nature communications

Article

https://doi.org/10.1038/s41467-025-64622-5

Gene expression signatures from whole blood predict amyotrophic lateral sclerosis case status and survival

Received: 27 February 2025

Accepted: 23 September 2025

Published online: 31 October 2025

Check for updates

Yue Zhao[1], Masha G. Savelieff[2], Xiayan Li[1], Kai Guo[3,4], Kai Wang[1], Minghua Li[1], Bo Li[1], Gayatri Iyer[1], Stacey A. Sakowski[3,4], Lili Zhao[5], Samuel J. Teener[4], Kelly M. Bakulski[6], John F. Dou[6], Bryan J. Traynor[7,8], Alla Karnovsky[1], Stuart A. Batterman[9], Junguk Hur[2], Stephen A. Goutman[3,4], Maureen A. Sartor[1,10,11] ✉ & Eva L. Feldman[3,4,11] ✉

Amyotrophic lateral sclerosis (ALS) is a rare and fatal neurodegenerative disease with a median survival of only 2 to 4 years from diagnosis. Improved tools are needed to shorten diagnostic delays and improve prognostication to benefit clinical care. Herein, we profiled whole blood gene expression by RNA sequencing in a large cohort of ALS participants ($n = 422$) versus controls ($n = 272$). Several machine learning classifiers trained on our detailed gene expression dataset accurately predicted case-control status, including in a fully independent external test cohort, achieving an area under the receiver operating characteristic curve of 0.894 with the best performing model. Integrating gene expression features with clinical variables improved our ability to discriminate ALS cases into shorter, intermediate, and longer survival in an external dataset. Finally, we identified ALS-relevant pathways in our blood transcriptomics dataset as well as "core genes" that overlapped with gene expression changes occurring in the primary disease tissue, facilitating a drug perturbation analysis that identified several candidates. Overall, our results highlight the potential diagnostic and prognostic applications of whole blood gene expression data, with important implications for improving ALS clinical care.

ALS is a progressive, fatal, neurodegenerative disease with a median survival of only 2 to 4 years from diagnosis. ALS remains difficult to identify in routine clinical practice[1,2]. Patients manifest symptoms and signs like other more common illnesses[3] and misdiagnoses and errors are relatively frequent[4]. The median time to a definitive ALS diagnosis is 5 to 15 months, depending on subtype, and can even take up to 19 months for some patients[5]. This diagnostic delay postpones treatment, which adversely affects survival since earlier initiation with standard-of-care riluzole[6] and multidisciplinary care[7] improves clinical outcomes. Moreover, delayed diagnoses exclude many

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [2]Department of Biomedical Sciences, University of North Dakota, Grand Forks, ND, USA. [3]Department of Neurology, University of Michigan, Ann Arbor, MI, USA. [4]NeuroNetwork for Emerging Therapies, University of Michigan, Ann Arbor, MI, USA. [5]Department of Preventive Medicine (Biostatistics Division), Northwestern University, Chicago, IL, USA. [6]Department of Epidemiology, University of Michigan, Ann Arbor, MI, USA. [7]Neuromuscular Diseases Research Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA. [8]Department of Neurology, Johns Hopkins University Medical Center, Baltimore, MD, USA. [9]Department of Environmental Health Sciences, University of Michigan, Ann Arbor, MI, USA. [10]Biostatistics Department, University of Michigan, Ann Arbor, MI, USA. [11]These authors contributed equally: Maureen A. Sartor, Eva L. Feldman. ✉e-mail: sartorma@umich.edu; efeldman@umich.edu

patients from ALS trials, which usually recruit participants with less advanced disease, and also leaves patients with less time to organize their financial, legal, psychological, and spiritual affairs.

While improved diagnostic approaches are urgently needed to shorten diagnostic delays for ALS patients, no ideal biomarkers have been developed for ALS. The frontrunner is neurofilament light chain (NfL), and both serum and cerebrospinal fluid (CSF) NfL levels are significantly higher in ALS patients versus controls and increase in presymptomatic at-risk individuals as they develop ALS[8]. Nevertheless, NfL has a critical shortcoming as a biomarker. As an indicator of neuronal damage, NfL levels are also elevated in other diseases, such as mild cognitive impairment[9], Alzheimer's disease[10], Parkinson's disease[11], multiple sclerosis[12], diabetic peripheral neuropathy[13], various additional neurodegenerative illnesses[14], and even all-cause mortality[15]. Thus, NfL lacks specificity as an ALS diagnostic tool.

In lieu of assessing a single biomarker measure, gene transcriptomic profiles are feasible as clinical tools with commercial viability. The PAM50 is a U.S. Food and Drug Administration (FDA)-cleared gene expression biomarker panel that leverages the expression of 50 genes to classify breast tumor subtypes[16,17]. A recently developed 18-gene qPCR array can diagnose high-grade prostate cancer from biofluid samples[18]. Gene signatures that predict ALS case-control status based on whole blood are also reported in the literature[19–21], but only one was tested in an independent external dataset and performed poorly (63.3% accuracy, 60.0% sensitivity, 66.7% specificity, 64.7% area under the curve [AUC])[21]. The performance and accuracy of the other reported gene signatures[19,20] were not tested in independent external datasets, rendering their use as diagnostic ALS biomarker panels uncertain.

The goal of the current study was to develop a gene classifier as a future ALS biomarker panel to expedite ALS diagnosis. We profiled gene expression from accessible whole blood samples from a large cohort of ALS cases versus controls. We employed RNA sequencing (RNA-seq), detecting over 22,000 protein-coding genes, long non-coding RNAs, and microRNAs. We leveraged this rich and detailed gene expression dataset to evaluate the ability of blood gene expression to differentiate ALS cases from controls, finding high accuracy using various machine learning classifiers. The best classifier could also predict case-control status in a fully independent external test cohort[21].

We extended this success in two ways. First, we integrated gene features with clinical variables to enhance ALS survival prediction, addressing another unmet need in the field. Second, we performed pathway enrichment analysis of our blood transcriptomics dataset, which revealed ALS-relevant pathways. We selected "core genes" that overlapped with differentially expressed genes (DEGs) in the primary disease tissue and then used these core genes as input for drug perturbation analysis, which identified drug candidates for evaluation as future ALS therapies.

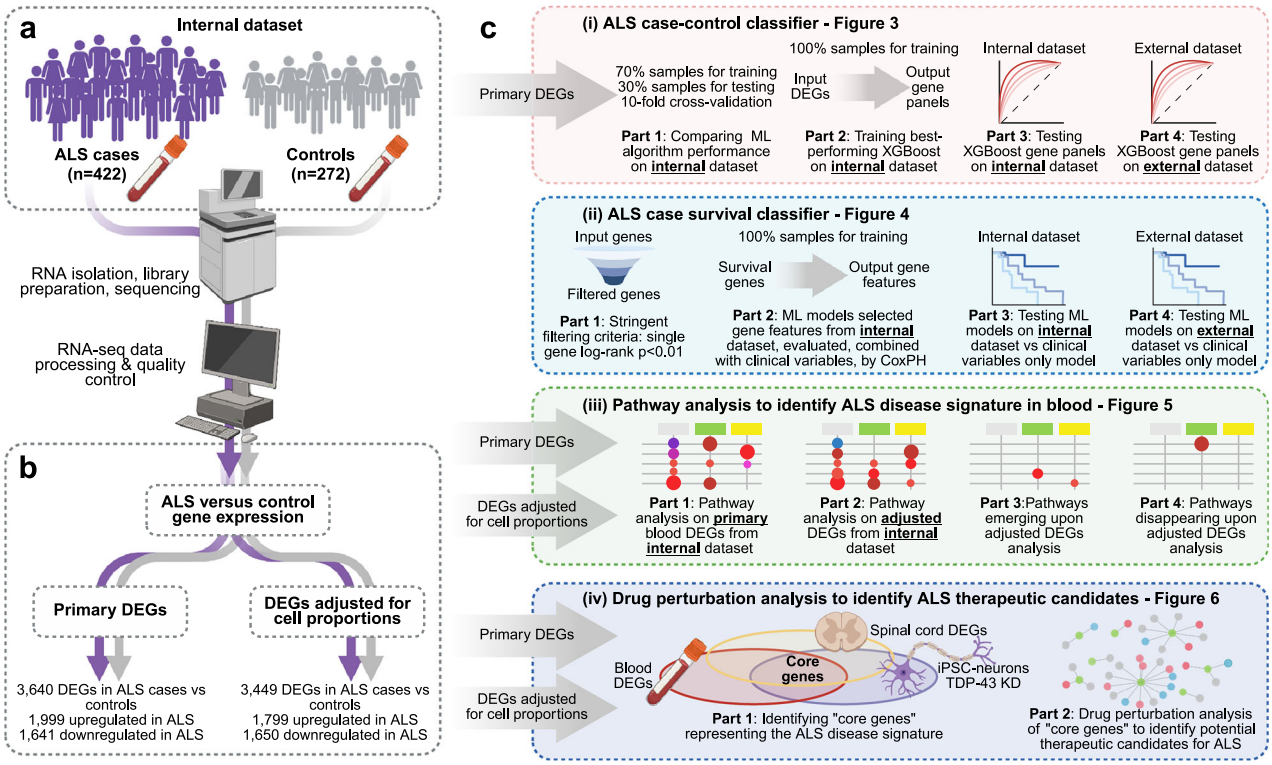


Fig. 1 | Study design overview. a Blood samples collected from cases and controls were processed for RNA-seq. **b** Differentially expressed genes (DEGs) in ALS cases ($n = 422$) versus controls ($n = 272$) were identified using both a primary (unadjusted) and an adjusted analysis for immune cell proportions. **c** (i) ALS case-control classifier: Comparison of seven machine learning (ML) algorithms (Part 1); Improving the performance of the best-performing ML algorithm XGBoost (Part 2); Testing the three gene panels plus a combined 46-gene panel on our internal and the external Grima et al.[21] datasets (Parts 3 and 4, respectively). Results shown in Fig. 3. (ii) ALS case survival classifier: Criteria applied for gene filtering (Part 1); Training two ML algorithms, XGBoost and stepwise, using gene features with clinical variables (Part 2); Comparing XGBoost and stepwise (gene features + clinical variables) to a clinical variables only model in our internal and the external Grima et al.[21] datasets (Parts 3 and 4, respectively). Results shown in Fig. 4. (iii) Pathway analysis to identify ALS disease signature in blood: Kyoto Encyclopedia of Genes and Genomes, Hallmark, and Gene Ontology pathway enrichment of primary DEGs (Part 1); Analyses on DEGs adjusted for cell proportions (Part 2); Pathway analyses of adjusted versus primary DEGs (Parts 3, 4). Results shown in Fig. 5. (iv) Drug perturbation analysis to identify ALS therapeutic candidates: Identification of "core genes" of adjusted and primary DEGs that overlap with DEGs from induced pluripotent stem cell (iPSC)-derived neurons with TDP-43 knockdown and ALS postmortem spinal cord (Part 1); Drug perturbation analysis of "core genes" (Part 2). Results shown in Fig. 6. Created in BioRender. Feldman, E. (2025) https://BioRender.com/zpmd31u.

https://doi.org/10.1038/s41467-025-64622-5

Table 1 | Clinical characteristics of ALS and control participants in primary study cohort

	ALS cases, $n = 422$	Controls, $n = 272$	p-value
Age at blood draw (years)	65.00 (57.00, 71.75)	61.00 (54.00, 66.25)	3.03×10^{-6}
Sex			7.89×10^{-4}
Female	178 (42%)	151 (56%)	
Male	244 (58%)	121 (44%)	
Race			0.39
White	404 (96%)	254 (93%)	
Black	12 (3%)	12 (4%)	
Other	6 (1%)	6 (2%)	
Family history of ALS			2.28×10^{-28}
Yes	41 (10%)	0 (0%)	
No	367 (87%)	272 (100%)	
Unknown	14 (3%)	0 (0%)	
Revised El Escorial criteria at diagnosis			
Definite	110 (26%)		
Probable	126 (30%)		
Probable, lab supported	104 (25%)		
Possible	54 (13%)		
Suspected	22 (5%)		
Unknown	6 (1%)		
Onset segment			
Bulbar	109 (26%)		
Cervical	135 (32%)		
Lumbar	164 (39%)		
Respiratory	4 (1%)		
Thoracic	7 (2%)		
Generalized / Cannot be determined	2 (0%)		
Unknown	1 (0%)		
***C9orf72* hexanucleotide repeat expansion status**			5.98×10^{-4}
Negative	268 (64%)	147 (54%)	
Unknown	126 (30%)	125 (46%)	
Intermediate	1 (0%)	0 (0%)	
Positive	27 (6%)	0 (0%)	
ALSFRS-R Score	37 (32, 41)		
Unknown	2		
Time from diagnosis to blood draw (months)	5.31 (3.27, 9.07)		
Unknown	4		
Time from symptom onset to diagnosis (months)	12.68 (7.97, 22.01)		
Unknown	3		

Median (25%,75%); n (%); two-sided Wilcoxon rank sum test; Pearson's Chi-squared test.

Results

Cohort characteristics

We profiled whole blood by RNA-seq from ALS cases ($n = 422$) versus controls ($n = 272$) (Fig. 1a; Table 1). Participants with ALS were older (median 65 versus 61 years, $p < 0.001$) with more males (58% versus 44%, $p < 0.001$) compared to controls. The ALS cohort exhibited typical disease characteristics with a median age of 65 years and a male-to-female ratio of 1.38[22]. 87% cited no known family history. Most were *C9orf72* hexanucleotide repeat expansion negative (64%) or unknown (30%). ALS participants were 26% bulbar, 32% cervical, and 39% lumbar, similar to prior cohorts[23], with a median ALS functional rating scale, revised (ALSFRS-R) score of 37 (range 32 to 41), indicating a moderate level of functional impairment.

