SUPPLEMENTARY INFORMATION for

Gene Expression Signatures from Whole Blood Predict Amyotrophic Lateral Sclerosis Case Status and Survival

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

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI

[2]Department of Biomedical Sciences, University of North Dakota, Grand Forks, ND, USA

[3]Department of Neurology, University of Michigan, Ann Arbor, MI

[4]NeuroNetwork for Emerging Therapies, University of Michigan, Ann Arbor, MI

[5]Department of Preventive Medicine (Biostatistics Division), Northwestern University Chicago, IL

[6]Department of Epidemiology, University of Michigan, Ann Arbor, MI

[7]Neuromuscular Diseases Research Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD

[8]Department of Environmental Health Sciences, University of Michigan, Ann Arbor, MI

[9]Biostatistics Department, University of Michigan, Ann Arbor, MI

*Equal corresponding co-authors

Correspondence to:

Eva L. Feldman, MD, PhD
Department of Neurology
109 Zina Pitcher Place
Ann Arbor, MI 48109
e-mail: efeldman@umich.edu

Maureen A. Sartor, PhD
Department of Computational Medicine and Bioinformatics
100 Washtenaw Ave.
Ann Arbor, MI 48109
e-mail: sartorma@umich.edu

Table of Contents

Supplementary Figure S1. DEGs and pathway overlap of our dataset with literature datasets.

2×2 contingency tables for up- and downregulated DEGs from our dataset of ALS cases (n=422) versus controls (n=272) with (**a**) Swindell et al.[1] and (**b**) Grima et al.[2] datasets, assessed by one-sided Fisher's exact test, with the alternative hypothesis set to "greater" to evaluate whether the overlap is higher than expected from chance. (**c**) KEGG and MSigDB Hallmark pathway enrichment analysis of our DEGs that are consistent and unique with Swindell et al.[1] and Grima et al.[2] literature DEGs.

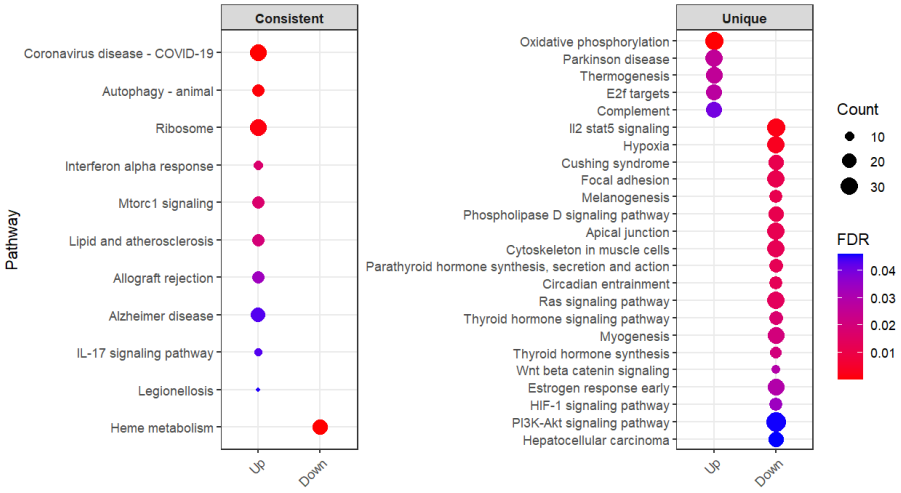
a

7,509 universe genes	**Upregulated in Swindell et al.[1]**	**NOT upregulated in Swindell et al.[1]**
Upregulated in our dataset	381	1,547
NOT upregulated in our dataset	371	5,210
Upregulated Swindell DEGs significantly enriched among our upregulated DEGs. Fisher's exact test: $p=1.31\times10^{-54}$, odds ratio=3.46 (95% confidence interval: 3.03, Inf)		
7,509 universe genes	**Downregulated in Swindell et al.[1]**	**NOT downregulated in Swindell et al.[1]**
Downregulated in our dataset	267	1,357
NOT downregulated in our dataset	497	5,388
Downregulated Swindell DEGs significantly enriched among our downregulated DEGs. Fisher's exact test: $p=2.22\times10^{-19}$, odds ratio 2.13 (95% confidence interval: 1.86, Inf)		

b

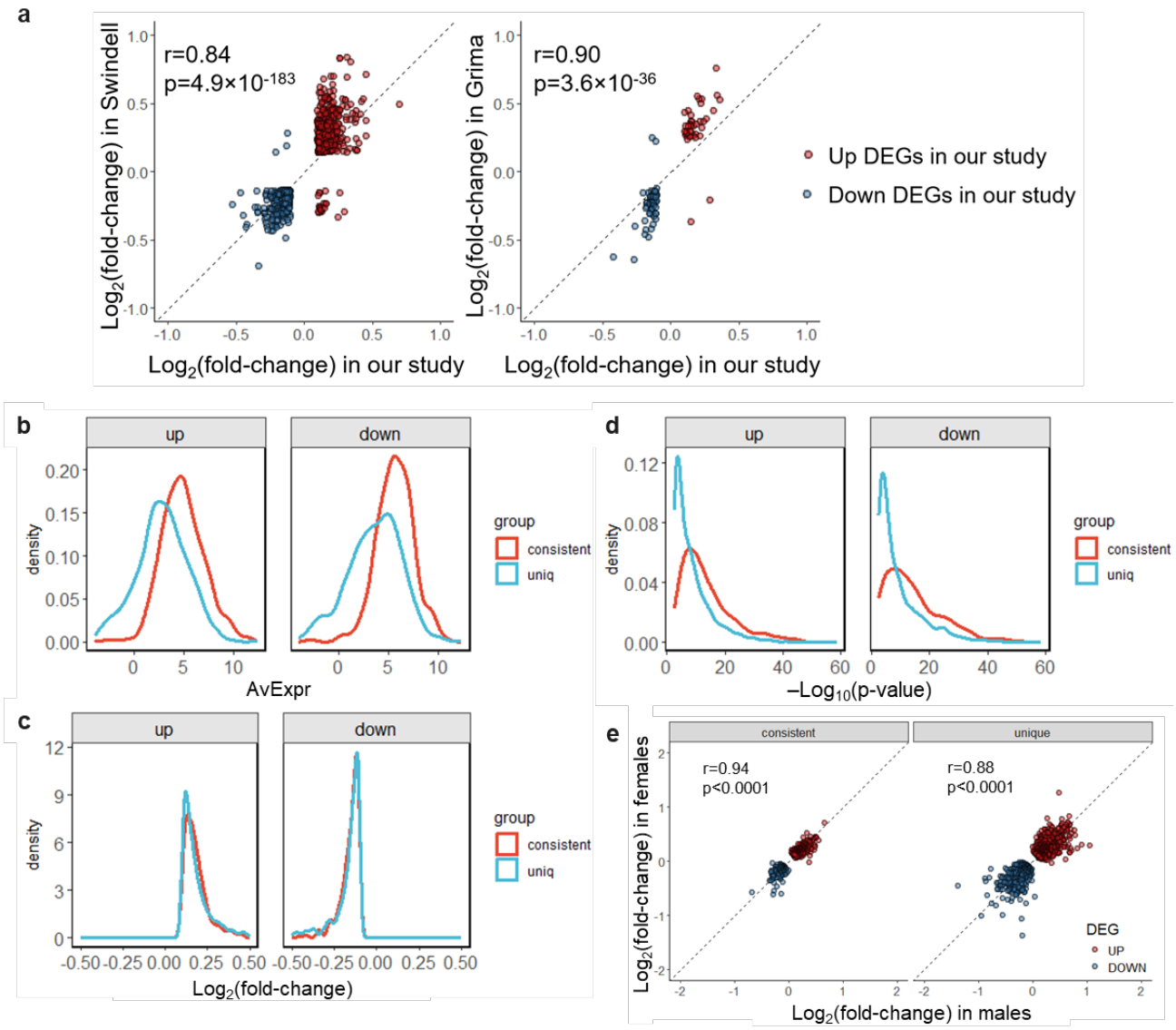
11,867 universe genes	**Upregulated in Grima et al.[2]**	**NOT upregulated in Grima et al.[2]**
Upregulated in our dataset	39	1,889
NOT upregulated in our dataset	46	9,893
Upregulated Grima DEGs significantly enriched among our upregulated DEGs. Fisher's exact test: $p=1.26\times10^{-10}$, odds ratio 4.44 (95% confidence interval: 3.02, Inf)		
11,867 universe genes	**Downregulated in Grima et al.[2]**	**NOT downregulated in Grima et al.[2]**
Downregulated in our dataset	56	1,568
NOT downregulated in our dataset	104	10,139
Downregulated Grima DEGs significantly enriched among our downregulated DEGs. Fisher's exact test: $p=5.71\times10^{-12}$, odds ratio 3.48 (95% confidence interval: 2.60, Inf)		

