB marker genes:
https://panglaodb.se/
ENCODE Blacklist:
https://github.com/Boyle-Lab/Blacklist/blob/master/lists/hg38-blacklist.v2.bed.gz
WGS QC pipeline:
https://github.com/jackhump/WGS-QC-Pipeline
QTL mapping pipeline:
https://github.com/RajLabMSSM/QTL-mapping-pipeline
DLPFC TWAS weights:
http://gusevlab.org/projects/fusion/#reference-functional-data
ExpansionHunter:
https://github.com/Illumina/ExpansionHunter
SNPNexus:
https://www.snp-nexus.org/v4/
VCFs of 1000 Genomes samples:
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20190312_biallelic_SNV_and_INDEL/

Methods-only references

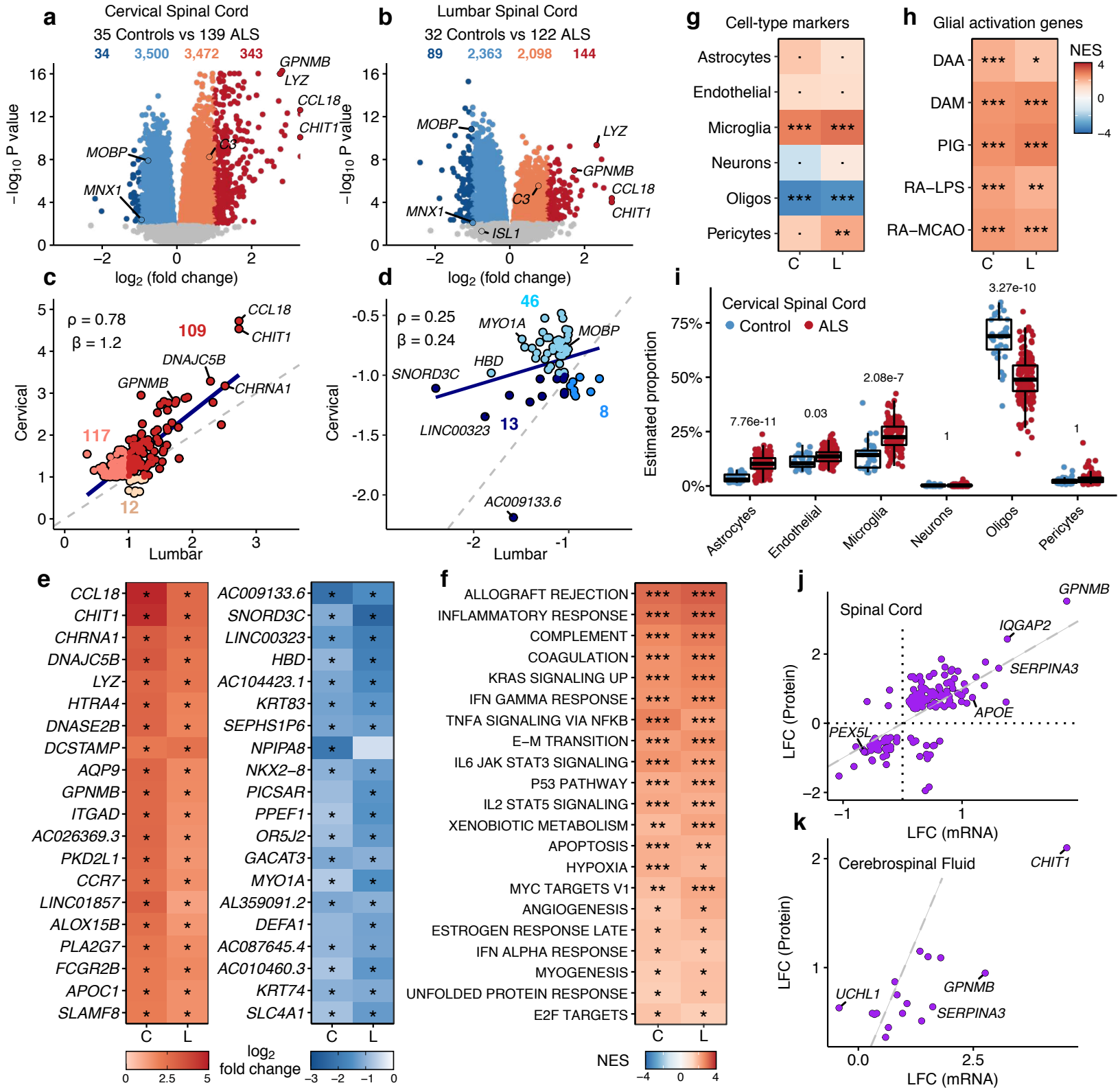
1. Dolzhenko, E. *et al.* ExpansionHunter: A sequence-graph based tool to analyze variation in short tandem repeat regions. Preprint at https://doi.org/10.1101/572545.
2. Tam, O. H. *et al.* Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia. *Cell Rep.* **29**, 1164–1177.e5 (2019).
3. Prudencio, M. *et al.* Truncated stathmin-2 is a marker of TDP-43 pathology in frontotemporal dementia. *J. Clin. Invest.* (2020) doi:10.1172/JCI139741.
4. Schroeder, A. *et al.* The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.* **7**, 3 (2006).
5. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
6. Bolduc, B. Quality Control of Reads Using Trimmomatic (Cyverse) v1 (protocols.io.ewbbfan). *protocols.io* Preprint at https://doi.org/10.17504/protocols.io.ewbbfan.
7. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
8. Harrow, J., Frankish, A., Gonzalez, J. M. & Frazer, K. A. GENCODE : The reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
9. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
10. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
11. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
12. Consortium, T. G. & The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* vol. 369 1318–1330 Preprint at https://doi.org/10.1126/science.aaz1776 (2020).
13. Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, 483 (2016).
14. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
15. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
16. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-content normalization for RNA-Seq data. *BMC Bioinformatics* **12**, 480 (2011).
17. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
18. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology* vol. 16 284–287 Preprint at https://doi.org/10.1089/omi.2011.0118 (2012).
19. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425 (2015).
20. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of*