A gene expression signature differentiates ALS cases from controls

Overall, we identified 3,640 DEGs in ALS cases ($n = 422$) versus controls ($n = 272$), of which 1999 were upregulated and 1641 downregulated in ALS (Fig. 1b). As expected in ALS[24–26], many DEGs were related to the immune system, e.g., *IL2RB*, *S100A8*, *S100A9*, *S100A12*. The large sample size allowed us to perform a sex-stratified DEG analysis in males ($n = 365$) and females ($n = 329$), which identified a similar number of up- and downregulated genes in ALS and aligned with reported sex differences in ALS in immune cell levels and activation state[24–26]. (Fig. 2a, Supplementary Data 1). Although most DEGs overlapped between males and females, several deviated in their effect size, i.e., $\log_2$ fold-change (Fig. 2b), most generally of low expression. Among DEGs deviating by sex, some had relevance to ALS, including the *U1* small nuclear RNA in *FUS* mutant ALS[27], the environmental toxin detoxifier *GSTM5*, and *RGS17*, a regulator of G protein protein-coupled receptor signaling cascades, including the muscarinic acetylcholine receptor and dopamine receptor[28].

The top 10 up- and 10 downregulated DEGs in ALS clearly differentiated ALS cases from controls (Fig. 2c), highlighting a distinct transcriptomic ALS signature, both in males and females. There were several interesting candidates among the top upregulated DEGs, including an E3 ubiquitin ligase (*MARCHF7*), vesicular and endosomal trafficking proteins (*SNX13*, *RAB8B*, *ZFYVE16*), autophagy proteins (*VMP1*), and muscle proteins (*CAPZA1*, *CAPZA2*). Among ALS downregulated DEGs were candidates linked to epigenetics (*DNMT1*, *EP400*), inflammation (*ILF3*), and apoptosis (*MADD*).

Next, we compared our DEGs to the published datasets on ALS whole blood transcriptomics. The first comprises ALS cases ($n = 396$) and controls ($n = 645$) by microarray, originally collected by van Rheenen et al.[19] and reanalyzed by Swindell et al.[20]; since we compared our DEGs to the Swindell DEGs, we refer to it hereon as the "Swindell" dataset. The second published dataset comprises ALS cases ($n = 86$) and controls ($n = 48$) profiled by RNA-seq by Grima et al.[21] (Supplementary Data 2). We matched 1928 of our DEGs based on gene symbol to literature DEGs. Of our upregulated ALS DEGs, 381 and 39 overlapped with approximately 50% of Swindell and 46% of Grima DEGs, respectively (Fig. 2d). The extent of overlap was lower among downregulated DEGs, with only 267 shared with Swindell (35%) and 56 with Grima (35%) (Fig. 2e). Nevertheless, both upregulated and downregulated Swindell DEGs were significantly enriched among our upregulated (odds ratio 3.46 [95% confidence interval (CI): 3.03, Inf], $p = 1.31 \times 10^{-54}$) and downregulated (odds ratio 2.13 [95%CI: 1.86, Inf], $p = 2.22 \times 10^{-19}$) DEGs, respectively (Supplementary Fig. S1a). Additionally, upregulated and downregulated Grima DEGs were significantly enriched among our upregulated (odds ratio 4.44 [95%CI: 3.02, Inf], $p = 1.26 \times 10^{-10}$) and downregulated (odds ratio 3.48 [95%CI: 2.60, Inf], $p = 5.71 \times 10^{-12}$) DEGs, respectively (Supplementary Fig. S1b). The DEGs consistent across the three studies were enriched with "Alzheimer's disease", as well as other relevant terms, including "IL-17 signaling pathway", "autophagy – animal", "mTORC1 signaling", and "interferon alpha response". Among neurodegenerative disease pathways, "Parkinson disease", "oxidative phosphorylation", and relevant immune pathways were uniquely enriched in our DEGs (Supplementary Fig. S1c).

Shared DEGs had higher average expression (Fig. 2f) and significance (Fig. 2g). Although shared DEGs were relatively independent of extent of fold-change (Fig. 2h) there was a strong correlation of our

https://doi.org/10.1038/s41467-025-64622-5

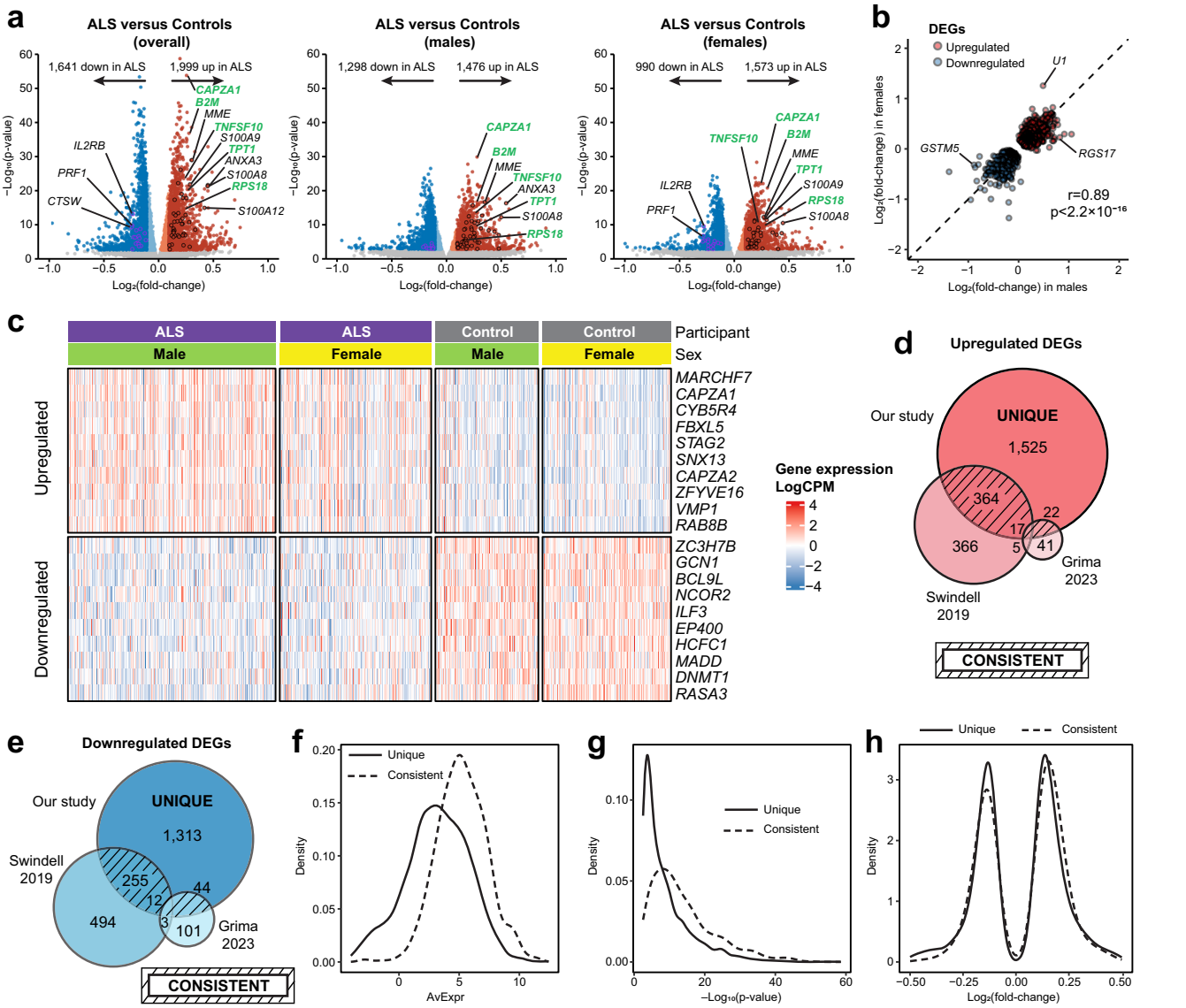


Fig. 2 | A gene expression signature differentiates ALS cases ($n = 422$) from controls ($n = 272$). **a** Volcano plots of DEGs in ALS cases versus controls in all participants (left), males (middle), and females (right). Bue and red dots represent upregulated and downregulated DEGs, respectively, in ALS meeting the criteria of absolute value of $\log_2$(fold-change)>0.1 (absLFC) and false discovery rate (FDR) < 0.01. Some immune-related DEGs annotated; neutrophil markers (black outlined circles); natural killer cell markers (purple outlined circles). Gray, light red, and light blue dots denote genes not meeting the DEG criteria. The five selected upregulated DEGs validated by qPCR are indicated in green font. **b** Scatter plot of male versus female LFC for overall significant DEGs. Circles represent upregulated and downregulated DEGs; r is the Pearson correlation coefficient; p-value by two-sided Pearson correlation test. **c** Heatmaps of 10 most upregulated (top) and downregulated (bottom) DEGs, ranked by FDR, in ALS cases versus controls by male and female. Legend represents gene expression by log(counts per million; CMP) from most upregulated to most downregulated in ALS. Overlap of (**d**) upregulated and (**e**) downregulated DEGs from our dataset with the Swindell et al.[20] 2019 and Grima et al.[21] 2023 datasets. The diagonal hatched area represents overlap consistent DEGs; the circle area is proportional to the approximate DEG numbers. Characteristics of unique and consistent DEGs in (**f**) average $\log_2$ gene expression (AvExpr), (**g**) p-value, and (**h**) LFC. All figure p-values are unadjusted for multiple comparisons.

DEGs to published DEGs (Supplementary Fig. S2a). DEG features did not differ whether they were up- or downregulated (Supplementary Fig. S2b–d). Overlapping DEGs were less likely to differ in effect size by sex (Supplementary Fig. S2e). Finally, to confirm our RNA-seq results, we selected a new, independent cohort of 29 ALS and 27 control participants (Table 2) and performed whole blood qPCR of 5 genes upregulated in ALS, chosen based on altered expression and biological relevance. Fold-changes in gene expression in ALS versus controls by qPCR were similar to fold-changes by RNA-seq (*B2M*, *CAPZA1*, *RPS18*, *TNFSF10*, *TPT1*; Supplementary Fig. S3). Overall, qPCR verified gene expression changes in ALS relative to controls by RNA-seq, validating our transcriptomic dataset.

A gene expression signature predicts ALS case status

Since ALS ($n = 422$) and control ($n = 272$) transcriptomes differed substantially, we next evaluated how well gene expression could predict case-control status (Fig. 1c, i). We compared the performance of seven machine learning algorithms, ridge, LASSO, elastic net, L1/2, SCAD, MCP, and XGBoost[29], trained on gene expression data from ALS cases ($n = 296$) and controls ($n = 191$) and tested using receiver operating characteristic (ROC) curve AUCs (ALS cases, $n = 126$; controls, $n = 81$). XGBoost had the significantly highest detection accuracy (DeLong's test between XGBoost and elastic net: $Z = 2.62$, $p = 0.0087$, ΔAUC 0.044 [95%CI: 0.011, 0.076]), with the largest AUC of 0.91 and, thus, was chosen for the next step (Supplementary

https://doi.org/10.1038/s41467-025-64622-5

Table 2 | Clinical characteristics of ALS and control participants in validation cohort

	ALS cases, $n = 29$	Controls, $n = 27$	p-value
Age at blood draw (years)	68.78 (62.87, 74.93)	68.52 (62.97, 74.71)	0.86
Sex			1
Female	11 (38%)	11 (41%)	
Male	18 (62%)	16 (59%)	
Race			0.38
White	27 (93%)	27 (100%)	
Black	1 (3%)	0 (0%)	
Other	1 (3%)	0 (0%)	
Family history of ALS			0.5
Yes	2 (7%)	0 (0%)	
No	27 (93%)	27 (100%)	
Revised El Escorial criteria at diagnosis			
Definite	15 (52%)		
Probable	9 (31%)		
Probable, lab supported	5 (17%)		
Possible	0 (0%)		
Suspected	0 (0%)		
Onset segment			
Bulbar	10 (34%)		
Cervical	10 (34%)		
Lumbar	9 (31%)		
Respiratory	0 (0%)		
Thoracic	0 (0%)		
Generalized / Cannot be determined	0 (0%)		
***C9orf72* hexanucleotide repeat expansion status**			5.37×10^{-133}
Negative	29 (100%)	0 (0%)	
Intermediate / unknown	0 (0%)	27 (100%)	
Positive	0 (0%)	0 (0%)	
ALSFRS-R Score	39.00(35.00, 43.00)		
Time from diagnosis to blood draw (years)	1.78 (1.49, 2.41)		
Time from symptom onset to diagnosis (years)	1.00 (1.00, 2.00)		
Unknown	1		

Median(25%,75%); *n* (%); two-sided Wilcoxon rank sum exact test; Pearson's Chi-squared test.

Fig. S4), although other algorithms also achieved good prediction accuracy, e.g., elastic net (AUC 0.87), LASSO (AUC 0.85), and ridge (AUC 0.84).

Next, we refined XGBoost, by first identifying input DEGs with false discovery rate (FDR) and average $\log_2$ of gene expression (AvExpr) that best discriminate ALS cases from controls using our full training cohort. The ultimate goal was to assess the predictive ability of XGBoost on a limited set of DEGs to assign case-control status on an independent, external testing cohort. We retrained XGBoost on three groups of DEGs from our entire cohort (ALS cases, $n = 422$; controls, $n = 272$) (Fig. 3a) filtered using the criteria FDR < 0.01 (3,640 DEGs), FDR < 0.01 with AvExpr>0 (3,261 DEGs), and FDR < 0.01 with AvExpr>2 (2,621 DEGs), yielding 27- (Fig. 3b), 30- (Fig. 3c), and 29-gene (Fig. 3d) panels, respectively. As anticipated, DEGs that contributed strongly to ALS case-control distinction were among the chosen top 20 DEGs (Fig. 1c). Fourteen DEGs were shared by all gene panels, with fewer DEGs shared by two panels, and some unique to each (Fig. 3e). We also merged all three panels into a "combined" 46-DEG panel.

AUCs from internal testing ranged from 0.969 to 0.972, with sensitivities from 93.2 to 94.2%, specificities from 86.0 to 87.9%, and accuracies of 91.1 to 91.2%. The XGBoost classifier continued to perform well in the external Grima et al.[21] test cohort, yielding an AUC of 0.894 from the combined gene panel (Fig. 3f, g). Additionally, we externally evaluated our classifier in the van Rheenan et al.[19] /Swindell et al.[20] cohort, attaining AUCs of 0.654 to 0.747 despite the microarray nature of the dataset, which also lacked 12 out of 46 gene features from our XGBoost models (Supplementary Fig. S5). Therefore, our classifier showed potential as a diagnostic biomarker panel in this proof-of-concept test.