C



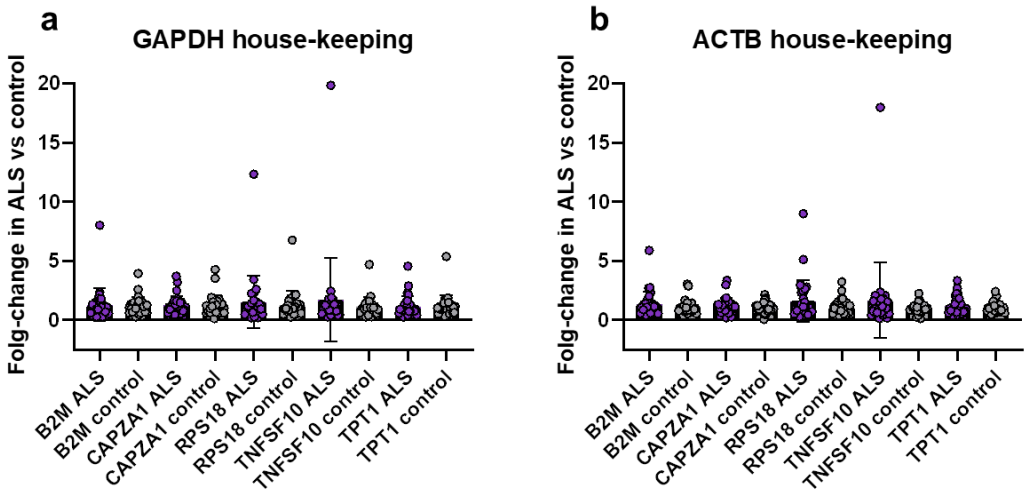
Supplementary Figure S2. Comparison of our blood DEGs to literature DEGs.

(**a**) Scatter plot of log_2(fold-change) in Swindell et al.[1] (left) and Grima et al.[2] (right) DEGs versus log_2(fold-change) in DEGs that overlap in our dataset of ALS cases (n=422) versus controls (n=272). Circle color represents up- and downregulated DEGs; black dashed line represents y=x. Characteristics of up- and down-regulated unique and consistent DEGs in ALS cases versus controls in (**b**) average log_2 gene expression (AvExpr), (**c**) log_2(fold-change), (**d**) $–log_{10}$(p-value). (**e**) Scatter plot of male log_2(fold-change) versus female log_2(fold-change) for unique DEGs (right panel) and consistent DEGs (left panel). Circle color represents up- and downregulated DEGs in overall analysis; black dashed line represents y=x.



Supplementary Figure S3. qPCR validation of five selected upregulated genes in an independent case-control cohort.

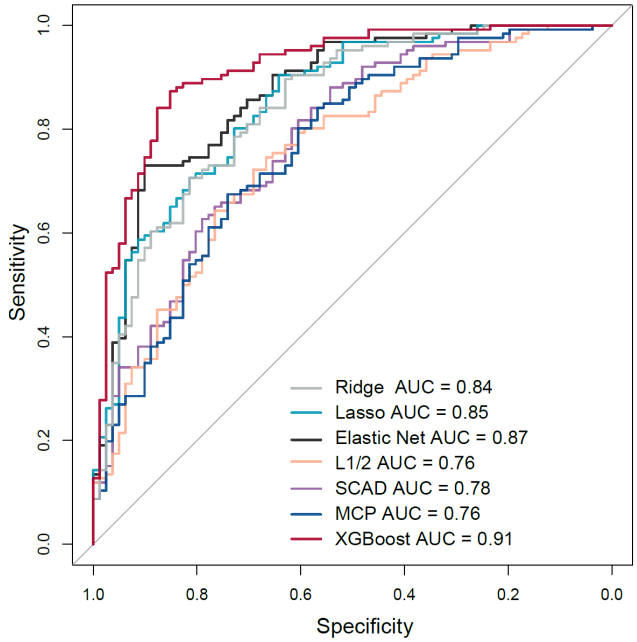
PCR of *B2M*, *CAPZA1*, *RPS18*, *TPT1*, and *TNFSF10* in ALS cases (n=29) versus controls (n=27) using (**a**) GAPDH and (**b**) ACTB as house-keeping genes. (**c**) Summary of fold-changes in the five genes. Data represented by mean ± standard deviation. Source data are associated with this figure.



c

		B2M	*CAPZA1*	*RPS18*	*TNFSF10*	*TPT1*
GAPDH	**ALS**	1.3092	1.2558	1.5385	1.7158	1.1534
	Control	1.1771	1.2267	1.2696	1.1516	1.1867
ACTB	**ALS**	1.3503	1.2208	1.6077	1.6980	1.2039
	Control	1.1124	1.0248	1.1551	0.9936	1.0713

Supplementary Figure S4. Comparison of different classifiers' performance on our internal testing dataset (n=207 total of which n=126 ALS, n=81 controls).



Supplementary Figure S5. XGBoost classifier performance metrics for predicting case status in the external Swindell dataset.