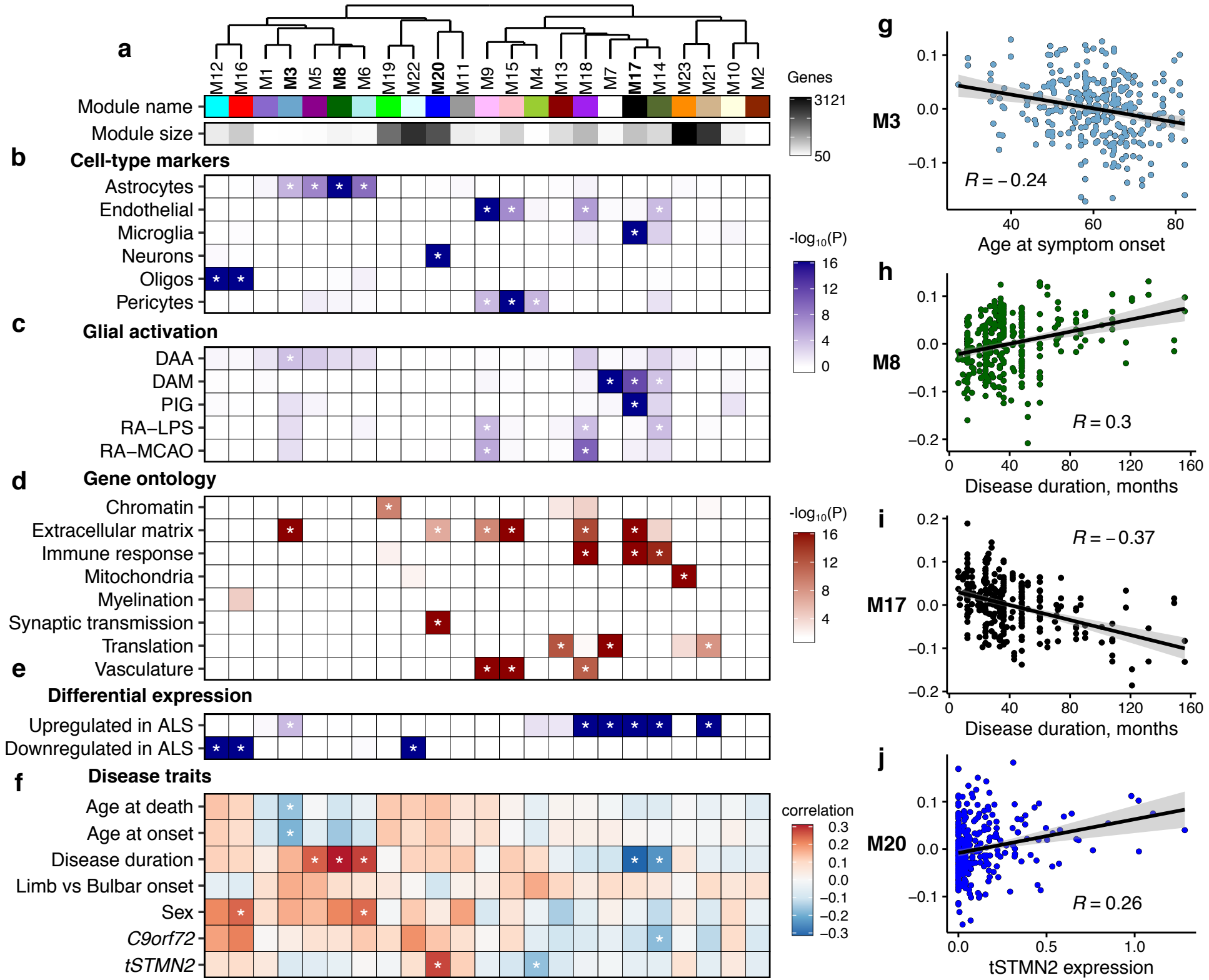
Sciences vol. 112 7285–7290 Preprint at https://doi.org/10.1073/pnas.1507125112 (2015).