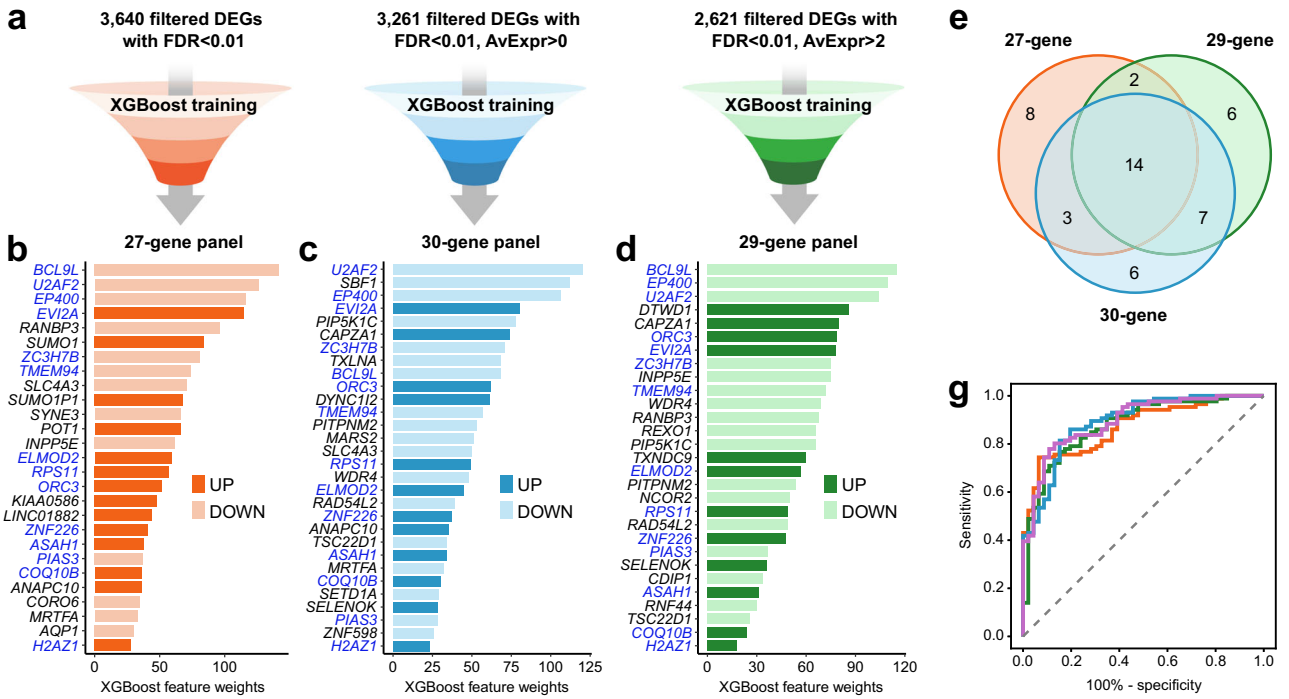
A gene expression signature predicts ALS survival

Besides aiding diagnosis, biomarkers can also inform prognosis (Fig. 1c, ii). We next appraised the ability of our models to predict ALS survival, focusing on stepwise and XGBoost versus a "clinical" only model of onset segment, symptom onset age, and sex (Fig. 4a). We trained models using all our ALS cases with available survival data ($n = 420$; Supplementary Data 3) and input of clinical variables and 575 genes filtered using stringent criteria [log-rank p-value < 0.01, average count per million mapped reads >10, protein-coding genes only]. In addition to all three clinical variables, stepwise and XGBoost selected 18 (Fig. 4b) and 8 (Fig. 4c) gene features, respectively, which had among the highest and lowest hazard ratios (Supplementary Data 4). However, onset segment and symptom onset age contributed most significantly to stepwise and XGBoost, respectively.

Next, we predicted survival in thirty random train-test splits of our internal dataset; compared to the clinical variables only model, stepwise had significantly larger AUC values for years 2 to 8, whereas XGBoost had higher AUCs for years 4 to 8 (Fig. 4d, Supplementary Fig. S6a, Supplementary Data 5). In fact, gene only models using the 18 stepwise- and 8 XGBoost-selected gene features significantly contributed to survival prediction independent of the clinical variables alone (Supplementary Fig. S6b). Our stepwise and XGBoost models similarly generated numerically larger AUC values for predicting 4-year survival in the external Grima et al.[21] dataset versus the clinical variables model (Fig. 4e, Supplementary Fig. S7). Comparing concordance index (C-index) between the models, a common approach to compare how well models predict the order of events, we found that XGBoost (C-index=0.69) performed significantly better than the clinical variables only model (C-index=0.66; $Z = 1.96$, $p = 0.050$, $\Delta C = 0.033$ [95% CI: 8.12×10^{-5}, 6.51×10^{-2}]), however the stepwise model performance (C-index=0.65) did not significantly differ from the clinical variables only model ($Z = -0.26$, $p = 0.79$, $\Delta C = -8.90 \times 10^{-3}$ [95%CI: −0.076, 0.058]) (Supplementary Fig. S6c).

Subsequently, we classified the 86 Grima cases into predicted shorter (top 25%), intermediate, and longer (bottom 25%) survival based on median predicted survival score, calculated using clinical variables only (clinical variables model) or combined clinical variables and gene features (stepwise and XGBoost models) (Supplementary Fig. S8). The clinical variables model predicted 1.11-year and 2.74-year median survival differences between the predicted shorter- versus intermediate- and intermediate- versus longer-surviving participants, respectively (Fig. 4f). Stepwise predicted 1.31-year and 2.46-year median survival differences between the predicted shorter- versus intermediate- and intermediate- versus longer-surviving participants, respectively (Fig. 4g), and performed the best at distinguishing the three survival groups earlier in the disease. XGBoost predicted 1.31-year and 3.20-year median survival differences between the predicted shorter- versus intermediate- and intermediate- versus longer-surviving participants, respectively (Fig. 4h). Overall, the XGBoost model outperformed the clinical variables only model for differentiating among survival groups in both the internal and external

https://doi.org/10.1038/s41467-025-64622-5



f

Internal training 10-fold cross-validation metrics (ALS n=422, controls n=272)					Metrics from testing in external Grima dataset (ALS n=86, controls n=46)						
Panel	Accuracy	Sensitivity	Specificity	ROC-AUC	Accuracy	Sensitivity	Specificity	Precision	ROC-AUC	Balanced accuracy	F1
27	91.07%	94.19%	85.98%	96.86%	75.00%	77.91%	69.57%	86.59%	87.23%	73.74%	84.52%
30	91.06%	93.34%	87.25%	97.00%	84.09%	86.05%	80.43%	88.37%	89.21%	83.24%	88.37%
29	91.22%	93.20%	87.91%	97.25%	80.30%	82.56%	76.09%	88.10%	87.69%	79.33%	87.06%
46	*	*	*	*	81.06%	82.56%	78.26%	87.65%	89.43%	80.41%	85.03%

External ROC-AUCs
27-gene, ROC-AUC 0.872
30-gene, ROC-AUC 0.892
29-gene, ROC-AUC 0.877
Combined, ROC-AUC 0.894

Fig. 3 | A gene expression signature predicts ALS case status. a XGBoost was trained on filtered DEGs in ALS cases ($n = 422$) versus controls ($n = 272$) using the criteria FDR < 0.01 (3,640 DEGs), FDR < 0.01 with AvExpr>0 (3,261 DEGs), and FDR < 0.01 with AvExpr>2 (2,621 DEGs), yielding 27-, 30-, and 29-gene panels, respectively. Funnels created in BioRender. Feldman, E. (2025) https://BioRender.com/56j577n. Candidate genes ranked by their importance to ALS case prediction for the (**b**) 27-gene, (**c**) 30-gene, and (**d**) 29-gene panels. **e** Overlap in genes between the three gene panels, including 14 shared by all models, presented by blue font in panels **b** to **d**. **f** Performance metrics for predicting case status by the three gene panels plus the combined panel on our internal dataset and the external Grima et al.[21] dataset. **g** Area under the receiver operating characteristic curves (ROC-AUC) for predicting case status by the three gene panels plus the combined panel on the external Grima dataset.

Grima et al. datasets (Supplementary Fig. S6a,c). This wider difference, reflecting a stronger overall distinction in survival classification by XGBoost, incorporated gene features, versus the clinical variables only model. Thus, although there was substantial overlap between the models in assigning cases as shorter- (Fig. 4i) or longer-surviving (Fig. 4j), there were differences, which would have important implications for patients.

Pathway analysis of the blood transcriptome reveals an ALS disease signature

After demonstrating the diagnostic and prognostic potential of our ALS blood transcriptomics dataset, we next performed pathway enrichment to identify disease-related pathways with the goal of creating an ALS disease signature observable in blood (Fig. 1c, iii). This analysis leveraged the dataset of all ALS cases ($n = 422$) and controls ($n = 272$). ALS is characterized by altered immune cell levels[30] and activation state[31]. Therefore, in addition to pathway enrichment of the primary whole blood transcriptomic analysis, we also performed enrichment of the transcriptomics dataset adjusted for cell type proportions, i.e., for the relative abundance of different immune cell types in the blood sample. This method extracts intrinsic, true biologically relevant pathways, independent of altered cell levels in ALS. Several methods exist to computationally adjust cell proportions, using DNA methylation (DNAm) or RNA expression data, which can be compared against experimentally measured cell proportions, e.g., flow cytometry. In addition to transcriptomics data from this study, DNAm[32] and flow cytometry[24–26] datasets were previously collected for this deeply phenotyped cohort. We computed the different proportions of blood cell types by DNAm (Supplementary Fig. S9a) and RNA expression (Supplementary Fig. S9b) data, which we correlated to experimentally measured cell proportions by flow cytometry for CD8 + T cells, CD4 + T cells, NK cells, monocytes, and neutrophils. Correlations to flow cytometry-measured cell proportions were stronger and more significant when computed by DNAm than by RNA expression (Supplementary Fig. S9c,d), making DNAm the adopted method.

Pathway enrichment using Kyoto Encyclopedia of Genes and Genomes (KEGG) and Hallmark terms of the primary whole blood transcriptomic analysis identified several enriched upregulated neurodegenerative pathways in ALS females versus control, including "amyotrophic lateral sclerosis", "Huntington disease", and "Parkinson disease." When adjusted for cell proportions, these pathways were more enriched and also appeared in ALS males (Fig. 5a, Supplementary Data 6). "Amyotrophic lateral sclerosis" was among the top upregulated pathway in both ALS males (FDR = 8.27×10^{-3}) and females (FDR = 6.90e-10), suggesting that blood gene expression contains ALS-relevant signatures. Other ALS-related pathways followed a similar pattern, e.g., "oxidative phosphorylation", "thermogenesis"[33], and "proteasome". ALS-related "nucleocytoplasmic transport", upregulated

https://doi.org/10.1038/s41467-025-64622-5

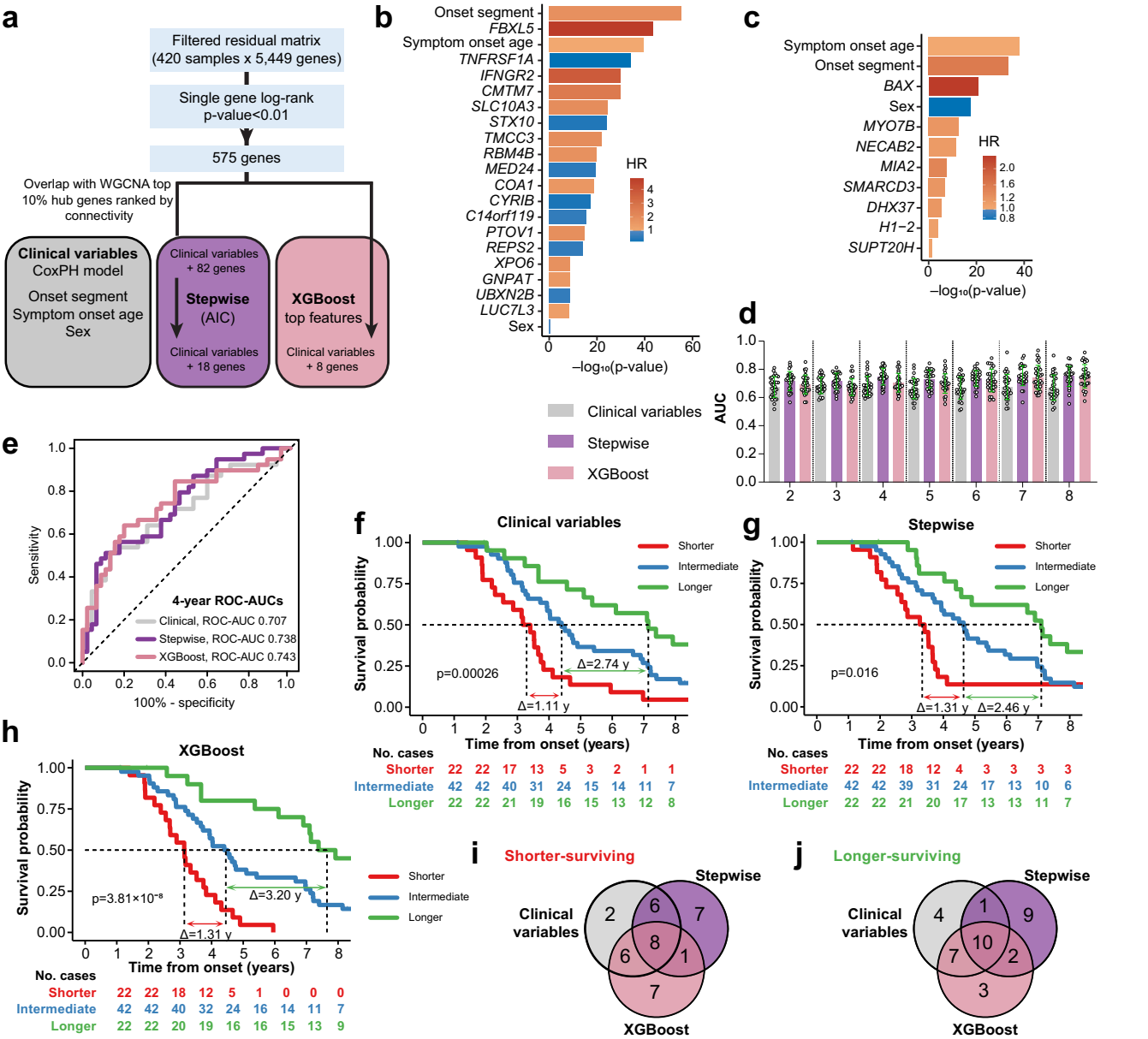


Fig. 4 | A gene expression signature may predict ALS survival ($n = 420$). **a** Workflow for training and validating machine learning models for predicting ALS survival; CoxPH, Cox proportional hazard; WGCNA, weighted gene co-expression network analysis. Gene features from the **b** stepwise and **c** XGBoost model, ranked by $-\log_{10}$(p-value), with color-coded hazard ratio (HR) (Supplementary Data 4). **d** Area under the time-dependent receiver operating characteristic curves (AUCs) represented by mean ± standard deviations for clinical variables, stepwise, and XGBoost models for predicting survival ($n = 30$ random train-test split in our internal dataset). Results reported in Supplementary Data 5. **e** Area under the receiver operating characteristic curves (ROC-AUC) for clinical variables, stepwise, and XGBoost models for predicting the most clinically relevant 4-year survival in the external Grima et al.[21] dataset. Kaplan-Meier curves for predicted shorter- versus intermediate- versus longer-surviving participants according to the **f** clinical variable, **g** stepwise, and **h** XGBoost models in the external Grima dataset. Δ represents the difference in median survival (years) between shorter- versus longer-surviving participants. Δ represents the difference in median survival (years) between shorter- versus intermediate- or intermediate- versus longer-surviving participants. Consistency across different models for assigning **i** shorter- and **j** longer-surviving participants in the external Grima dataset. All figure p-values are unadjusted for multiple comparisons.

in ALS females, newly appeared after adjusting DEGs for cell proportions, along with, interestingly, many immune and infection pathways, e.g., "human cytomegalovirus infection", "shigellosis" (Fig. 5b). Several immune and infection pathways, e.g., "complement", "interferon gamma response", disappeared after adjusting for cell proportions (Fig. 5c).