Performance metrics for predicting case status by the three XGBoost-selected gene panels plus the combined panel on the external Rheenan et al.[3]/Swindell et al.[1] datasets. AUC, area under the receiver operating characteristic curves.

	van Rheenen et al.[3]/Swindell et al.[1] (233 ALS cases versus 508 controls, microarray, GSE112676)				van Rheenen et al.[3]/Swindell et al.[1] (164 ALS cases versus 137 controls, microarray, GSE112680)			
	27-gene panel	**30-gene panel**	**29-gene panel**	**Combined 46-gene panel**	**27-gene panel**	**30-gene panel**	**29-gene panel**	**Combined 46-gene panel**
Accuracy	58.97%	64.10%	59.38%	60.73%	62.13%	67.44%	64.78%	64.78%
Balanced accuracy	60.32%	64.76%	61.09%	61.95%	61.99%	67.18%	64.62%	64.62%
Sensitivity	63.95%	66.52%	65.67%	65.24%	63.41%	70.12%	66.46%	66.46%
Specificity	56.69%	62.99%	56.50%	58.66%	60.58%	64.23%	62.77%	62.77%
AUC	65.42%	70.40%	66.54%	67.71%	69.05%	74.65%	70.78%	71.79%
Precision	40.38%	45.19%	40.91%	41.99%	65.82%	70.12%	68.12%	68.12%
F1 Score	49.50%	53.82%	50.41%	51.09%	64.60%	70.12%	67.28%	67.28%

Supplementary Figure S6. Comparison of stepwise, XGBoost, and clinical variables only models for predicting survival.

(**a**) Wilcoxon signed-rank pairwise comparisons of time-dependent ROC-AUCs at each year across 30 random train-test splits for models: stepwise (18-stepwise selected genes + clinical variables), XGBoost (8-XGBoost selected genes + clinical variables), and clinical variables only (clinical variables). (**b**) Time-dependent ROC-AUC performance for our dataset using 30 random train-test splits. (**c**) Log rank adjusted p-values of pairwise comparisons among longer, intermediate, and shorter survival groups based on predictions from the three models using the external Grima et al. dataset.[2] The adjusted p-values were calculated by Bonferroni correction.

a

	Stepwise vs clinical variables only			XGBoost vs clinical variables only			XGBoost vs stepwise		
Year	**ΔAUC***	**95%CI**	**Adjusted p-value#**	**ΔAUC***	**95%CI**	**Adjusted p-value#**	**ΔAUC***	**95%CI**	**Adjusted p-value#**
1	-0.096	(-0.193, 0.036)	0.46	-0.03	(-0.078, 0.012)	0.33	0.059	(-0.066, 0.182)	0.87
2	0.041	(0.021, 0.063)	**0.00057**	0.019	(0.001, 0.037)	0.12	-0.022	(-0.045, 0)	0.15
3	0.031	(0.011, 0.051)	**0.015**	-0.01	(-0.027, 0.006)	0.59	-0.042	(-0.06, -0.024)	**0.00024**
4	0.067	(0.045, 0.088)	**1.50×10^{-5}**	0.025	(0.011, 0.038)	**0.006**	-0.044	(-0.062, -0.023)	**0.00027**
5	0.071	(0.049, 0.09)	**4.30×10^{-6}**	0.038	(0.022, 0.053)	**7.10×10^{-5}**	-0.032	(-0.052, -0.011)	**0.012**
6	0.071	(0.046, 0.097)	**9.70×10^{-6}**	0.053	(0.035, 0.071)	**1.50×10^{-5}**	-0.017	(-0.043, 0.01)	0.69
7	0.08	(0.059, 0.103)	**7.60×10^{-7}**	0.069	(0.049, 0.087)	**1.20×10^{-6}**	-0.011	(-0.032, 0.007)	0.74
8	0.081	(0.058, 0.105)	**7.60×10^{-7}**	0.08	(0.059, 0.101)	**2.40×10^{-7}**	0.001	(-0.021, 0.023)	1

* ΔAUC = mean($AUC_{Group1} - AUC_{Group2}$) for Group 1 versus Group 2; # Adjusted p-values were calculated by Bonferroni correction; statistically significant values (adjusted p-value<0.05) in bold. CI, confidence interval.

b

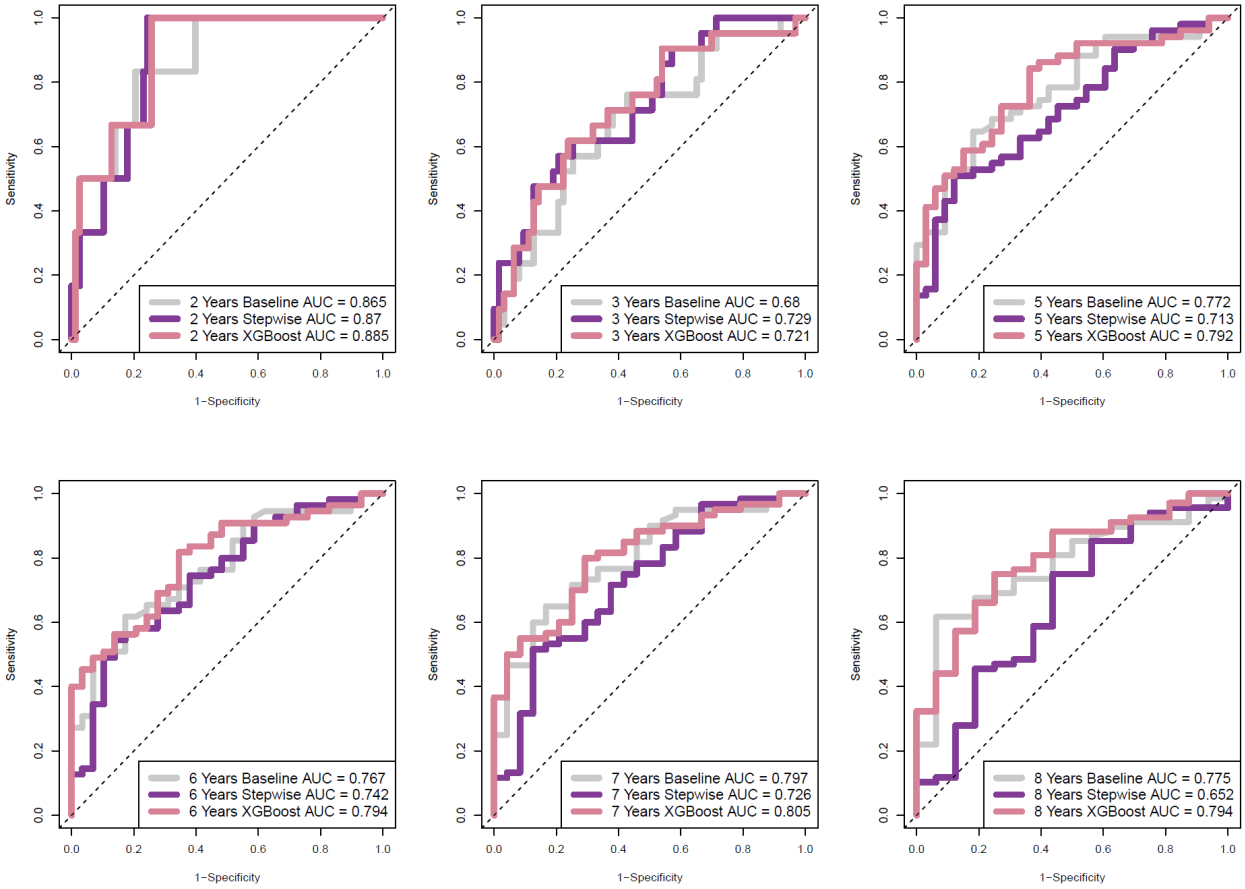
Time	**Clinical variables only model**		**18 stepwise genes + clinical variables**		**8 XGBoost genes + clinical variables**		**18 stepwise genes only**		**10 XGBoost genes only**	
Year	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
2	0.675	0.074	0.717	0.068	0.693	0.067	0.660	0.058	0.637	0.071
3	0.686	0.059	0.716	0.057	0.674	0.059	0.655	0.058	0.582	0.057
4	0.684	0.067	0.748	0.060	0.705	0.067	0.704	0.062	0.624	0.069
5	0.664	0.072	0.731	0.063	0.699	0.069	0.690	0.062	0.634	0.072
6	0.668	0.079	0.736	0.055	0.721	0.080	0.699	0.056	0.673	0.082
7	0.675	0.089	0.754	0.070	0.743	0.087	0.740	0.063	0.734	0.084
8	0.670	0.090	0.750	0.071	0.751	0.083	0.732	0.066	0.750	0.081