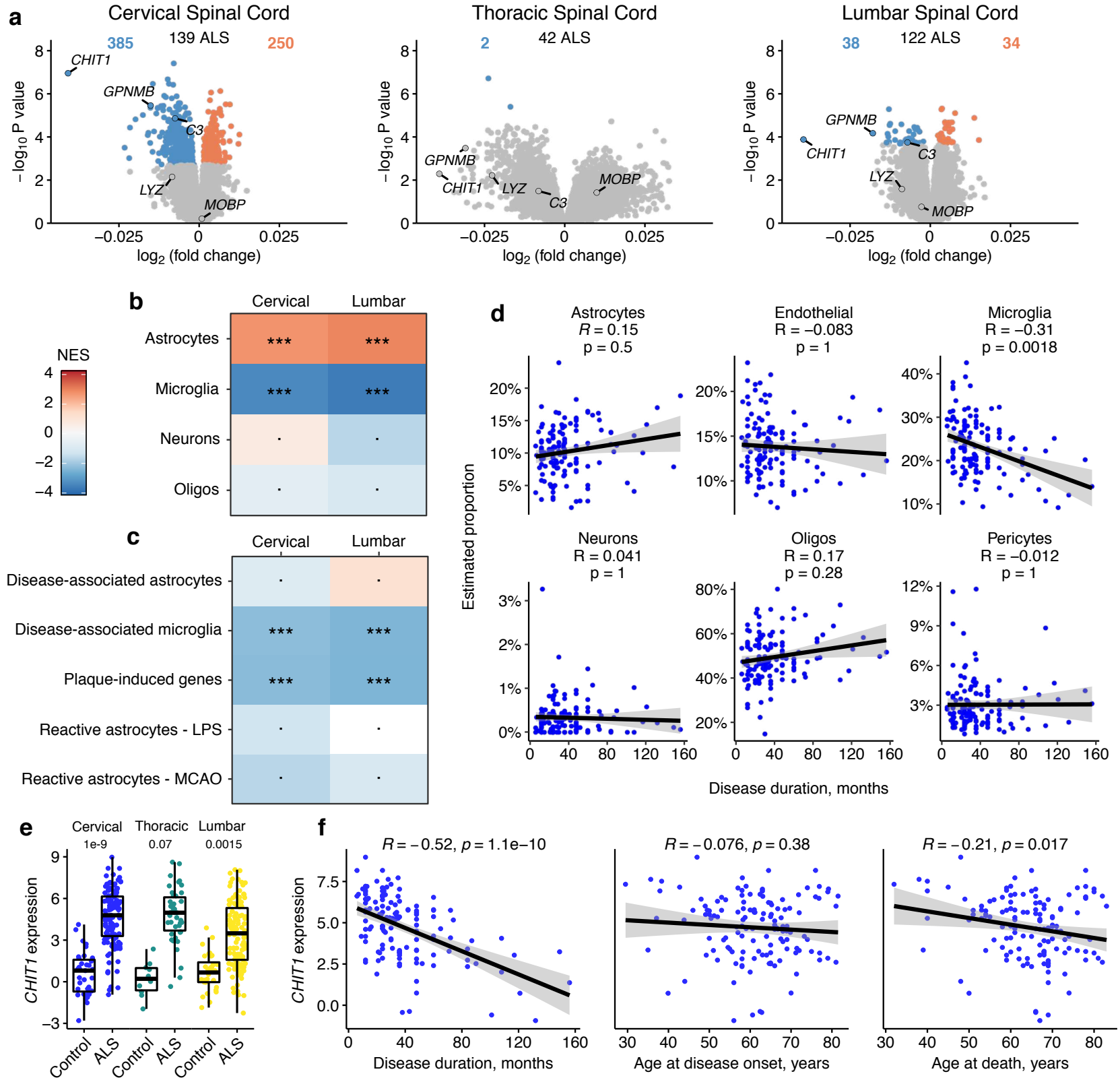
21. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).
22. Kelley, K. W., Nakao-Inoue, H., Molofsky, A. V. & Oldham, M. C. Variation among intact tissue samples reveals the core transcriptional features of human CNS cell classes. *Nat. Neurosci.* **21**, 1171–1184 (2018).
23. Franzén, O., Gan, L.-M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* **2019**, (2019).
24. Mancarci, B. O. *et al.* Cross-Laboratory Analysis of Brain Cell Type Transcriptomes with Applications to Interpretation of Bulk Tissue Data. *eNeuro* **4**, (2017).
25. Keren-Shaul, H. *et al.* A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease. *Cell* **169**, 1276–1290.e17 (2017).
26. Habib, N. *et al.* Disease-associated astrocytes in Alzheimer's disease and aging. *Nat. Neurosci.* **23**, 701–706 (2020).
27. Chen, W.-T. *et al.* Spatial Transcriptomics and In Situ Sequencing to Study Alzheimer's Disease. *Cell* **182**, 976–991.e19 (2020).
28. Zamanian, J. L. *et al.* Genomic analysis of reactive astrogliosis. *J. Neurosci.* **32**, 6391–6410 (2012).
29. Mancarci, O. & French, L. Homologene: quick access to homologene and gene annotation updates. *R package version* **1**, 68 (2019).
30. Oeckl, P. *et al.* 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 (2020).
31. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**, 380 (2019).
32. Hunt, G. J., Freytag, S., Bahlo, M. & Gagnon-Bartsch, J. A. dtangle: accurate and robust cell type deconvolution. *Bioinformatics* **35**, 2093–2099 (2019).
33. Skene, N. G. & Grant, S. G. N. Identification of Vulnerable Cell Types in Major Brain Disorders Using Single Cell Transcriptomes and Expression Weighted Cell Type Enrichment. *Front. Neurosci.* **10**, 16 (2016).
34. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* **35**, W193–W200 (2007).
35. Regier, A. A. *et al.* Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat. Commun.* **9**, 4038 (2018).
36. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at https://doi.org/10.1101/201178.
37. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci. Rep.* **9**, 9354 (2019).
38. Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv* 787903 (2019) doi:10.1101/787903.
39. Adelson, R. P. *et al.* Empirical design of a variant quality control pipeline for whole genome sequencing data using replicate discordance. *Sci. Rep.* **9**, 16156 (2019).
40. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
41. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

42. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* vol. 4 Preprint at https://doi.org/10.1186/s13742-015-0047-8 (2015).
43. Köster, J. & Rahmann, S. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
44. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
45. Fort, A. *et al.* MBV: a method to solve sample mislabeling and detect technical bias in large combined genotype and sequencing assay datasets. *Bioinformatics* **33**, 1895–1897 (2017).
46. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
47. Love, M. I., Soneson, C. & Robinson, M. D. Importing transcript abundance datasets with tximport. *dim (txi. inf. rep $ infReps $ sample1)* **1**, 5 (2017).
48. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500 (2012).
49. Taylor-Weiner, A. *et al.* Scaling computational genomics to millions of individuals with GPUs. *Genome Biol.* **20**, 228 (2019).
50. Feng, Y.-Y. *et al.* RegTools: Integrated analysis of genomic and transcriptomic data for discovery of splicing variants in cancer. *bioRxiv* 436634 (2018) doi:10.1101/436634.
51. Li, Y. I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).
52. Storey, J. D. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Stat.* **31**, 2013–2035 (2003).
53. Nicolas, A. *et al.* Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. *Neuron* **97**, 1268–1283.e6 (2018).
54. Oscanoa, J. *et al.* SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update). *Nucleic Acids Res.* **48**, W185–W192 (2020).
55. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
56. Myers, T. A., Chanock, S. J. & Machiela, M. J. LDlinkR: An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations. *Front. Genet.* **11**, 157 (2020).
57. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
58. Lowy-Gallego, E. *et al.* Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project. *Wellcome Open Res* **4**, 50 (2019).
59. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
60. Li, Y. I., Wong, G., Humphrey, J. & Raj, T. Prioritizing Parkinson's disease genes using population-scale transcriptomic data. *Nat. Commun.* **10**, 994 (2019).
61. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis.* (Springer-Verlag New York, 2009).
62. Slowikowski, K. ggrepel: Repulsive Text and Label Geoms for 'ggplot2', 2016. *R package version 0. 5*.
63. Tang, Y., Horikoshi, M. & Li, W. ggfortify: Unified interface to visualize statistical results of popular R packages. *R J.* **8**, 474 (2016).
64. Pedersen, T. L. patchwork: The Composer of Plots. *R package version* **1**, 410 (2019).
65. Xu, S. *et al.* Use ggbreak to Effectively Utilize Plotting Space to Deal With Large Datasets and Outliers. *Front. Genet.* **12**, 774846 (2021).