Gene ontology (GO) pathway analysis of the DEGs from whole blood revealed enriched upregulated pathways linked to RNA processing and splicing, and energy production through the electron transport chain, both relevant to ALS pathophysiology (Supplementary Fig. S10, Supplementary Data 7). Adjusting for cell proportions eliminated a few pathways, e.g., ion transport, glucose catabolic processes (Supplementary Fig. S10b), while revealing new ones, e.g., nucleotide and nucleoside synthesis and metabolism, protein catabolic processes (Supplementary Fig. S10c). As expected, adjusting for cell proportions decreased enrichment of immune cell markers, most notably for neutrophils and monocytes (Supplementary Fig. S10d).

We also examined pathway enrichment of primary DEGs and DEGs adjusted for cell proportions by disease severity by analyzing the lowest (most functionally impaired) versus the highest (least functionally impaired) quartile of ALSFRS-R scores. Overall, we found

https://doi.org/10.1038/s41467-025-64622-5

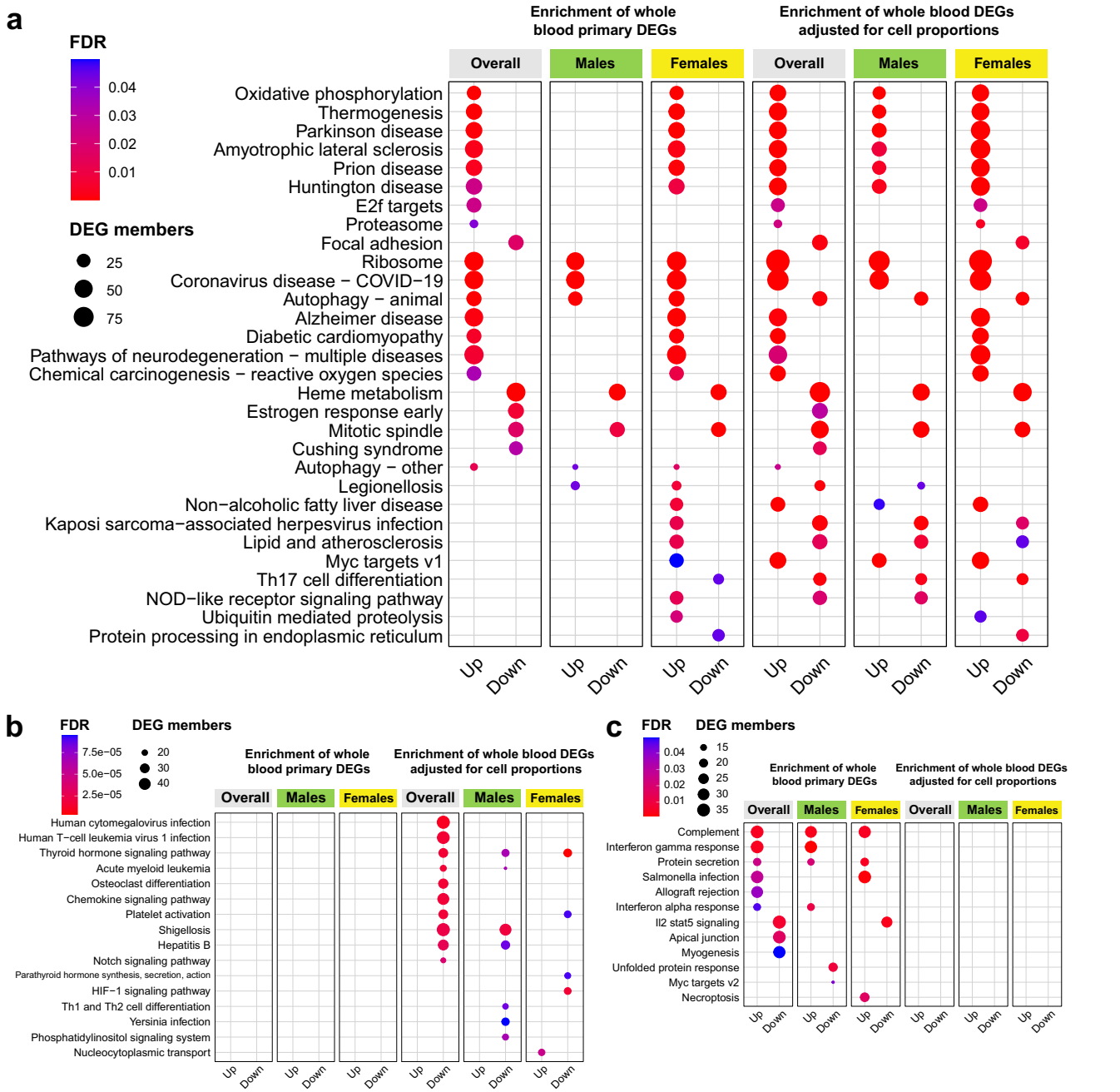


Fig. 5 | Pathway analysis of the blood transcriptome reveals ALS-related pathways. a Enriched KEGG and Hallmark pathways in ALS cases ($n = 422$) versus controls ($n = 272$) shared after pathway analysis from primary DEGs (left) and DEGs adjusted (right) for cell proportions, shown overall, in males, and in females. Up- and downregulated pathways are annotated at the bottom of the dot plot. **b** Enriched KEGG and Hallmark pathways in ALS cases versus controls that appeared after pathway analysis from DEGs adjusted for cell proportions. **c** Enriched KEGG and Hallmark pathways in ALS cases versus controls that disappeared after pathway analysis from DEGs adjusted for cell proportions. Dot color represents significance according to FDR, dot size represents the number of DEGs selected by KEGG and Hallmark pathways following enrichment analysis.

pathways linked to more severe disease reflected in lower ALSFRS-R scores were related to neurodegeneration, immune dysregulation, and metabolic stress (Supplementary Fig. S11), key ALS pathophysiological processes.

Overall, enrichment analysis of the whole blood transcriptome revealed ALS-relevant pathways beyond immune-related pathways. Adjusting for cell proportions in blood further enriched these ALS pathways and confirmed the presence of an ALS disease signature in blood. Thus, we next sought to leverage the blood transcriptomic dataset to identify potential drug candidates by drug perturbation analysis.

Drug perturbation analysis of the blood transcriptome reveals ALS therapeutic candidates

We launched drug perturbation analysis by first identifying "core genes" most strongly linked to the disease process in relevant tissues (Fig. 1c, iv). We identified early "core genes" by examining DEG overlap between adjusted and primary analysis of our blood transcriptomic dataset with induced pluripotent stem cell (iPSC)-derived neurons with TDP-43 knockdown, an established in vitro model of ALS (iPSC-neurons hereon)[34], modeling the disease process. We found 472 and 473 DEGs from adjusted and primary blood datasets, respectively, overlapped with iPSC-neurons (Fig. 6a); these overlapping DEGs were not the most

https://doi.org/10.1038/s41467-025-64622-5

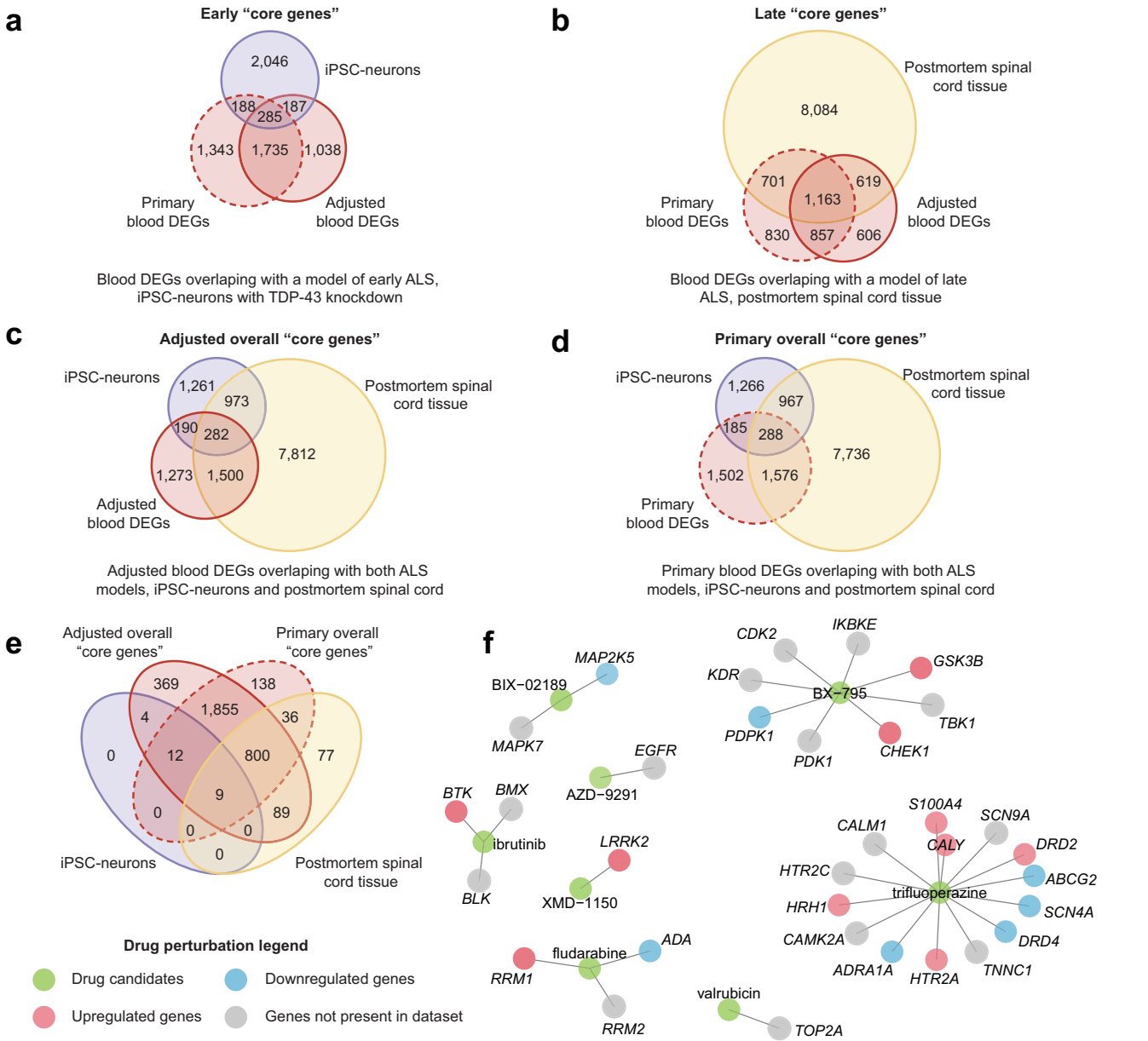


Fig. 6 | Drug perturbation analysis of the blood transcriptome reveals ALS therapeutic candidates. Venn diagram of our blood primary DEGs and DEGs adjusted for cell proportions in ALS cases ($n = 422$) versus controls ($n = 272$), shared with (**a**) iPSC-neuron and (**b**) postmortem spinal cord DEGs. Venn diagram of our blood **c** DEGs adjusted for cell proportions and **d** primary DEGs, shared with iPSC-neuron and postmortem spinal cord DEGs. **e** Venn diagram of overlap in candidates identified by drug perturbation of DEGs from the "core genes" (adjusted), "core genes" (primary), iPSC-neurons, and postmortem spinal cord tissue. **f** Gene-drug network of the 8 candidates selected by drug perturbation analysis that reversed gene expression of the overall disease course, shared across "core genes" (adjusted, primary), iPSC-neurons, and postmortem spinal cord.

significant or with the largest fold-change in the iPSC-neuron dataset (Supplementary Fig. S12a). We found 1782 and 1864 DEGs in adjusted and primary blood datasets, respectively, overlapped with postmortem spinal cord[35], relevant to end-stage ALS (Fig. 6b, Supplementary Fig. S12b), which, again, were not among the most significant or with the largest fold-changes. Finally, we identified overall "core genes" between blood, iPSC-neurons, and spinal cord, finding 282 and 288 DEGs in our adjusted and primary datasets (Fig. 6c, d).

Equipped with the overall "core genes", we next performed drug perturbation analysis using the LINCS database, a collection of 473,647 unique transcriptomic signatures across diverse cell types resulting from 42,080 "perturbagens", i.e., potential drug candidates. Our input was the overall "core genes", along with 150 up- and 150 down-regulated DEGs from iPSC-neurons and spinal cord, with the greatest fold-changes. Drug perturbation analysis identified the most, 3138 signatures, for cell proportions-adjusted "core genes", followed by 2850 in primary "core genes" (Supplementary Data 8).