c

		Clinical variables only	**Stepwise**	**XGBoost**
Pairwise log-rank test adjusted p-values	Longer versus intermediate	0.036	0.177	0.00068
	Longer versus shorter	0.0003	0.033	1.30×10^{-7}
	Intermediate versus shorter	0.1019	0.261	0.0064
Compare C-index	C-index	0.66	0.65	0.69
	p-value (vs. clinical variables only)		0.79	0.05

Supplementary Figure S7. Additional analyses for predicting ALS survival in the external Grima dataset.

Time-dependent receiver operating characteristic curves (AUCs) for predicting ALS case survival at 2, 3, 5, 6, 7, and 8 years in the external Grima et al. dataset[2] by clinical variables, stepwise, and XGBoost models. AUCs for 4-year survival are in main text **Figure 4e**.



Supplementary Figure S8. Formulas to calculate risk scores for each model.

(**a**) Equations for calculating predicted survival scores by the clinical variables, stepwise, and XGBoost models. (**b**) Survival score table and (**c**) distribution plot for the external Grima et al. dataset.[2]

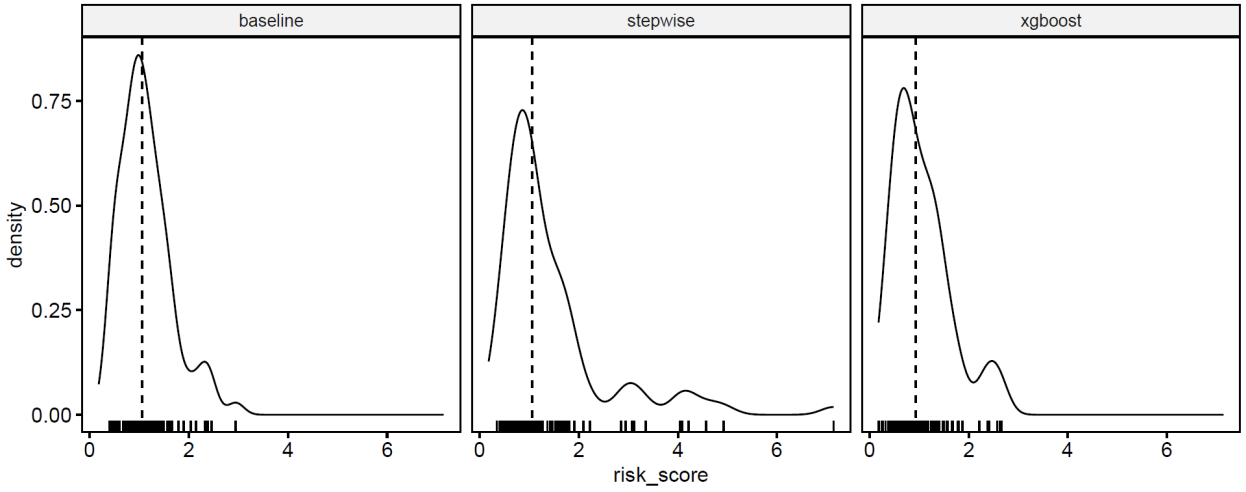
a

- **Clinical variables CoxPH**:
  risk score = exp[0.481 * *onset_segment* + 0.029 * (*onset_age*-61.840) – 0.090 * *sex*]
- **Stepwise CoxPH**:
  risk score = exp[0.641 * *onset_segment* + 0.024 * (*onset_age*-61.840) – 0.018 * *sex* + 1.582*(*FBXL5*- 0.000498) + …]
- **XGBoost CoxPH**:
  risk score = exp[0.463 * *onset_segment* + 0.022 * (*onset_age*-61.840) – 0.289 * *sex* + 0.838*(*BAX*-0.000372) + …]
- Onset_segment is 1 for “bulbar”, is 0 for “not bulbar”.
- Sex is 1 for “male”, is 0 for “female”.