66. Yin, T., Cook, D. & Lawrence, M. ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biol.* **13**, R77 (2012).

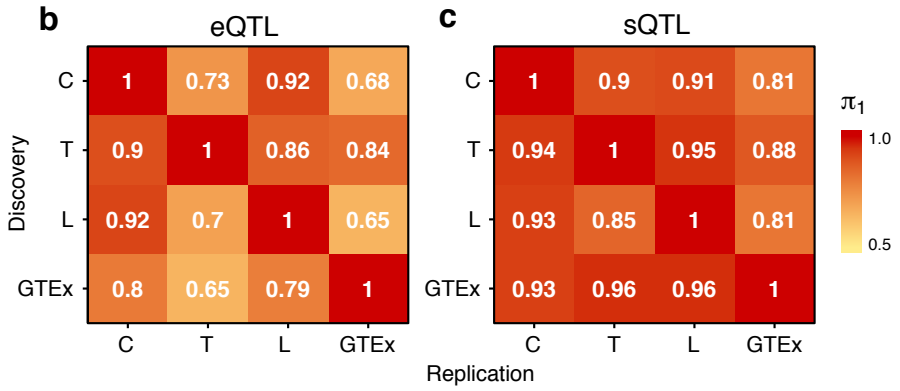

a
Cervical Spinal Cord
35 Controls vs 139 ALS
34
3,500
3,472
343
GPNMB
LYZ
CCL18
CHIT1
MOBP
C3
MNX1
−log10 P value
log2 (fold change)
b
Lumbar Spinal Cord
32 Controls vs 122 ALS
89
2,363
2,098
144
MOBP
LYZ
GPNMB
CCL18
CHIT1
C3
MNX1
ISL1
c
ρ = 0.78
β = 1.2
109
CCL18
CHIT1
DNAJC5B
GPNMB
CHRNA1
117
12
Cervical
Lumbar
d
ρ = 0.25
β = 0.24
46
MYO1A
MOBP
HBD
SNORD3C
8
LINC00323
13
AC009133.6
Cervical
Lumbar
g
Cell-type markers
Astrocytes
Endothelial
Microglia
Neurons
Oligos
Pericytes
C
L
h
Glial activation genes
NES
DAA
DAM
PIG
RA–LPS
RA–MCAO
i
Cervical Spinal Cord
Control
ALS
Estimated proportion
7.76e-11
0.03
2.08e-7
1
3.27e-10
1
Astrocytes
Endothelial
Microglia
Neurons
Oligos
Pericytes
e
CCL18
CHIT1
CHRNA1
DNAJC5B
LYZ
HTRA4
DNASE2B
DCSTAMP
AQP9
GPNMB
ITGAD
AC026369.3
PKD2L1
CCR7
LINC01857
ALOX15B
PLA2G7
FCGR2B
APOC1
SLAMF8
AC009133.6
SNORD3C
LINC00323
HBD
AC104423.1
KRT83
SEPHS1P6
NPIPA8
NKX2–8
PICSAR
PPEF1
OR5J2
GACAT3
MYO1A
AL359091.2
DEFA1
AC087645.4
AC010460.3
KRT74
SLC4A1
C
L
log2 fold change
f
ALLOGRAFT REJECTION
INFLAMMATORY RESPONSE
COMPLEMENT
COAGULATION
KRAS SIGNALING UP
IFN GAMMA RESPONSE
TNFA SIGNALING VIA NFKB
E–M TRANSITION
IL6 JAK STAT3 SIGNALING
P53 PATHWAY
IL2 STAT5 SIGNALING
XENOBIOTIC METABOLISM
APOPTOSIS
HYPOXIA
MYC TARGETS V1
ANGIOGENESIS
ESTROGEN RESPONSE LATE
IFN ALPHA RESPONSE
MYOGENESIS
UNFOLDED PROTEIN RESPONSE
E2F TARGETS
C
L
NES
j
Spinal Cord
GPNMB
IQGAP2
SERPINA3
APOE
PEX5L
LFC (Protein)
LFC (mRNA)
k
Cerebrospinal Fluid
CHIT1
UCHL1
GPNMB
SERPINA3
LFC (Protein)
LFC (mRNA)


a
M12
M16
M1
M3
M5
M8
M6
M19
M22
M20
M11
M9
M15
M4
M13
M18
M7
M17
M14
M23
M21
M10
M2
Module name
Module size
Genes
3121
50
b
Cell-type markers
Astrocytes
Endothelial
Microglia
Neurons
Oligos
Pericytes
-log10(P)
16
12
8
4
0
c
Glial activation
DAA
DAM
PIG
RA–LPS
RA–MCAO
d
Gene ontology
Chromatin
Extracellular matrix
Immune response
Mitochondria
Myelination
Synaptic transmission
Translation
Vasculature
-log10(P)
16
12
8
4
e
Differential expression
Upregulated in ALS
Downregulated in ALS
f
Disease traits
Age at death
Age at onset
Disease duration
Limb vs Bulbar onset
Sex
C9orf72
tSTMN2
correlation
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
g
M3
R = −0.24
Age at symptom onset
h
M8
R = 0.3
Disease duration, months
i
M17
R = −0.37
Disease duration, months
j
M20
R = 0.26
tSTMN2 expression