We focused on candidates that overlapped between overall "core genes" (adjusted and primary) with both iPSC-neurons and spinal cord, of which eight out of nine reversed gene expression in the overall disease course (Fig. 6e). These potential candidates spanned diverse mechanisms of action, mutant EGFR inhibitor AZD-9291, MEK5/ERK5 inhibitor BIX-02189, multi-kinase inhibitor, including TANK-binding kinase 1 (TBK1) BX-795, DNA synthesis inhibitor fludarabine, Bruton's tyrosine kinase (BTK) inhibitor ibrutinib, anti-adrenergic and anti-dopaminergic trifluoperazine, semisynthetic doxorubicin analog valrubicin, and autophagy targeting XMD−1150[36]. These candidates interacted with several genes, which, with few exceptions, e.g., *TBK1*,

interestingly, do not have well-established roles in ALS (Fig. 6f). Thus, in addition to potential therapeutic candidates, the analysis pinpointed new potentially disease-related genes.

Discussion

ALS remains difficult to recognize and lacks a disease-specific biomarker, leading to diagnostic delays[4]. Moreover, patients often lack accurate information regarding their personalized anticipated disease course. Generally, however, most ALS patients face a survival of only 2 to 4 years due to the lack of effective disease-modifying treatments[22]. In the current study, we sought to address these unmet needs by developing blood-based gene expression signatures of ALS risk and survival along with therapeutic drug candidates using drug perturbation analysis.

First, we capitalized on altered whole blood gene expression in ALS with our overarching goal of developing diagnostic or clinical status biomarkers. We identified 3640 DEGs in ALS cases versus controls, confirming ALS-related altered gene expression in blood. As anticipated, many DEGs were linked to the immune system, and there were some, albeit primarily minor, male versus female differences, possibly linked to sex differences in the immune system in ALS[24–26]. Top DEGs clearly differentiated ALS cases from controls, many related to ALS pathophysiology, spanning vesicular and endosomal trafficking, autophagy, epigenetics, inflammation, muscle, and apoptosis. Our DEGs substantially overlapped with 35 to 50% of published blood transcriptomic datasets[19,20]; nevertheless, we had over 2800 unique DEGs, likely resulting from detailed RNA-seq analysis of our large ALS cohort. We then leveraged this highly detailed RNA-seq dataset to train several classifiers. XGBoost, the best performing classifier, accurately predicted case-control status with an AUC of 0.91.

We next refined XGBoost with more stringent criteria for input DEGs, generating 27-, 29-, and 30-DEG gene panels and a combined 46-DEG panel, all amenable to a future commercial multiplex-format PCR array as an ALS biomarker panel. All panels performed very well, with AUCs of 0.969 to 0.972, sensitivities of 93.2 to 94.2%, specificities of 86.0 to 87.9%, and accuracies of 91.1 to 91.2%. We validated our XGBoost classifier externally on the Grima dataset[21], with AUCs 0.872 to 0.894, underscoring proof-of-concept evidence as a potential diagnostic tool.

Accurate prediction of case-control status in a fully independent external testing cohort[21] is a long-standing desired goal for eventual clinical diagnosis. Pioneering work by Saris et al. first detected ALS-specific gene expression profiles detected in blood[37], prompting use for ALS case-control classification. van Rheenen et al., from blood-based gene microarray profiles, developed a classifier with an AUC of 0.90 from an internal test, which was not externally validated[19]. Re-analysis of the dataset differentiated ALS cases from ALS mimics and controls with 87% accuracy, 86% sensitivity, and 87% specificity[20], but still was not externally validated. Grima et al.[21], based on expression of 20 genes, generated a classifier that distinguished sporadic ALS cases from controls internally with 0.829 AUC, 78% accuracy, 79% sensitivity, and 75% specificity[21]. However, in an independent external dataset, performance dropped significantly to 0.647 AUC, 63% accuracy, 60% sensitivity, and 67% specificity. Thus, to our knowledge, our classifier outperforms externally compared to any literature reports and should be further investigated as a potential ALS biomarker panel to improve diagnostic accuracy and decrease diagnostic delay.

Secondly, we combined gene expression features with clinical variables to predict overall survival, both internally and externally, in a fully independent test cohort[21]. Including gene features improved AUCs and, in the external dataset[21], better differentiated shorter-, intermediate-, and longer-surviving cases by median survival. Moreover, there was substantial overlap in cases classified as shorter-, intermediate-, or longer-surviving between the gene-incorporating stepwise and XGBoost models versus the clinical variables only models. However, there were some differences, which would hold important ramifications in a clinical scenario for patients regarding their personal anticipated disease course. Currently, ALS prognosis is based on disease progression, monitored clinically by ALSFRS-R, including via its domain for respiratory function[3]. New scoring, scaling, and staging tools have become available, but none are in widespread clinical use[3,38]. In parallel, no biomarker has been clinically developed for ALS prognosis, although baseline serum NfL level can correlate with progression[39] and has been adopted as a secondary outcome in several recent clinical trials[40–42]. Reanalysis of the van Rheenen et al.[19] dataset by Swindell et al. identified a 61-gene blood expression signature that could predict survival with a mean concordance index of 0.60, which increased to 0.74 when they included covariates, specifically onset segment, onset age, sex, and batch[20]; however, this 61-gene signature was not externally validated. Therefore, our externally validated survival prediction models that incorporate blood gene expression features with clinical criteria are unique in the literature.

Our last goal was to uncover ALS disease-relevant pathways in the blood transcriptome and identify potential candidates by drug perturbation analysis. Although blood is not the primary affected tissue in ALS, enrichment analysis detected several upregulated KEGG terms for neurodegenerative pathways, including “amyotrophic lateral sclerosis”, along with pathways related to ALS pathophysiology, such as “oxidative phosphorylation”, “thermogenesis”[33], and “proteasome”. Although the “amyotrophic lateral sclerosis” pathway also included genes related to mitochondrial energy production, proteasome, and nucleocytoplasmic transport, which enriched individually as pathways, it also contained ALS genes, including *C9orf72*, *TBK1*, and *CHMP2B*. Therefore, ALS-causing DEGs were also represented in the pathway enrichment of our blood transcriptomic dataset. Adjusting for cell proportions in blood further enhanced the ALS-relevant pathways. As expected, several immune pathways were represented, independent of cell proportion adjustments, while some disappeared and newly appeared following adjustment. Our GO enrichment analysis also highlighted ALS-relevant pathways, such as RNA processing and splicing. Our results are aligned with van Rheenen et al., who identified dysregulated RNA binding, intracellular transport, and protein transport and localization as most characteristic of ALS[19]. Swindell et al. similarly found altered RNA metabolism, along with immune system pathways[20], overlapping with Grima et al., reporting altered metabolic, transcriptional regulation, immune response, and apoptotic pathways[21]. Saris et al. identified mitochondrial dysfunction, related to oxidative phosphorylation, as well as neurodegenerative and mRNA processing pathways in blood from ALS cases versus controls[37].

Our pathway analysis demonstrated that comprehensive ALS signals are present in the blood of individuals with ALS. Indeed, our whole blood DEGs in ALS cases versus controls overlapped with DEGs from ALS iPSC-neurons and postmortem spinal cord. These “core genes” underscore key shared features between blood and the primary disease tissue, reflected by enrichment of ALS-relevant pathways in our blood gene expression dataset. We leveraged these “core genes” for a drug perturbation analysis and discovered eight drugs that overlapped with drug perturbation of iPSC-neurons and postmortem spinal cord DEGs and reversed the disease transcriptomic signature.

Among the candidates was an FDA-approved drug, the phenothiazine trifluoperazine, previously identified by a drug repurposing effort in ALS[43]. Trifluoperazine is a typical antipsychotic primarily used to treat schizophrenia, a neuropsychiatric disorder that may overlap with ALS[44]. While there are reports that use of antipsychotics may lower the risk of ALS[45] and other age-related neurodegenerative diseases[46], some studies report no association[47]. Trifluoperazine is also a potent allosteric modulator of the human purinergic P2X7 receptor[48]. P2X7 receptor antagonism has been proposed as an ALS therapeutic strategy by modulating neuroinflammation[49].

A second identified candidate was an FDA-approved irreversible BTK inhibitor, ibrutinib, that blocks B-cell proliferation and survival and is reported to delay symptom onset, reduce muscular atrophy, and decrease pro-inflammatory brain cytokine production in *SOD1*^{G93A} mice[50]. Bosutinib, another tyrosine kinase inhibitor approved for chronic myelogenous leukemia, was repurposed for ALS based on an iPSC screen[51]. A small open-label, dose-escalation phase I trial found the drug was safe and potentially effective in an ALS participant subset[52]. Overall, several kinase inhibitors have or are being considered for ALS[53]. Our drug perturbation also selected BX-795, an inhibitor of TBK1 kinase, a known ALS gene linked to innate immunity, autophagy, and cell cycle[54]. A recent Mendelian randomization of ALS genome-wide association studies proposed TBK1 as a drug repurposing target[55].

Some candidates had no known literature linked to ALS, including AZD-9291, BIX-02189, fludarabine, valrubicin, and XMD − 1150[36], suggesting future possible research avenues. Indeed, XMD-1150 may potentially stimulate autophagy[36], impaired in ALS[54], and inhibit LRRK2 (leucine-rich repeat kinase 2)[56], also linked to autophagy in neurodegeneration, especially Parkinson's disease[57]. Overall, our drug perturbation analysis promotes a strategy for identifying therapeutic candidates and supports further investigation into three drugs previously suggested as potential ALS therapies[58,59].

Our study had several weaknesses. First, we did not include ALS mimics or presymptomatic ALS mutation carriers. In a clinical scenario, a diagnostic tool would need to differentiate patients with ALS from patients without ALS that manifest similar symptoms, i.e., ALS mimics. Moreover, a diagnostic tool capable of identifying imminent phenoconversion to ALS in mutation carriers is lacking; however, our study did not include presymptomatic mutation carriers. Second, we faced the tissue issue, relying on blood as the most accessible sample; however, blood harbors ALS-related changes in immune system dysfunction[24–26,60]. To overcome this weakness, we adjusted our blood transcriptomic analysis for cell proportions, which enhanced enrichment of multiple ALS-relevant disease pathways, validating blood as a viable biofluid. Third, the clinical criteria used in our prediction model relied on onset segment, symptom onset age, and sex; however, this list may not necessarily comprise all clinical variables that affect ALS survival[38]. Moreover, AUCs for models that incorporated gene features were only numerically higher than AUCs for the clinical variables only model; however, gene feature-based models maximized median survival differences, underscoring the added benefit of gene expression data. Fourth, our drug perturbation analysis was based on overlapping DEGs between our blood gene expression dataset with ALS iPSC-neurons and spinal cord, covering only a fraction of the dysregulated ALS transcriptome, so selected drugs may not necessarily modify the disease process. Additionally, drug perturbation analysis was conducted on averaged perturbagen-induced gene expression changes across all database cell types, which may not reflect transcriptomic changes the perturbagen would induce in blood or spinal cord tissue. Finally, we had an imbalance in sex and RNA processing methods in our cases versus controls, but we adjusted for these variables in our analyses.

Despite limitations, our study also had numerous strengths. First, we had a large sample size, especially of ALS cases, the largest to date in blood transcriptomics. Our large sample size enabled sex-specific analyses, relevant to sex differences in ALS[24–26]. Second, we used high-coverage RNA-seq, and detected 22,332 genes encompassing long non-coding RNAs and microRNAs, as well as protein-coding genes. This contrasts with the published microarray data of only 9822 protein-coding genes[20]. Third, we used stringent criteria to identify DEGs, relying on FDR < 0.01 and absolute $\log_2$(fold-change)>0.1, and filtered DEGs by rigorous criteria as input to train our classifiers. Fourth, we externally validated our case-control classifiers and our survival prediction models, demonstrating the robustness of our findings. Finally, we leveraged immune phenotyping and DNAm data to adjust our blood transcriptomics data for specific subpopulations of blood cells, revealing intrinsic disease pathways and facilitating drug perturbation analysis.

In conclusion, there is a real clinical need for better diagnostic and prognostic tools in ALS to improve patient care. The breast cancer PAM50 gene panel demonstrates the feasibility of gene expression data for disease subtype classification and prognosis[61]. Although no such gene expression panel is available for ALS diagnosis and/or prognosis, our findings strongly support the ability of gene expression profiles from whole blood to differentiate ALS cases from controls, to potentially predict survival, and to infer, by pathway enrichment, relevant disease pathways and identify potential drug candidates. Towards this goal, we developed an accurate, externally validated blood-based gene expression signature panel for ALS classification, underscoring potential clinical diagnostic use as a biomarker. We also combined gene expression features with clinical variables for survival prediction, highlighting potential prognostic utility. Now that our study has shown blood transcriptomics can accurately predict case-control status in external cohorts, we can begin to assess potential utility for classification versus ALS mimics and in presymptomatic ALS mutation carriers prior to phenoconversion. We envision a possible translation to a clinically viable platform that could accelerate ALS diagnosis and provide patients, based on their personalized gene expression profiles, with anticipated survival times.

Methods

Ethics statement

This study was conducted according to all relevant ethical regulations; all participants were over 18 years of age and provided informed consent in English. The University of Michigan Institutional Review Board (IRBMED HUM28826) granted ethical approval for this research.

Participants

The study recruited participants with ALS ($n = 422$) meeting the Gold Coast definition of ALS during their clinical visit at the University of Michigan Pranger ALS Clinic between 2011 and 2021. Control participants ($n = 272$) without ALS and without first- or second-degree relatives with ALS were identified and recruited via a University of Michigan research interest database, random address direct mailings, and Meta/Facebook advertisements, and were compensated USD $50 via a check in the mail[62–64]. ALS participants did not receive monetary compensation. Demographic information was collected from ALS and control participants and disease characteristics were collected from ALS participants, including age at onset and diagnosis, disease onset segment, revised El Escorial criteria, and ALSFRS-R score (Table 1). Gold Coast criteria were used to group participants as cases for classification purposes, while the revised El Escorial criteria were collected as part of their diagnostic workup. Possible or suspected El Escorial criteria at diagnosis did not affect case status prediction relative to definite, probable, or probable, laboratory supported criteria (Supplementary Fig. S13).