b

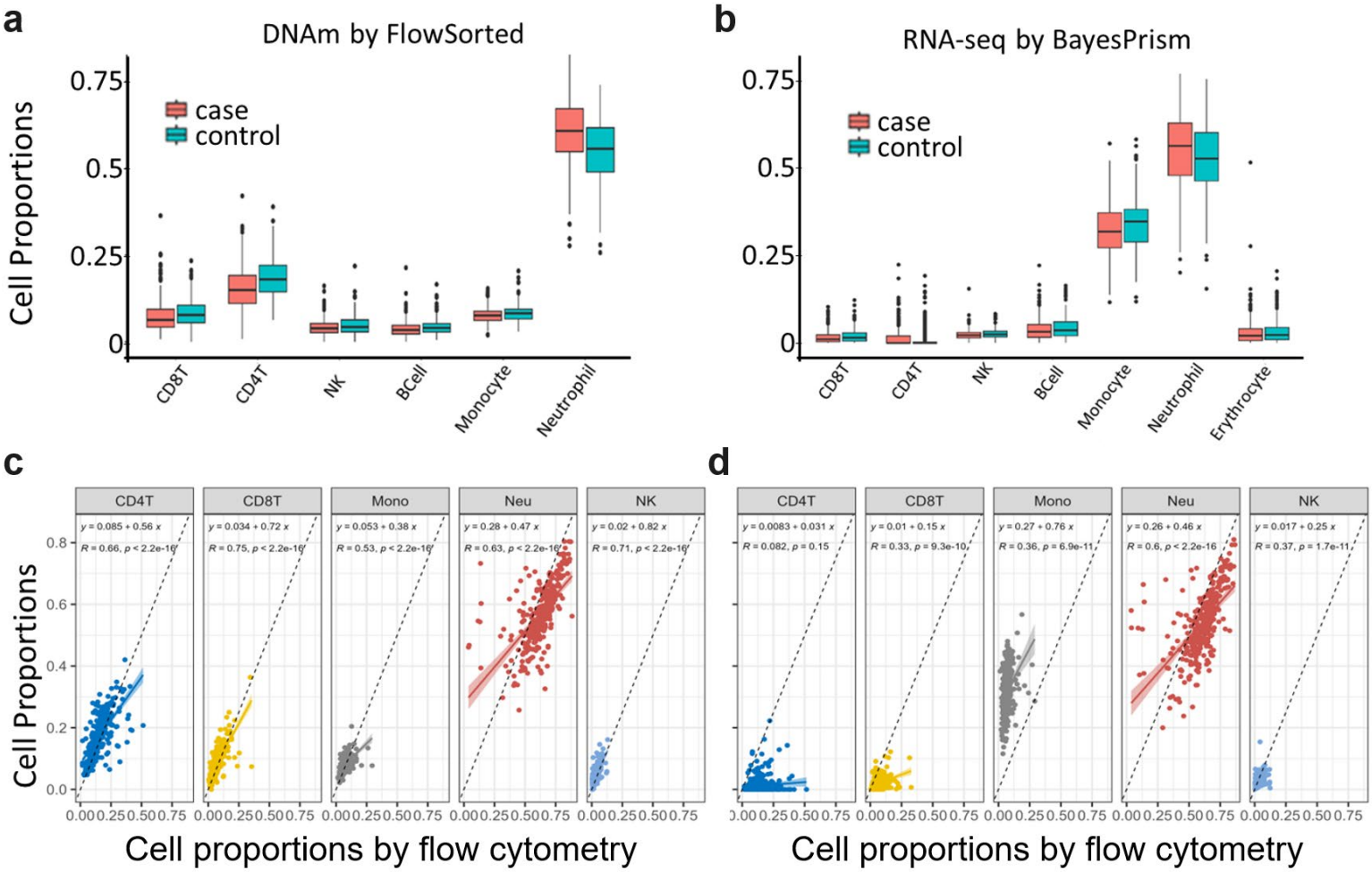
Survival score cutoff	Clinical variables only	Stepwise	XGBoost
Longer survival (bottom 25% in score)	0.807	0.764	0.609
Shorter survival (top 25% in score)	1.397	1.678	1.319

c



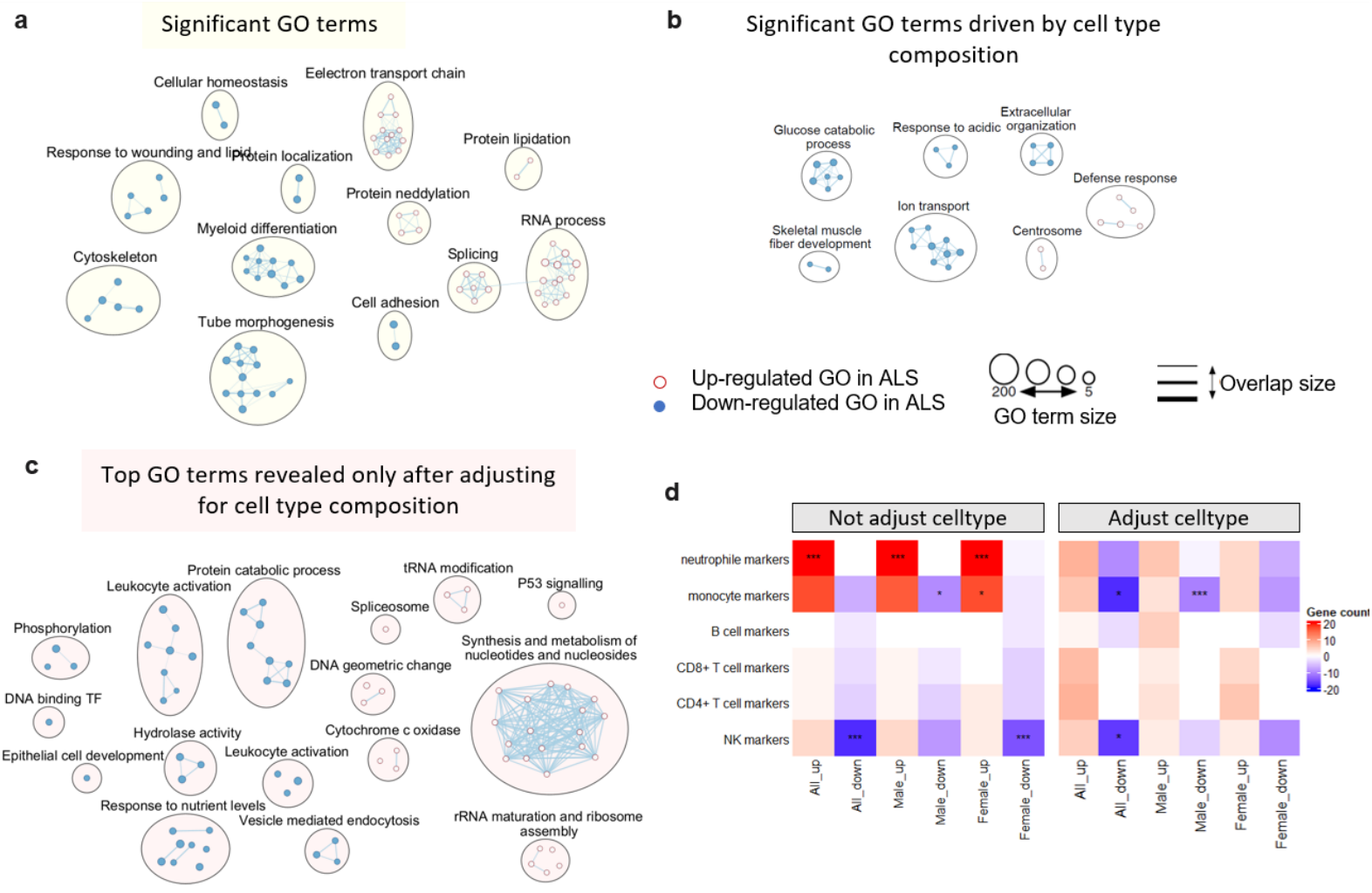
Supplementary Figure S9. Immune cell proportions deconvolution and Pearson's correlation with flow cytometry results (225 ALS cases and 119 controls).

(**a**) Immune cell proportions computed by DNA methylation (DNAm) EPIC data deconvolution. (**b**) Immune cell proportions computed by RNA-seq deconvolution. Data in box plots represented by horizontal line for median, box for first and third quartiles, and whiskers for minimum and maximum values from 1.5 of the interquartile range from the first and third quartiles, respectively. Pearson's correlation of measured cell proportions from flow cytometry and computed cell proportions by (**c**) DNAm and (**d**) RNA-seq. Source data are associated with this figure.