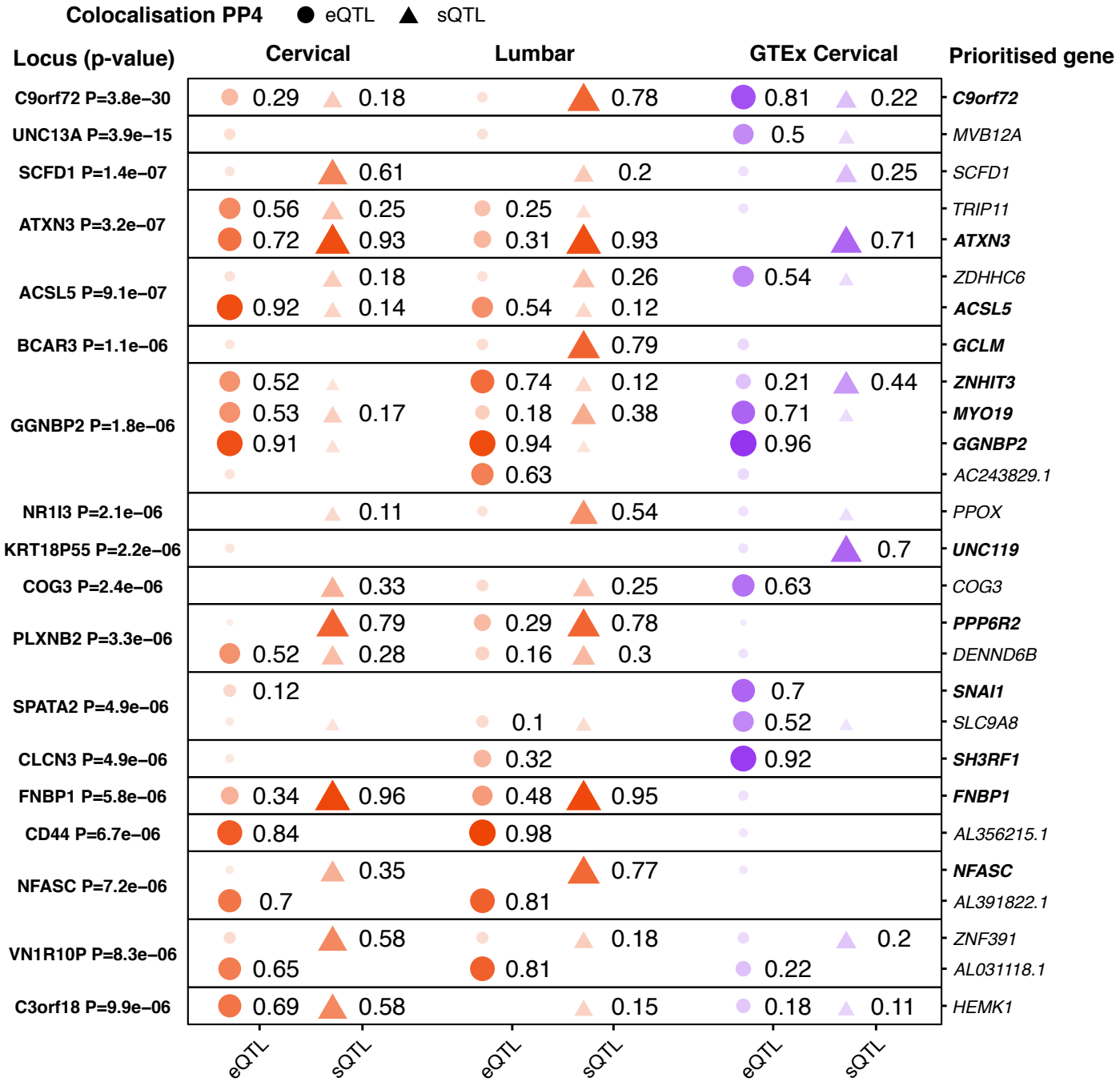

a
Cervical Spinal Cord
385
139 ALS
250
Thoracic Spinal Cord
2
42 ALS
Lumbar Spinal Cord
38
122 ALS
34
−log10 P value
log2 (fold change)
CHIT1
GPNMB
C3
LYZ
MOBP
b
Cervical
Lumbar
NES
Astrocytes
Microglia
Neurons
Oligos
c
Disease-associated astrocytes
Disease-associated microglia
Plaque-induced genes
Reactive astrocytes - LPS
Reactive astrocytes - MCAO
d
Astrocytes
R = 0.15
p = 0.5
Endothelial
R = −0.083
p = 1
Microglia
R = −0.31
p = 0.0018
Neurons
R = 0.041
p = 1
Oligos
R = 0.17
p = 0.28
Pericytes
R = −0.012
p = 1
Estimated proportion
Disease duration, months
e
Cervical
Thoracic
Lumbar
1e-9
0.07
0.0015
CHIT1 expression
Control
ALS
f
R = −0.52, p = 1.1e−10
R = −0.076, p = 0.38
R = −0.21, p = 0.017
CHIT1 expression
Disease duration, months
Age at disease onset, years
Age at death, years