Sample collection and RNA-seq

Blood samples for RNA-seq were collected from participants with ALS and controls into PAXgene tubes (catalog no. 762165, Qiagen, Germantown, MD), per manufacturer instructions. Blood samples were frozen and stored at −80 °C until RNA extraction. RNA was extracted from the PAXgene tubes using the PAXgene Blood miRNA Kit (catalog no. 763134, Qiagen). RNA quality was evaluated by RNA integrity number (RIN) using TapeStation (Agilent, Santa Clara, CA) and Qubit RNA broad-range assay (catalog no. Q10211, Thermo Fisher Scientific, Waltham, MA). Depending on the RIN, total RNA samples were further processed either by ribosomal RNA (rRNA) depletion or messenger RNA (mRNA) selection.

rRNA depletion from samples. When total RNA samples (90 ng) met the criteria RI*n* < 5.5 or DV200 in the range of 50 to ~75%, they were depleted of rRNA using the NEBNext Globin & rRNA Depletion Kit (Human/Mouse/Rat) (catalog no. E7750X, New England Biolabs, Ipswich, MA). This was the protocol adopted for 261 ALS and 126 control samples. rRNA-depleted RNA samples were then fragmented for 5–10 min based on RIN of the input RNA. Fragments were then copied into first-strand cDNA by reverse transcription with random primers, and 3′ ends were adenylated and ligated to adapters. The products were amplified by 16 PCR cycles to generate the final cDNA library. Library preparation was performed using xGen Broad-Range RNA Library Preparation Kit (catalog no. 10010145, Integrated DNA Technologies, Coralville, IA) and xGen Normalase UDI Primers (catalog no. 10009795, 10009800, 10009811, 10009812, Integrated DNA Technologies). RNA-seq analysis adjusted for RNA processing method.

polyA selection of mRNA from samples. When total RNA samples (90 ng) met the criteria RI*n* ≥ 5.5 and DV200 > 75%, mRNA was isolated by polyA purification using the NEBNext Poly**a** mRNA Magnetic Isolation Module (catalog no. E7490L, New England Biolabs). This was the protocol adopted for 161 ALS and 146 control samples. mRNA-purified samples were then fragmented and copied into first-strand cDNA by reverse transcription with random primers, and 3′ ends were adenylated and ligated to adapters. The products were amplified by 16 PCR cycles to generate the final cDNA library. Library preparation was performed using xGen Broad-range RNA Library Prep (catalog no. 10010145, Integrated DNA Technologies), and xGen Normalase UDI Primers (catalog no. 10009795, 10009800, 10009811, 10009812, Integrated DNA Technologies). RNA-seq analysis adjusted for RNA processing method.

The quality of all final libraries, both rRNA-depleted and mRNA-purified, was assessed by Qubit dsDNA (Thermo Fisher Scientific) and LabChip (PerkinElmer, Waltham, MA). Samples were pooled and sequenced on an Illumina NovaSeq S4 using 150 bp paired-end reads (Illumina, San Diego, CA). BCL Convert Conversion Software v3.9.3 (Illumina) demultiplexed FASTQ files. RNA-seq was performed by the University of Michigan Advanced Genomics Core.

RNA-seq data processing and quality control

Raw RNA-seq reads (FASTQ files) were trimmed in Cutadapt v2.3, aligned to the human reference genome hg38 in STAR v2.5.3a, and quantified in featureCounts v2.0.3 (using "-s 2" for reverse strand reads). Rigorous RNA-seq quality control was implemented by FastQC v0.11.9, RSeQC v5.01, and FastQ Screen v0.15.2, and results were summarized with MultiQC v1.7. mRNA samples with fewer than 10 million reads aligned to non-globin genes were re-sequenced. Expression of the X chromosome gene *XIST* and 49 male-specific Y chromosome genes was analyzed to compare observed sex to recorded sex[65], filtering out five samples with discrepancies, which were removed. Thirteen globin genes were filtered because they are highly expressed in red blood cells, which are present in blood samples even when nucleated cells (e.g., white blood cells) are the primary target of analysis. Finally, genes with low expression, defined as fewer than five raw read counts in over 50% of samples, and genes only present in one library were excluded. Overall, this yielded 22,332 genes from 694 samples.

Identification of the relative abundance of different cell types in blood (cell proportions)

When a disease alters immune cell levels, such as occurs in ALS[24–26,30], gene expression changes in whole blood transcriptomics can arise secondary to changes in the relative abundance of the various immune cell types, i.e., in cell proportions. Adjusting the whole blood transcriptomic dataset for immune cell proportions, i.e., adjusting for gene expression changes that arise merely from differences in immune cell proportions, reveals intrinsic gene expression changes linked to the disease process rather than only to immune cell levels. Herein, the primary whole blood transcriptomics dataset (termed "primary") was employed for case-control and survival prediction, since these gene expression changes only need to differentiate case from control and shorter- versus longer-surviving cases, regardless of whether these gene expression changes arise from altered immune cell levels or from intrinsic biological processes. The whole blood transcriptomic dataset adjusted for cell proportions (termed "adjusted") was used for pathway and drug perturbation analyses, since these methods require assessment of intrinsic disease-related biological pathways.

Adjusting the whole blood transcriptomic dataset requires accurate accounting of cell proportions. Cell proportions can be computationally inferred by deconvolution of RNA-seq data or deconvolution of DNAm data. To identify the optimal approach in this dataset, deconvolution by DNAm and RNA-seq was compared to immune cell profiling by flow cytometry. These data were available for a substantial proportion of this deeply phenotyped cohort; DNAm was previously profiled by microarray in 428 ALS cases and 288 controls[32], while immune cell levels and activation states were previously profiled by flow cytometry in 225 ALS cases and 119 controls[24–26].

Cell type deconvolution using DNAm was conducted using the FlowSorted.Blood.EPIC v1.4.1 R package[66] for CD8 + T cells, CD4 + T cells, NK cells, B cells, monocytes, and neutrophils. Cell type deconvolution using RNA-seq was conducted in BayesPrism v2.1.2[67] for CD8 + T cells, CD4 + T cells, NK cells, B cells, monocytes, neutrophils, and erythrocytes. Flow cytometry profiled CD8 + T cells, CD4 + T cells, NK cells, monocytes, and neutrophils. Correlation of cell proportions by DNAm and RNA-seq with cell proportions determined by flow cytometry was performed using Pearson's correlation analysis. Based on the correlation results, the cell proportions estimated by DNAm were used for differential gene expression analyses adjusted for cell proportions.

Differential gene expression analysis

To improve the DEG analysis, unknown sources of variation in gene expression, such as differences in library preparation, were first eliminated[68] by implementing a modified version of surrogate variable analysis (SVA), a batch correction method[69–71]. SVA estimates the principal components (PCs) after adjusting for biological variables (case-control status, age, sex). However, SVA selects the top k factors as the criteria for selecting PCs, while our modified method selects the top k PCs based on an elbow plot (Supplementary Fig. S14a; max. variance explained), excluding PCs that correlate with the main biological variable, i.e., case-control status. Indeed, after applying SVA, which included surrogate variables in the DEG model, the residual principal component analysis (PCA) plot still showed clustering based on library preparation type (Supplementary Fig. S14b), indicating remaining uncaptured unknown variance. However, when PCs were carefully selected for the DEG model, the residual PCA plot did not exhibit any differences resulting from library preparation types (Supplementary Fig. S14c). PCs were carefully selected by correlation analysis using Pearson's linear correlation and logistic regression adjusted for all other model covariates. PC1 and PC3 significantly correlated with case-control status in logistic regression (Bonferroni-adjusted p-value < 0.05), so they were omitted from the DEG model. Gene counts were normalized for library size in count per million (CPM), $\log_2$-transformed, and normalized between samples by the trimmed mean of M-values (TMM) normalization method in edgeR v4.2.1.

DEGs between ALS cases versus controls were identified with the model **y ~ group + age + sex + genetic_PCs$_{1\text{-}4}$ + RNA_PCs$_{2\text{-}10}$ + batch + library_type** in the limma R package (v3.48.3; functions: voom, lmFit, eBayes). This model controlled for variation in demographics (age, sex) as well as technical variations due to library preparation, genetic heterogeneity (the first four genetic PCs), and unknown variance

(selected RNA PCs) (Supplementary Fig. S14d). DEGs met the criteria of absolute value of log_2(fold-change)>0.1 (absLFC) and FDR < 0.01. DEG analysis for ALS cases versus controls was also performed stratified by sex. DEGs identified by the model above, without adjusting for cell proportions, were used as input features for classifiers predicting ALS case status and survival.

DEGs between ALS cases versus controls adjusted for cell proportions were identified with the model **y ~ group + age + sex + CD8T + Mono + NK + BCell + Neu + genetic_PCs$_{1\text{-}4}$ + RNA_PCs$_{2,4\text{-}10}$ + batch + library_type** in the limma R package (v3.48.3; functions: voom, lmFit, eBayes), where covariates *CD8T* through *Neu* represent cell type proportions estimated by DNAm data. This model controlled for all the variables included in the model that were not adjusted for cell proportions, as well as for variation in cell proportions. The proportions of the different immune cells sum up to a constant, near 1, indicating linear dependence among them. To address this, the CD4T proportion was removed from the model due to a large variance-inflation factor. Accounting for cell proportions ensures that DEGs and corresponding pathway enrichment reflect disease-associated alterations in gene activity rather than ALS-related changes in the numbers of immune cells, i.e., the cell proportions.

Correlation analysis between RNA PCs and estimated cell proportions (neutrophils, NKs, CD8 T cells, monocytes, B cells) by Pearson correlation identified seven PC–cell type pairs that met the criteria | r | >0.25 and Bonferroni-adjusted $p < 0.05$ (Supplementary Fig. S15a, b). Removing the PCs that correlated with cell type proportions (PC_3, PC_4, & PC_9) reintroduced batch effects between library preparation types (Supplementary Fig. S15c); therefore, the original model was retained, called the "primary" analysis, which accounted, in part only, for cell proportions, in contrast to the analysis that adjusted for the remainder of cell proportions (i.e., referred to as adjusted for cell proportions). There was substantial overlap in DEGs for our primary analysis and adjusted for cell proportions, but the two DEGs analyses also identified hundreds of unique up- and down-DEGs (Supplementary Fig. S16).

Gene expression signature for predicting ALS case status

To compare machine learning classifiers for this task, a random 70% ($n = 487$ total with $n = 296$ ALS, $n = 191$ controls) of the total gene expression dataset ($n = 694$) was used for 10-fold cross-validation training with the remaining 30% held out for testing ($n = 207$ total with $n = 126$ ALS, $n = 81$ controls). The same training and testing sets were used across all seven machine learning algorithms to ensure comparability. TMM-normalized log_2CPM values were used as input for prediction. Seven machine learning algorithms were compared: penalized logistic regression with different regularization parameters, including ridge ($\alpha = 0$, where α is the mixing parameter that determines the balance between L1 and L2 regularization)[72], least absolute shrinkage and selection operator (LASSO) ($\alpha = 1$)[73], and elastic net ($\alpha = 0.5$)[74], L1/2[75], smoothly clipped absolute deviation (SCAD)[76], maximum concave penalty (MCP)[77], and extreme gradient boosting (XGBoost). These classification algorithms were chosen because they are more robust than deep learning for tabular data with our sample number.

Next, given the top performance of XGBoost, using 100% of RNA-seq samples in the training dataset ($n = 694$ total of which $n = 422$ ALS, $n = 272$ controls), the residuals of TMM-normalized log_2CPM values (adjusted for library type, batch, and technical PCs 2, 4, and 5) were used to train an XGBoost classifier in Python packages xgboost v2.1.1, scikit-learn v1.3.0, and kneed v0.8.5. Different sets of input genes, which defined the universe of potential features, were tested and the classifier was trained using 10-fold cross validation. Specifically, FDR < 0.01 and average log_2 gene expression (AvExpr) were identified as the most important criteria for filtering primary DEGs to use as input genes to train XGBoost models. Three different criteria, (i) FDR < 0.01, (ii) FDR < 0.01 and AvExpr>0, and (iii) FDR < 0.01 and AvExpr>2, were tested for filtering DEGs, and 10-fold cross-validation training was performed for each. Using each set of filtered DEGs as input, three initial models were trained with a learning rate of 0.1, max depth 20, lambda 0.0003, alpha 0.0003, and learning task set to binary classification by logistic regression, outputting probability. For feature selection for each of the three models, based on the relative importance of each feature, we repeatedly fit the models with the same hyperparameters by excluding the least important features in each iteration (i.e., recursive feature elimination). By plotting the average ROC AUC score from 10-fold cross-validation versus the number of features in each model, the kneedle algorithm selected the best knee point, corresponding to the final number of features in the candidate gene classifiers. These three candidate classifiers [trained on (i) FDR < 0.01, (ii) FDR < 0.01, AvExpr>0, (iii) FDR < 0.01, AvExpr>2] were further combined into one ensemble classifier based on averaged prediction probabilities.

The external Grima et al. test dataset was downloaded from NCBI Gene Expression Omnibus (GSE234297)[21], containing RNA-seq of peripheral blood from 86 ALS cases versus 48 matched controls. Genes with >5 counts in >10 samples were retained. TMM-normalized log_2CPM values were adjusted for batch, RIN, and technical PCs 2 to 7 using the approach described above. The three XGBoost classifiers and the ensemble classifier were tested on the external dataset and were assessed by accuracy, sensitivity, specificity, and AUC.

Gene expression signature for predicting ALS survival

In our case-control classification models, feature selection was performed on DEGs adjusted for age and sex and filtered based on average expression. Slightly different pre-processing and filtering criteria were employed to optimize the predictions for this case-only analysis. Low expressed genes were filtered out using the filterByExpr function in the edgeR v4.2.1 R package, and then further filtered out average count per million mapped reads (CPM) < 10 and CPM < 1 in >10% of the 422 ALS samples. Only protein-coding genes also present in the external dataset Grima et al.[21] were retained, resulting in 5,449 genes for feature selection. TMM-normalized (within ALS case samples) log_2CPM values were adjusted for library type, batch, and technical PCs 1 and 2 using regression and subsequently used as input for prediction. Additionally, clinical variables, including onset segment, symptom onset age, and sex, were used for feature selection, selected based on shared data availability between this study with the external test Grima cohort[21]. Two machine-learning methods, stepwise selection and XGBoost, were applied to select gene panels combined with clinical variables to predict survival, which were compared to a clinical variable-only model. A log-rank test was conducted for each gene, using its median expression value to stratify ALS cases ($n = 420$ with survival data) into high- versus low-expression groups. This analysis, conducted without clinical variables, identified 575 significant genes that predicted survival (p-value < 0.01).