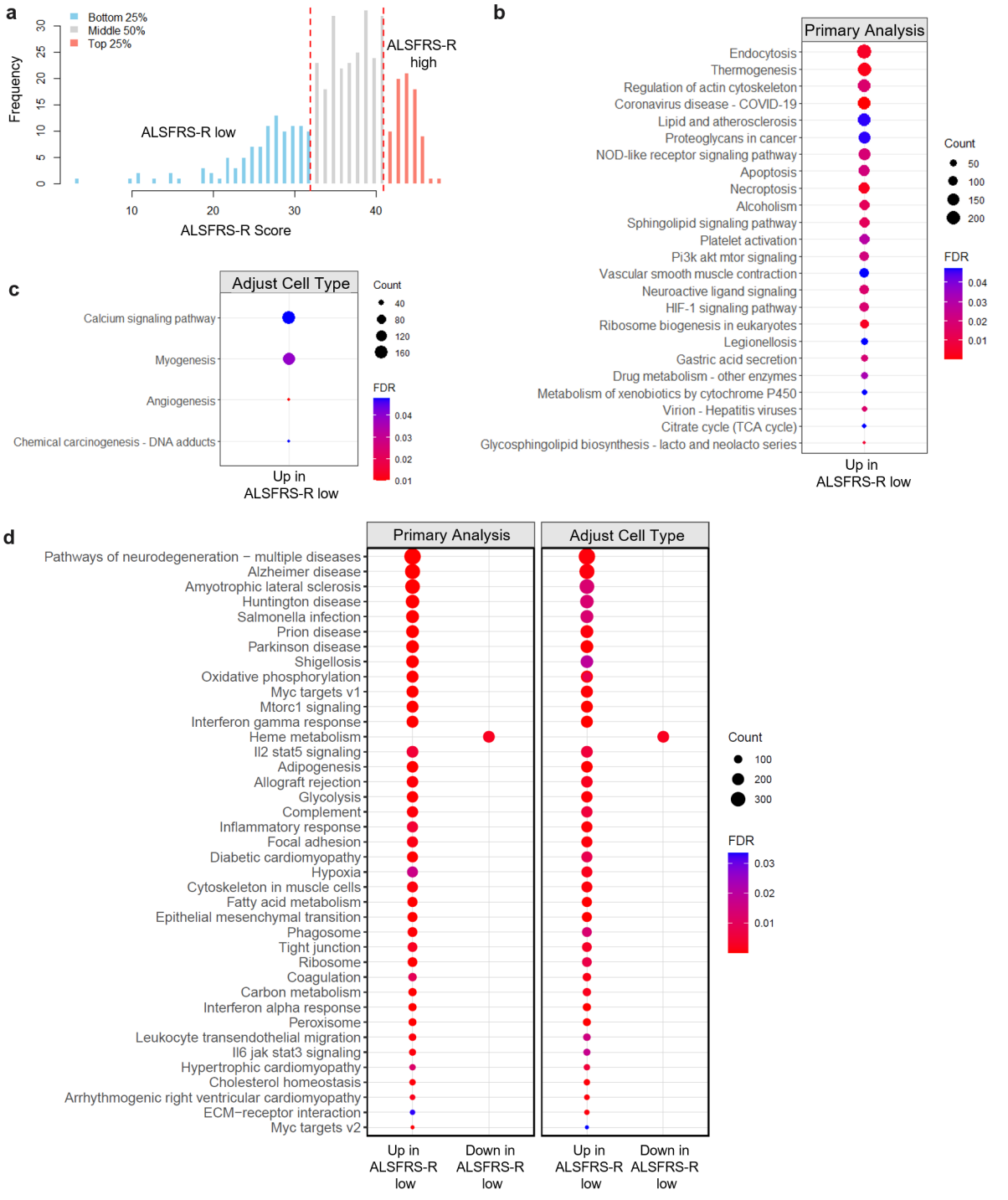
Supplementary Figure S10. Adjusting cell type proportions revealed disease relevant GO terms not driven by cell type composition.

(**a**) Fisher's exact test of significant Gene Ontology (GO) Biological Process (BP) terms that were overrepresented in all ALS cases (n=422) versus controls (n=272) after enrichment analysis of DEGs both with and without cell type composition adjustment. Nodes represent enriched GO BP terms, which are clustered and annotated by their similarity according to related terms. Node size is proportional to the total number of genes within each gene set. Edge thickness is proportional to the number of shared genes between two nodes (i.e., gene sets). Red nodes represent upregulated and blue nodes represent downregulated GO BP terms in ALS. (**b**) The network constructed by the top 40 GO BP terms (ranked by adjusted p-values for up- and downregulated GO BP terms separately) that only appeared in the analysis without adjusting for cell composition. (**c**) The network constructed by the top 40 GO BP terms (ranked by adjusted p-values for up- and downregulated GO BP terms separately) that only appeared in the analysis with adjusting for cell composition. (**d**) Fisher's exact test for the cell type signature gene sets. Color intensity reflect the number of overlapped DEGs and cell markers. Red represents upregulated DEGs, and blue represents down-regulated DEGs. Asterisks reflect the magnitude of Benjamini-Hochberg-adjusted p-values false discovery rate (FDR): *FDR<0.05, ***FDR<0.0001.



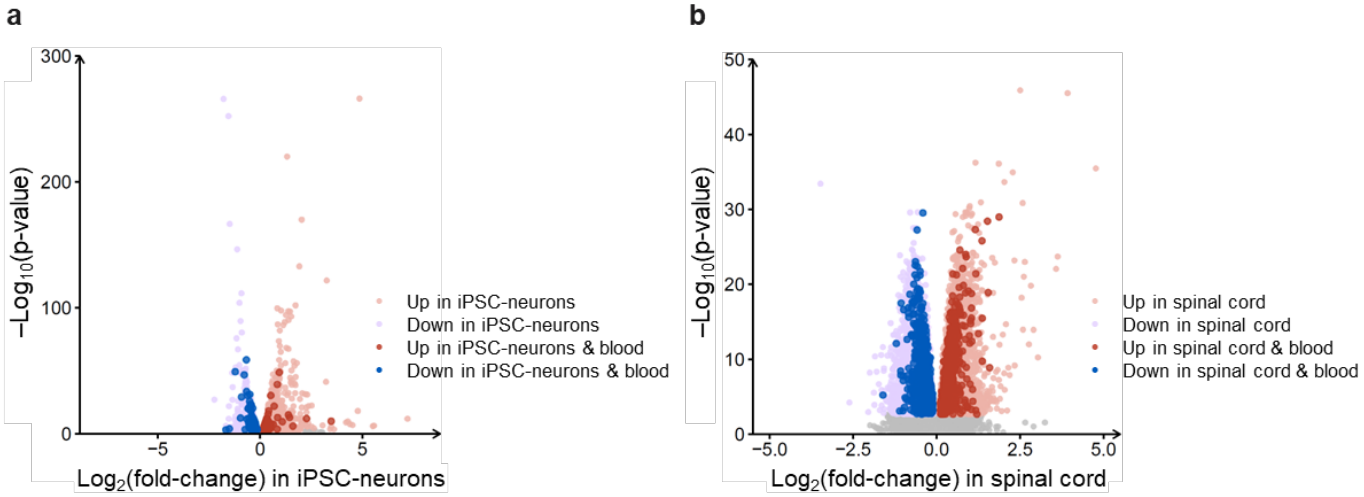
Supplementary Figure S11. Pathway analysis by gene set enrichment analysis of the blood transcriptome in lower versus higher ALSFRS-R scores from ALS participants.