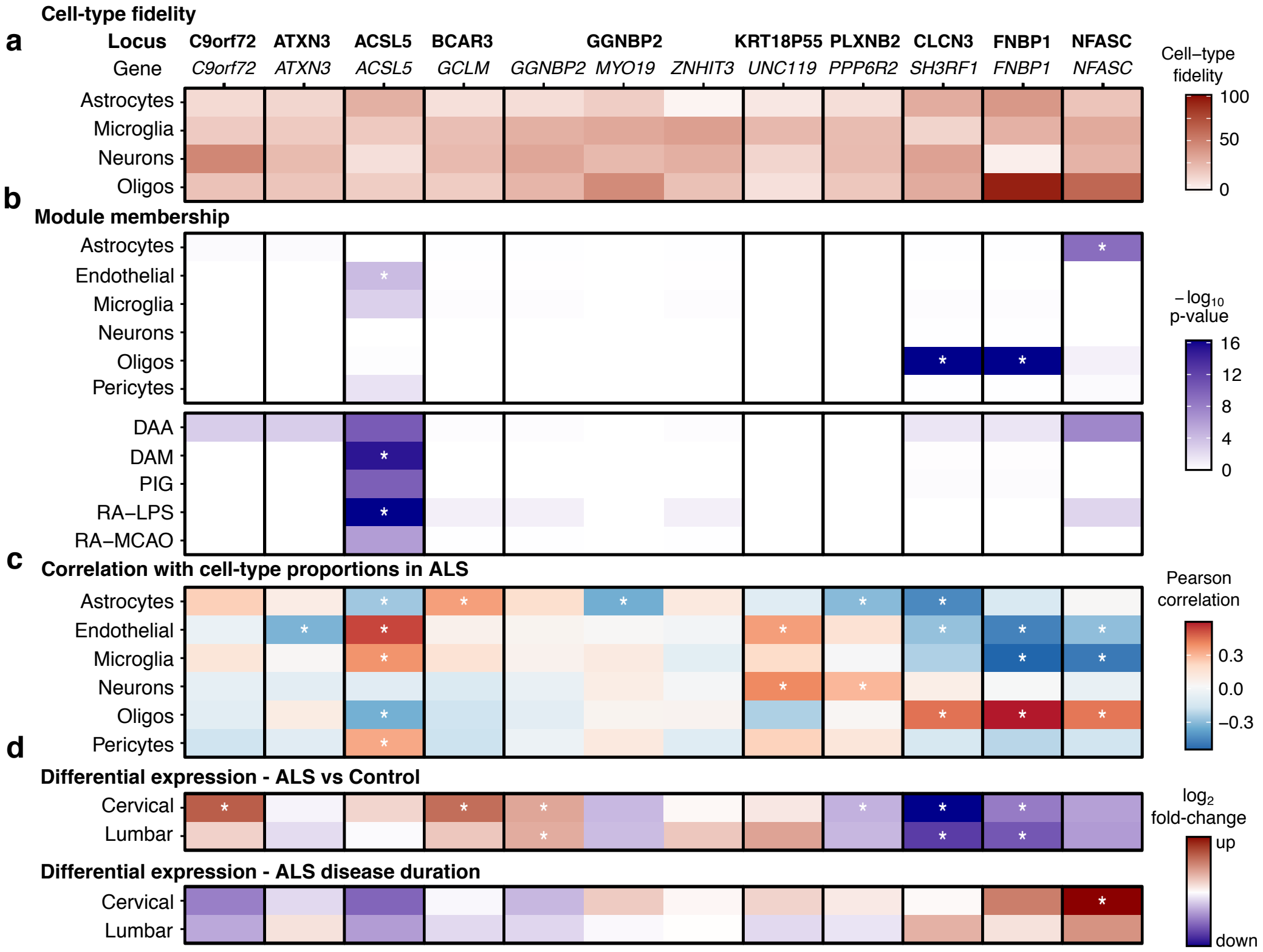
a

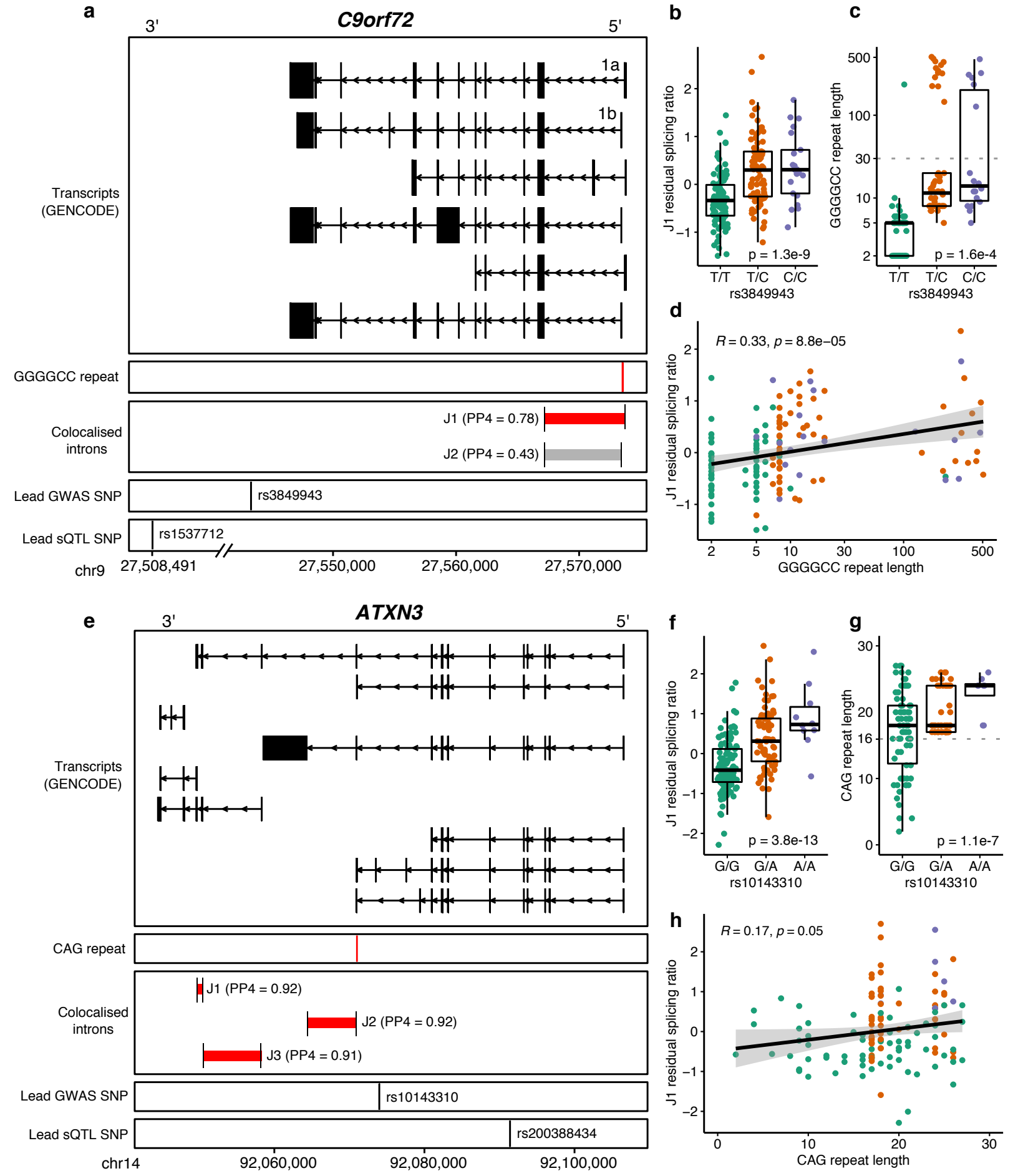
Tissue	N	eGenes	sGenes
Cervical	216	7890	4730
Lumbar	197	6751	4302
Thoracic	68	962	1387
GTEx Cervical	*126*	*3414*	*965*



d




a
Cell-type fidelity
Locus
Gene
C9orf72
ATXN3
ACSL5
BCAR3
GGNBP2
KRT18P55
PLXNB2
CLCN3
FNBP1
NFASC
C9orf72
ATXN3
ACSL5
GCLM
GGNBP2
MYO19
ZNHIT3
UNC119
PPP6R2
SH3RF1
FNBP1
NFASC
Astrocytes
Microglia
Neurons
Oligos
Cell–type fidelity
100
50
0
b
Module membership
Astrocytes
Endothelial
Microglia
Neurons
Oligos
Pericytes
DAA
DAM
PIG
RA–LPS
RA–MCAO
– log10 p-value
16
12
8
4
0
c
Correlation with cell-type proportions in ALS
Astrocytes
Endothelial
Microglia
Neurons
Oligos
Pericytes
Pearson correlation
0.3
0.0
–0.3
d
Differential expression - ALS vs Control
Cervical
Lumbar
log2 fold-change
up
down