Stepwise selection training. The stepwise model was trained on 100% of the internal ALS dataset ($n = 420$) for survival analysis. To identify input genes, first, weighted gene co-expression network analysis (WGCNA)[78] by WGCNA v1.73 R package was performed on the 575 genes, which established 15 clusters based on similarity in gene expression, including Module 0, that contained genes that did not co-express with genes in other modules. A soft threshold of 6 was set to achieve a scale independence of 0.8 while ensuring that the adjacency matrix exhibited relatively high average connectivity. To reduce multicollinearity in the 575 genes from the univariate log-rank model, we selectively kept the top 10% hub genes ranked by connectivity within each module as well as genes without highly correlated counterparts (within WGCNA Module 0). This resulted in 82 WGCNA genes as input for stepwise feature selection. Next, the stepwise Akaike information criterion approach implemented in the MASS v7.3.60.2[79] R package was used to further enhance feature selection for fitting the

multivariate Cox proportional hazards model with both forward and backward selection. Stepwise selected 18 gene features in addition to the three clinical variables.

XGBoost training. The XGBoost algorithm from the xgboost v0.1.0[80] R package was trained on 100% of the internal ALS dataset ($n = 420$) for survival analysis. A boosted tree model including clinical variables and the 575 genes selected by the log-rank test (p-value < 0.01) was generated using the entire training dataset. A grid search was performed in the hyperparameter space, and the final model was trained with the best hyperparameters, including the learning rate 0.1, max depth 6, lambda 1, alpha 0.01, gamma 0.1, and learning task survival analysis using a Cox proportional hazards model evaluated by the negative log partial likelihood. The top 10 features based on their importance (gain) in predicting survival risk were selected, which included 8 gene features and 2 clinical variables.

Testing on the external dataset. The external Grima[21] test dataset was downloaded from NCBI Gene Expression Omnibus (GSE234297). The RNA-seq raw read counts from 86 sporadic ALS cases were TMM-normalized and the $\log_2$CPM values were adjusted for batch, RIN, library size, and technical PCs 2 to 5 using regression and subsequently used as input. Models constituted all clinical variables in the clinical variables model (onset segment, symptom onset age, sex) and all clinical variables plus the selected gene features in the stepwise (18 selected genes) and XGBoost model (8 selected genes). Cox proportional hazards models were fitted using the coxph function within the survival v3.7.0 R package[81] and the predict function calculated risk scores for the internal and external test data. The timeROC v0.4 R package[82] evaluated the time-dependent ROC curves for the three survival models to comprehensively assess the predictive accuracy of survival models over time. AUCs of the ROC curves were calculated at various time points (2, 3, 4, 5, 6, 7, 8 years) for prediction in five random train-test splits of our internal dataset and in the Grima[21] dataset to compare the discriminative power of the models. The Kaplan-Meier plots were generated using the survminer v0.4.9 R package to visualize the predicted survival probabilities over time for each model. C-index of the different CoxPH models were compared using the compare v1.3.2 R package.

Pathway analysis

Over-representation analysis was performed using a one-sided Fisher's exact test from the functions enrichKEGG, enricher, and enrichGO in the clusterProfiler v4.12.2 R package on KEGG pathways, MSigDB Hallmark gene sets, and GO terms. DEGs with FDR < 0.01 and absLFC>0.1 were used for ALS cases versus control comparisons in all participants, and in sex-stratified analyses. KEGG pathways and Hallmark gene sets with FDR < 0.05 were selected for visualization in dot plots. GO terms with FDR < 0.01 were clustered and visualized using the EnrichmentMap v3.5.0 module in Cytoscape v3.10.0. Gene set enrichment analysis was performed on participants in the lowest quartile of ALSFRS-R (most functionally impaired) versus the highest quartile of ALSFRS-R scores (least functionally impaired). $-\text{Log}_{10}$(p-value) x sign(log[fold-change]) was used as the statistic to rank the gene list for inputs to gene set enrichment analysis, where p-values and sign(logFC) were from primary DEGs and DEGs adjusted for cell type proportions.

Drug perturbation analysis

DEGs were obtained from previous publications of human iPSC-derived cortical-like neurons with TDP-43 knockdown ($n = 3$) versus controls ($n = 4$)[34] and postmortem spinal cord (214 ALS versus 57 controls)[35], and examined for overlap with whole blood DEGs (422 ALS versus 272 controls) from this study, both primary and adjusted for cell proportions. The same criteria were used for defining DEGs in all datasets: FDR < 0.01 and absLFC>0.1. Overlapping DEGs across the three datasets were used as input for drug perturbation analysis. For iPSC-neurons and postmortem spinal cord, the top 150 upregulated and 150 downregulated DEGs (ranked by LFC) were used as input, the maximum number of input DEGs allowed, per the clue.io platform. Drug perturbation analysis was performed using the latest Touchstone dataset from L1000 via the clue.io platform to identify candidate compounds targeting pathways involved in ALS, using averaged perturbations across all available cell types[83]. Normalized connectivity scores (norm_cs; −2 to 2), a measure of the extent of matching between input DEGs and database compounds, were calculated to assess compound-DEG relationships, with positive scores indicating promotive effects, negative scores indicating suppressive effects. Only compounds with defined mechanisms of action (MOA ≠ "") , passing quality control (qc_pass = 1), and FDR < 0.01 were considered. This approach identified high-confidence candidate compounds for further investigation in ALS. R packages cmapR v1.21.0, igraph v2.1.4, and GGally v2.2.1 were used for drug perturbation analysis and network visualization.

qPCR validation. Select DEGs were validated in an independent cohort ($n = 29$ ALS cases, $n = 27$ controls) by qPCR. Participants providing samples for the qPCR validation underwent the same recruitment, enrollment, and consent as described earlier. RNA was isolated from participant blood samples using the PAXGene miRNA kit (catalog no. 763134, Qiagen). cDNA was then generated from 500 ng input RNA using iScript cDNA Synthesis kit (catalog no. 1708890, Bio-Rad, Hercules, CA). qPCR was performed using TaqMan Gene Expression Master Mix (catalog no. 4369016, Thermo Fisher Scientific) with TaqMan Gene Expression Assay primers for *B2M* (catalog no. 4331182, Assay ID Hs00187842_m1), *CAPZA1* (catalog no. 4331182, Assay ID Hs04187789_g1), *RPS18* (catalog no. 4331182, Assay ID Hs01375212_g1), *TPT1* (catalog no. 4331182, Assay ID Hs00372008_m1), and *TNFSF10* (catalog no. 4331182 Assay ID Hs00921974_m1) using both *GAPDH* (catalog no. 4331182 Assay ID Hs02786624_g1) and *ACTB* (catalog no. 4331182 Assay ID Hs01060665_g1) as endogenous controls, and data were analyzed by the delta-delta C_T method.

Statistics & reproducibility. This was a case-control study of peripheral whole blood by RNA-seq transcriptomics, with an additional prospective cohort study of survival. The sample size was determined based on budget constraints and available samples. The sample size is the largest compared to all previously published studies of ALS participant blood by RNA-seq. Sequencing of the samples was performed in a blinded manner. Randomization was not a feature of the study design. Expression of the X chromosome gene XIST and 49 male-specific Y chromosome genes was analyzed to compare observed sex to recorded sex, filtering out five samples with discrepancies, which were removed. Otherwise, all samples were included. No replication was undertaken; however, select differentially expressed genes between ALS and control were validated by qPCR in an independent cohort ($n = 29$ ALS cases, $n = 27$ controls). Case-control and survival prediction models were validated in an independent external ALS case-control blood RNA-seq dataset.

Details of statistical and computational analyses were detailed in their respective sections in "Differential gene expression analysis" and "Pathway analysis".

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The RNA-seq data generated in this study have been deposited in the European Genome-Phenome Archive (EGA) database under accession

code EGAS50000001019. The RNA-seq data and associated clinical/demographical data are available under restricted access for ALS disease-specific research, and access can be obtained by submitting requests through EGA and completing a Data Use Agreement that outlines requirements for use. Requests are limited to not-for-profit organizations. Source data are provided with this paper.

Code availability

Code has been made available at https://github.com/yzhao80/ALS_RNA.

References

1. Gwathmey, K. G. et al. Diagnostic delay in amyotrophic lateral sclerosis. *Eur. J. Neurol.* **30**, 2595–2601 (2023).
2. Vazquez-Costa, J. F. et al. Analysis of the diagnostic pathway and delay in patients with amyotrophic lateral sclerosis in the Valencian Community. *Neurologia (Engl. Ed.)* **36**, 504–513 (2021).
3. Feldman, E. L. et al. Amyotrophic lateral sclerosis. *Lancet* **400**, 1363–1380 (2022).
4. Richards, D., Morren, J. A. & Pioro, E. P. Time to diagnosis and factors affecting diagnostic delay in amyotrophic lateral sclerosis. *J. Neurol. Sci.* **417**, 117054 (2020).
5. Chiò, A., Calvo, A., Moglia, C., Mazzini, L. & Mora, G. Phenotypic heterogeneity of amyotrophic lateral sclerosis: a population based study. *J. Neurol. Neurosurg. Psychiatry* **82**, 740–746 (2011).
6. Thakore, N. J., Lapin, B. R., Mitsumoto, H. & Pooled Resource Open-Access Als Clinical Trials, C. Early initiation of riluzole may improve absolute survival in amyotrophic lateral sclerosis. *Muscle Nerve* **66**, 702–708 (2022).
7. Miller, R. G. et al. Practice parameter update: the care of the patient with amyotrophic lateral sclerosis: multidisciplinary care, symptom management, and cognitive/behavioral impairment (an evidence-based review): report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology* **73**, 1227–1233 (2009).
8. Benatar, M., Wuu, J., Andersen, P. M., Lombardi, V. & Malaspina, A. Neurofilament light: A candidate biomarker of presymptomatic amyotrophic lateral sclerosis and phenoconversion. *Ann. Neurol.* **84**, 130–139 (2018).
9. Giacomucci, G. et al. Future perspective and clinical applicability of the combined use of plasma phosphorylated tau 181 and neurofilament light chain in Subjective Cognitive Decline and Mild Cognitive Impairment. *Sci. Rep.* **14**, 11307 (2024).
10. Lista, S. et al. A critical appraisal of blood-based biomarkers for Alzheimer's disease. *Ageing Res Rev.* **96**, 102290 (2024).
11. Zarkali, A. et al. Neuroimaging and plasma evidence of early white matter loss in Parkinson's disease with poor outcomes. *Brain Commun.* **6**, fcae130 (2024).
12. Bittner, S. et al. Clinical implications of serum neurofilament in newly diagnosed MS patients: A longitudinal multicentre cohort study. *EBioMedicine* **56**, 102807 (2020).
13. Määttä, L. L. et al. Longitudinal Change in Serum Neurofilament Light Chain in Type 2 Diabetes and Early Diabetic Polyneuropathy: ADDITION-Denmark. *Diabetes Care* **47**, 986–994 (2024).
14. Leckey, C. A. et al. CSF neurofilament light chain profiling and quantitation in neurological diseases. *Brain Commun.* **6**, fcae132 (2024).
15. Beydoun, M. A. et al. Serum neurofilament light chain as a prognostic marker of all-cause mortality in a national sample of US adults. *Eur. J. Epidemiol.* https://doi.org/10.1007/s10654-024-01131-7 (2024).
16. Parker, J. S. et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *J. Clin. Oncol.* **41**, 4192–4199 (2023).
17. Nielsen, T. et al. Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens. *BMC Cancer* **14**, 177 (2014).
18. Tosoian, J. J. et al. Development and Validation of an 18-Gene Urine Test for High-Grade Prostate Cancer. *JAMA Oncol.* **10**, 726–736 (2024).
19. van Rheenen, W. et al. Whole blood transcriptome analysis in amyotrophic lateral sclerosis: A biomarker study. *PLoS One* **13**, e0198874 (2018).
20. Swindell, W. R., Kruse, C. P. S., List, E. O., Berryman, D. E. & Kopchick, J. J. ALS blood expression profiling identifies new biomarkers, patient subgroups, and evidence for neutrophilia and hypoxia. *J. Transl. Med.* **17**, 170 (2019).
21. Grima, N. et al. RNA sequencing of peripheral blood in amyotrophic lateral sclerosis reveals distinct molecular subtypes: Considerations for biomarker discovery. *Neuropathol. Appl Neurobiol.* **49**, e12943 (2023).
22. Goutman, S. A. et al. Recent advances in the diagnosis and prognosis of amyotrophic lateral sclerosis. *Lancet Neurol*, https://doi.org/10.1016/S1474-4422(21)00465-8 (2022).
23. Goutman, S. A. et al. Metabolomics identifies shared lipid pathways in independent amyotrophic lateral sclerosis cohorts. *Brain*, https://doi.org/10.1093/brain/awac025 (2022).
24. Murdock, B. J., Goutman, S. A., Boss, J., Kim, S. & Feldman, E. L. Amyotrophic Lateral Sclerosis Survival Associates With Neutrophils in a Sex-specific Manner. *Neurol. Neuroimmunol. Neuroinflamm.* **8**, e953 (2021).
25. Murdock, B. J. et al. Natural killer cells associate with amyotrophic lateral sclersois in a sex- and age-dependent manner. *JCI Insight*, https://doi.org/10.1172/jci.insight.147129 (2021).
26. Murdock, B. J. et al. Peripheral Immune Profiles Predict ALS Progression in an Age- and Sex-Dependent Manner. *Neurol. Neuroimmunol. Neuroinflamm* **11**, e200241 (2024).
27. Jutzi, D. et al. Aberrant interaction of FUS with the U1 snRNA provides a molecular mechanism of FUS induced amyotrophic lateral sclerosis. *Nat. Commun.* **11**, 6341 (2020).
28. Mao, H. et al. RGS17/RGSZ2, a novel regulator of Gi/o, Gz, and Gq signaling. *J. Biol. Chem.* **279**, 26314–26322 (2004).
29. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*). Association for Computing Machinery (2016).
30. Murdock, B. J. et al. Correlation of Peripheral Immunity With Rapid Amyotrophic Lateral Sclerosis Progression. *JAMA Neurol.* **74**, 1446–1454 (2017).
31. Figueroa-Romero, C. et al. Tofacitinib suppresses natural killer cells in vitro and in vivo: implications for amyotrophic lateral sclerosis. *Front Immunol.* **13**, 773288 (2022).
32. Zhao, Y. et al. Epigenetic age acceleration is associated with occupational exposures, sex, and survival in amyotrophic lateral sclerosis. *EBioMedicine* **109**, 105383 (2024).
33. Braun, M. C. et al. Defective daily temperature regulation in a mouse model of amyotrophic lateral sclerosis. *Exp. Neurol.* **311**, 305–312 (2019).
34. Brown, A. L. et al. TDP-43 loss and ALS-risk SNPs drive mis-splicing and depletion of UNC13A. *Nature* **603**, 131–137 (2022).
35. Ziff, O. J. et al. Integrated transcriptome landscape of ALS identifies genome instability linked to TDP-43 pathology. *Nat. Commun.* **14**, 2176 (2023).
36. Zhou, F., Liu, Y., Liu, D., Xie, Y. & Zhou, X. Identification of basement membrane-related signatures for estimating prognosis, immune infiltration landscape and drug candidates in pancreatic adenocarcinoma. *J. Cancer* **15**, 401–417 (2024).
37. Saris, C. G. et al. Weighted gene co-expression network analysis of the peripheral blood from Amyotrophic Lateral Sclerosis patients. *BMC Genomics* **10**, 405 (2009).