(**a**) Lower ALSFRS-R scores (n=107, ALSFRS-R≤32, bottom 25%) and higher ALSFRS-R scores (n=112, ALSFRS-R≥41, top 25%) from ALS participants based on ALSFRS-R score at first clinic visit. Pathways unique to enrichment analysis of (**b**) primary DEGs and (**c**) DEGs adjusted for cell type proportions. (**d**) Top 30 pathways shared by enrichment analysis of both primary DEGs and (**c**) DEGs adjusted for cell type proportions.



Supplementary Figure S12. Overlap of blood DEGs with iPSC-neuron and spinal cord DEGs.

Volcano plot of our common blood DEGs shared with (**a**) iPSC-neurons with TDP-43 knockdown and (**b**) postmortem spinal cord DEGs highlighted in darker red (upregulated) or blue (downregulated).



Supplementary Figure S13. Revised El Escorial criteria at diagnosis does not affect case prediction in our internal training data by 10-fold cross validation.

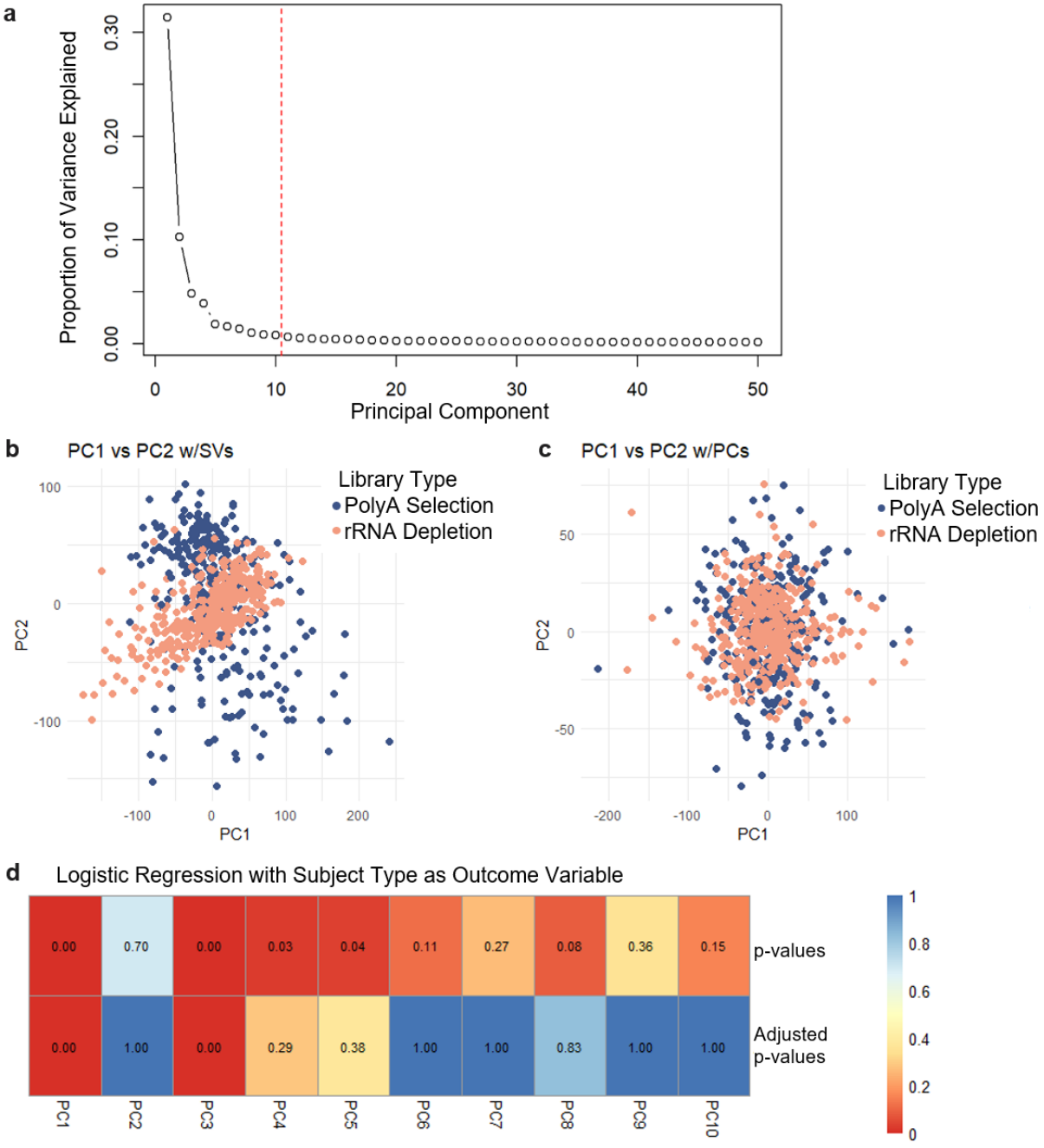
Cases with “definite”, “probable”, and “probable, lab supported” revised El Escorial criteria at diagnosis were assigned as Group 1, and cases with “possible” and “suspected” criteria as Group 2. Fisher’s exact test results show that incorrect classifications were not enriched in Group 2 versus Group 1 for any of the three 27-, 30-, and 29-gene panel XGBoost classifiers.

		27-gene classifier		30-gene classifier		29-gene classifier		
	El Escorial criteria at diagnosis	**Incorrect**	**Correct**	**Incorrect**	**Correct**	**Incorrect**	**Correct**	**Total**
Group1	**Definite**	5	105	6	104	5	105	110
	Probable	2	124	8	118	7	119	126
	Probable, lab supported	8	96	9	95	9	95	104
Group2	**Possible**	2	52	2	52	1	53	54
	Suspected	2	20	1	21	2	20	22
	All	19	397	26	390	24	392	416
	Fisher’s exact test p-value	0.47		0.89		0.85		
	Fisher’s exact test odds ratio	0.83 (95%CI: 0, 2.80)		1.76 (95%CI: 0, 7.09)		1.60 (95%CI: 0, 6.47)		

CI, confidence interval.

Supplementary Figure S14. Selection of principal components for differential gene expression analysis.