Differential expression - ALS disease duration
Cervical
Lumbar


a
C9orf72
3'
5'
1a
1b
Transcripts (GENCODE)
GGGGCC repeat
Colocalised introns
J1 (PP4 = 0.78)
J2 (PP4 = 0.43)
Lead GWAS SNP
rs3849943
Lead sQTL SNP
rs1537712
chr9
27,508,491
27,550,000
27,560,000
27,570,000
b
J1 residual splicing ratio
p = 1.3e-9
T/T
T/C
C/C
rs3849943
c
GGGGCC repeat length
p = 1.6e-4
T/T
T/C
C/C
rs3849943
d
R = 0.33, p = 8.8e−05
J1 residual splicing ratio
GGGGCC repeat length
e
ATXN3
3'
5'
Transcripts (GENCODE)
CAG repeat
Colocalised introns
J1 (PP4 = 0.92)
J2 (PP4 = 0.92)
J3 (PP4 = 0.91)
Lead GWAS SNP
rs10143310
Lead sQTL SNP
rs200388434
chr14
92,060,000
92,080,000
92,100,000
f
J1 residual splicing ratio
p = 3.8e-13
G/G
G/A
A/A
rs10143310
g
CAG repeat length
p = 1.1e-7
G/G
G/A
A/A
rs10143310
h
R = 0.17, p = 0.05
J1 residual splicing ratio
CAG repeat length