38. Westeneng, H. J. et al. Prognosis for patients with amyotrophic lateral sclerosis: development and validation of a personalised prediction model. *Lancet Neurol.* **17**, 423–433 (2018).
39. Benatar, M. et al. Validation of serum neurofilaments as prognostic and potential pharmacodynamic biomarkers for ALS. *Neurology* **95**, e59–e69 (2020).
40. Miller, T. M. et al. Trial of Antisense Oligonucleotide Tofersen for SOD1 ALS. *N. Engl. J. Med.* **387**, 1099–1110 (2022).
41. Camu, W. et al. Repeated 5-day cycles of low dose aldesleukin in amyotrophic lateral sclerosis (IMODALS): A phase 2a randomised, double-blind, placebo-controlled trial. *EBioMedicine* **59**, 102844 (2020).
42. Benatar, M. et al. Design of a randomized, placebo-controlled, phase 3 trial of tofersen initiated in clinically presymptomatic SOD1 Variant Carriers: the ATLAS Study. *Neurotherapeutics* **19**, 1248–1258 (2022).
43. Kadena, K. & Ouzounoglou, E. Drug repurposing for amyotrophic lateral sclerosis based on gene expression similarity and structural similarity: a cheminformatics, genomic and network-based analysis. *BioMedInformatics* **4**, 1713–1724 (2024).
44. McLaughlin, R. L. et al. Genetic correlation between amyotrophic lateral sclerosis and schizophrenia. *Nat. Commun.* **8**, 14774 (2017).
45. Stommel, E. W., Graber, D., Montanye, J., Cohen, J. A. & Harris, B. T. Does treating schizophrenia reduce the chances of developing amyotrophic lateral sclerosis?. *Med Hypotheses* **69**, 1021–1028 (2007).
46. Cortes-Flores, H., Torrandell-Haro, G. & Brinton, R. D. Association between CNS-active drugs and risk of Alzheimer's and age-related neurodegenerative diseases. *Front Psychiatry* **15**, 1358568 (2024).
47. D'Ovidio, F. et al. Amyotrophic Lateral Sclerosis Incidence and Previous Prescriptions of Drugs for the Nervous System. *Neuroepidemiology* **47**, 59–66 (2016).
48. Hempel, C. et al. The phenothiazine-class antipsychotic drugs prochlorperazine and trifluoperazine are potent allosteric modulators of the human P2X7 receptor. *Neuropharmacology* **75**, 365–379 (2013).
49. Ruiz-Ruiz, C., Calzaferri, F. & García, A. G. P2X7 Receptor Antagonism as a Potential Therapy in Amyotrophic Lateral Sclerosis. *Front Mol. Neurosci.* **13**, 93 (2020).
50. Zheng, C. et al. Ibrutinib Delays ALS Installation and Increases Survival of SOD1(G93A) Mice by Modulating PI3K/mTOR/Akt Signaling. *J. Neuroimmune Pharm.* **18**, 383–396 (2023).
51. Imamura, K. et al. The Src/c-Abl pathway is a potential therapeutic target in amyotrophic lateral sclerosis. *Sci Transl Med.* **9**, eaaf3962 (2017).
52. Imamura, K. et al. Safety and tolerability of bosutinib in patients with amyotrophic lateral sclerosis (iDReAM study): A multicentre, open-label, dose-escalation phase 1 trial. *EClinicalMedicine* **53**, 101707 (2022).
53. Palomo, V., Nozal, V., Rojas-Prats, E., Gil, C. & Martinez, A. Protein kinase inhibitors for amyotrophic lateral sclerosis therapy. *Br. J. Pharm.* **178**, 1316–1335 (2021).
54. Goutman, S. A. et al. Emerging insights into the complex genetics and pathophysiology of amyotrophic lateral sclerosis. *Lancet Neurol*, https://doi.org/10.1016/S1474-4422(21)00414-2 (2022).
55. Duan, Q. Q. et al. TBK1, a prioritized drug repurposing target for amyotrophic lateral sclerosis: evidence from druggable genome Mendelian randomization and pharmacological verification in vitro. *BMC Med.* **22**, 96 (2024).
56. Kondapuram, S. K. & Coumar, M. S. Pan-cancer gene expression analysis: Identification of deregulated autophagy genes and drugs to target them. *Gene* **844**, 146821 (2022).
57. Erb, M. L. & Moore, D. J. LRRK2 and the Endolysosomal System in Parkinson's Disease. *J. Parkinsons Dis.* **10**, 1271–1291 (2020).
58. Moore, A. S. & Holzbaur, E. L. Dynamic recruitment and activation of ALS-associated TBK1 with its target optineurin are required for efficient mitophagy. *Proc. Natl Acad. Sci. USA* **113**, E3349–E3358 (2016).
59. Kubat Oktem, E., Aydin, B., Yazar, M. & Arga, K. Y. Integrative Analysis of Motor Neuron and Microglial Transcriptomes from SOD1(G93A) Mice Models Uncover Potential Drug Treatments for ALS. *J. Mol. Neurosci.* **72**, 2360–2376 (2022).
60. Beers, D. R. & Appel, S. H. Immune dysregulation in amyotrophic lateral sclerosis: mechanisms and emerging therapies. *Lancet Neurol.* **18**, 211–220 (2019).
61. Ohnstad, H. O. et al. Prognostic value of PAM50 and risk of recurrence score in patients with early-stage breast cancer with long-term follow-up. *Breast Cancer Res.* **19**, 120 (2017).
62. Goutman, S. A. et al. High plasma concentrations of organic pollutants negatively impact survival in amyotrophic lateral sclerosis. *J. Neurol. Neurosurg. Psychiatry* **90**, 907–912 (2019).
63. Goutman, S. A. et al. Environmental risk scores of persistent organic pollutants associate with higher ALS risk and shorter survival in a new Michigan case/control cohort. *J. Neurol. Neurosurg. Psychiatry* **95**, 241–248 (2024).
64. Goutman, S. A. et al. Associations of self-reported occupational exposures and settings to ALS: a case-control study. *Int Arch. Occup. Environ. Health* **95**, 1567–1586 (2022).
65. Godfrey, A. K. et al. Quantitative analysis of Y-Chromosome gene expression across 36 human tissues. *Genome Res.* **30**, 860–873 (2020).
66. Syage, A. R. et al. Single-cell RNA sequencing reveals the diversity of the immunological landscape following central nervous system infection by a murine coronavirus. *J. Virol.* **94** (2020).
67. Chu, T., Wang, Z., Pe'er, D. & Danko, C. G. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat. Cancer* **3**, 505–517 (2022).
68. Li, S. et al. Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotechnol.* **32**, 888–895 (2014).
69. Zhou, H. J., Li, L., Li, Y., Li, W. & Li, J. J. PCA outperforms popular hidden variable inference methods for molecular QTL mapping. *Genome Biol.* **23**, 210 (2022).
70. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* **3**, 1724–1735 (2007).
71. Leek, J. T. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res* **42**, e161 (2014).
72. Hoerl, A. E. & Kennard, R. W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
73. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.: Ser. B (Methodol.)* **58**, 267–288 (1996).
74. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B: Stat. Methodol.* **67**, 301–320 (2005).
75. Xu, Z., Zhang, H., Wang, Y., Chang, X. & Liang, Y. L1/2 regularization. *Sci. China Inf. Sci.* **53**, 1159–1169 (2010).
76. Fan, J. & Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96**, 1348–1360 (2001).
77. Zhang, C.-H. Nearly Unbiased Variable Selection Under Minimax Concave Penalty. *Ann. Stat.* **38**, 894–942 (2010).
78. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinforma.* **9**, 559 (2008).
79. Venables W. N., Ripley B. D. *Modern applied statistics with S*. Springer (2010).
80. Li, K. et al. Efficient gradient boosting for prognostic biomarker discovery. *Bioinformatics* **38**, 1631–1638 (2022).
81. Therneau T., Grambsch P. *Modeling survival data: Extending the cox model*, 1 edn. Springer (2010).

82. Blanche, P., Dartigues, J. F. & Jacqmin-Gadda, H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat. Med.* **32**, 5381–5397 (2013).
83. Subramanian, A. et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000. *Profiles Cell* **171**, 1437–1452.e1417 (2017).

Acknowledgements

We are grateful to the study participants and their families. We also thank Crystal Pacut, Diana M. Rigan, Jayna Duell, RN, Caroline Piecuch, Alyssa Braun, Aanchal Gopalan, Nicole Bopp, Raymond G. Cavalcante, PhD, and Adam Patterson for study support. We also acknowledge Professor Yuanfang Guan, PhD, for discussions on survival prediction. This study was supported by the NINDS (R01NS127188 to SAB, SAG, MAS, ELF; R01NS120926 to SAG), National ALS Registry/CDC/ATSDR (1R01TS000289 to ELF; R01TS000327 to SAG and ELF), NIEHS (P30ES017885 to SAB; R01ES030049 to SAB, SAG, and ELF), National Center for Advancing Translational Sciences at the National Institutes of Health (UL1TR002240 to ELF), Intramural Research Program of the National Institute on Aging (ZIA AG000933 to BJT), ALS Association (20-IIA-532 to SAG), James and Margaret Hiller (to SAG), Eric and Linda Novak (to SAG), the Coleman Therapeutic Discovery Fund (to ELF), the Peter R. Clark Fund for ALS Research (to ELF), the Sinai Medical Staff Foundation (to ELF), the Scott L. Pranger ALS Clinic Fund (to ELF), the Dr. Randall W. Whitcomb Fund for ALS Genetics (to ELF), the Richard Stravitz Foundation (to ELF), the Stanford Morris ALS Research Fund (to SAG and ELF), and the NeuroNetwork for Emerging Therapies, University of Michigan (to ELF).

Author contributions

S.A.G., M.A.S., and E.L.F. planned and oversaw all aspects of the study. Y.Z. performed the differential gene expression, case-control prediction, survival prediction, and pathway analyses, and contributed substantially to the conception of the work, methodology, data curation, visualization, and interpretation. M.G.S. contributed substantially to the conception of the work, interpretation of data, and data visualization. X.L. performed the case-control prediction and cell-type deconvolution analyses, and data curation and visualization. K.G. performed the drug perturbation analysis and data visualization. K.W. performed RNA sequencing preprocessing and contributed substantially to the cell-type deconvolution analysis, data curation, and visualization. M.L. contributed substantially to the survival prediction analysis and data visualization. B.L. contributed substantially to the survival prediction analysis and data visualization. G.I. contributed to the pathway analysis. S.A.S. oversaw the project management of the study. S.J.T. isolated RNA, submitted samples for RNA sequencing, performed qPCR, and curated the study database. L.Z., K.M.B., A.K., S.A.B., and J.H. contributed substantially to the methodology and interpretation of data. J.F.D. and B.J.T. contributed substantially to data curation. Y.Z., M.G.S., and E.L.F. wrote the manuscript with input and substantial revisions from all authors.

Competing interests

YZ: None. MGS: None. XL: None. KG: None. KW: None. ML: None. BL: None. GI: None. SAS: Listed as inventors on a patent, Issue number US10660895, held by the University of Michigan, titled "Methods for Treating Amyotrophic Lateral Sclerosis," that targets immune pathways for use in ALS therapeutics. SAS has received unrelated research funding from NIH R01DK129320, which has no competing interest with this work. LZ: None. SJT: None. KMB None. JFD: None. BJT: BJT holds patents on the diagnostic and therapeutic implications of the C9orf72 repeat expansion. BJT has a patent pending (U.S. Patent Application No. 63/717,807) on the diagnostic testing for ALS based on the proteomic panel. BJT has received unrelated research funding from Cerevel Therapeutics and the ALS Association, which have no competing interest with this work. BJT holds a leadership role in an Advisory Committee of the American Neurological Association. BJT serves on the editorial board of EClinicalMedicine and JNNP and is an associate editor for Brain. BJT also has collaborative research agreements with Ionis Pharmaceuticals, Roche, and Optimeos. AK: None. SAB: None. JH: None. SAG: Listed as inventors on a patent, Issue number US10660895, held by the University of Michigan titled "Methods for Treating Amyotrophic Lateral Sclerosis" that targets immune pathways for use in ALS therapeutics. Scientific consulting for Evidera. Payment from the American Academy of Neurology. MAS: None. ELF: ELF is listed as an inventor on a patent, Issue number US10660895, held by the University of Michigan titled "Methods for Treating Amyotrophic Lateral Sclerosis" that targets immune pathways for use in ALS therapeutics. ELF has received unrelated funding from NIH, CDC/ATSDR, DoD, Breakthrough T1D, and the American Heart Association, which have no competing interest with this work. ELF was a member of the National Academy of Sciences, Engineering, and Medicine, Committee on Making ALS a Livable Disease.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-64622-5.

Correspondence and requests for materials should be addressed to Maureen A. Sartor or Eva L. Feldman.

Peer review information *Nature Communications* thanks Karin Danzer, who co-reviewed with Veselin Grozdanovand the other, anonymous, reviewers for their contribution to the peer review of this work. A peer review file is available.