(**a**) Elbow plot of the first 50 principal components (PCs). Principal component analysis (PCA) of gene expression residuals from (**b**) differential expression analysis linear model with surrogate variables (SVs); (**c**) differential expression analysis linear model with PCs. Each dot is a sample colored by library preparation types, polyA selection, rRNA depletion. (**d**) Logistic regression with case-control status (i.e., subject_type) as a response variable, each PC as a predictor, with the other model covariates that included age, sex, genetic PCs (1-4), cell type proportions, batch, and library types.



Supplementary Figure S15. Correlation of RNA principal components to cell type proportions.

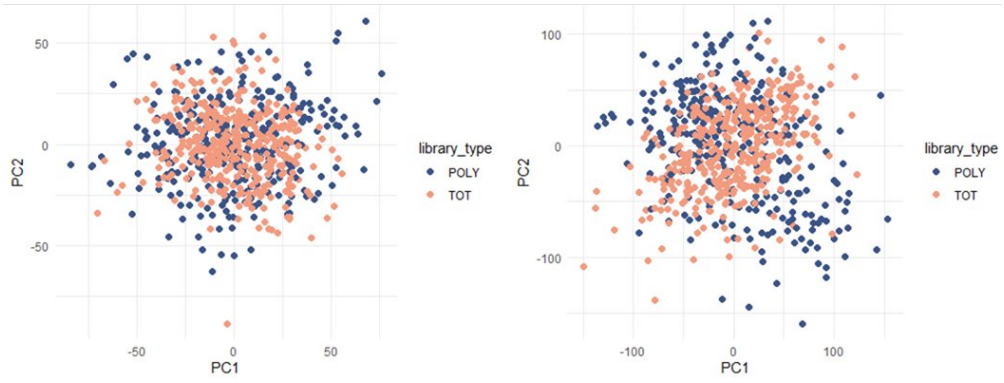
(**a**) Heatmap of Pearson's correlation results between RNA PCs and cell type proportions. Legend and box colors indicate Pearson's correlation coefficients; box numbers indicate Bonferroni-adjusted p-values. (**b**) RNA PCs that correlate to cell type proportions. (**c**) Principal component analysis (PCA) of gene expression residuals from (left) differential expression analysis linear model with RNA $PC_{2\text{-}10}$ and (right) differential expression analysis linear model with RNA PC_2, $PC_{5\text{-}8}$, and PC_{10}. Each dot is a sample colored by library preparation types, polyA selection, rRNA depletion.

a

b

Principal components	Cell types	r	Adjusted p-values
PC_3	Neutrophils	0.602098	5.56E-68
PC_3	B cells	-0.35685	1.43E-20
PC_3	CD8 T cells	-0.31906	3.46E-16
PC_4	Neutrophils	-0.28032	2.68E-12
PC_9	Monocytes	0.386587	1.85E-24
PC_9	NK cells	0.304874	1.07E-14
PC_9	CD8 T cells	0.272732	1.33E-11

c **Primary analysis with $PC_{2\text{-}10}$** **Primary analysis with $PC_{2,5\text{-}8,10}$**



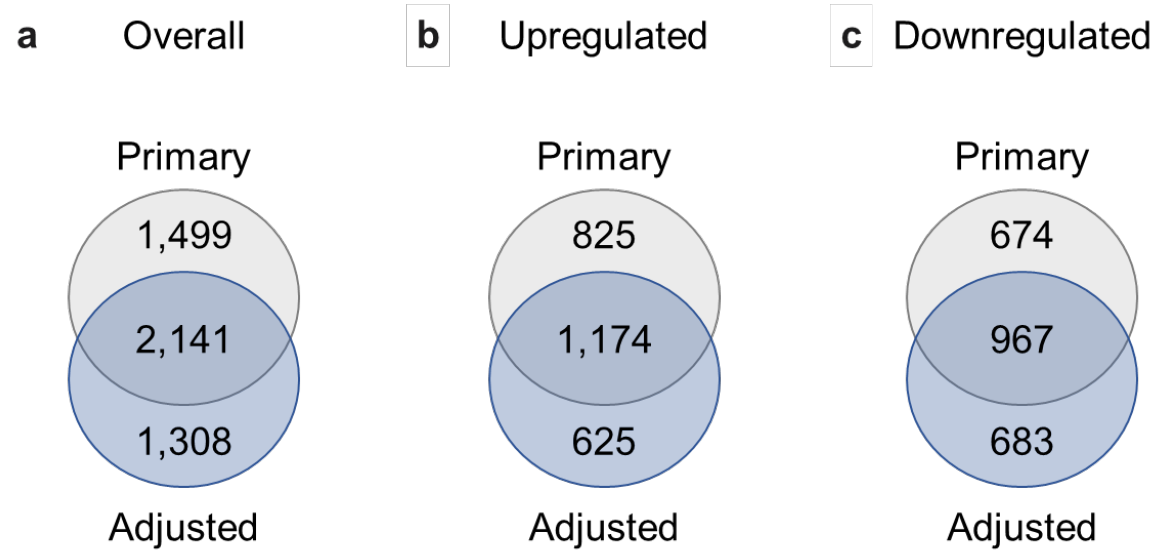
Supplementary Figure S16. Overlap of DEGs in ALS cases (n=422) versus control (n=272) blood samples in primary analyses and analyses adjusted for cell proportions.

Venn diagram of overlap between Ensembl IDs of primary DEGs and DEGs adjusted for cell proportions (**a**) overall, (**b**) upregulated, and (**c**) downregulated. Overlap quantitatively assessed by Fisher's exact tests with calculated overlap scores (Jaccard index) separately for upregulated, downregulated, and all DEGs.

All DEGs (up + down): Fisher's exact test showed a significant enrichment ($p<2.2x10^{-16}$, odds ratio 19.0 [95%CI: 17.39, 20.73]) with a Jaccard index of 0.43.

Upregulated DEGs: Fisher's exact test showed a significant enrichment ($p<2.2x10^{-16}$, odds ratio 44.8 [95%CI: 39.76, 50.65]) with a Jaccard index of 0.45.

Downregulated DEGs: Fisher's exact test showed a significant enrichment ($p<2.2x10^{-16}$, odds ratio 42.0 [95%CI: 37.06, 47.74]) with a Jaccard index of 0.42.



References

1. Swindell WR, Kruse CPS, List EO, Berryman DE, Kopchick JJ. ALS blood expression profiling identifies new biomarkers, patient subgroups, and evidence for neutrophilia and hypoxia. *J Transl Med*. May 22 2019;17(1):170. doi:10.1186/s12967-019-1909-0
2. Grima N, Liu S, Southwood D, et al. RNA sequencing of peripheral blood in amyotrophic lateral sclerosis reveals distinct molecular subtypes: Considerations for biomarker discovery. *Neuropathol Appl Neurobiol*. Dec 2023;49(6):e12943. doi:10.1111/nan.12943
3. van Rheenen W, Diekstra FP, Harschnitz O, et al. Whole blood transcriptome analysis in amyotrophic lateral sclerosis: A biomarker study. *PLoS One*. 2018;13(6):e0198874. doi:10.1371/journal.pone.0198874