

Disease related changes in ATAC-seq of iPSC-derived motor neuron lines from ALS patients and controls

Received: 5 January 2024

Accepted: 9 April 2024

Published online: 02 May 2024

 Check for updates

Stanislav Tsitkov¹, Kelsey Valentine¹, Velina Kozareva¹, Aneesh Donde¹, Aaron Frank², Susan Lei², the Answer ALS Consortium*, Jennifer E. Van Eyk³, Steve Finkbeiner^{4,5,6}, Jeffrey D. Rothstein^{7,8}, Leslie M. Thompson^{9,10,11,12}, Dhruv Sareen^{2,13}, Clive N. Svendsen¹³ & Ernest Fraenkel¹ ✉

Amyotrophic Lateral Sclerosis (ALS), like many other neurodegenerative diseases, is highly heritable, but with only a small fraction of cases explained by monogenic disease alleles. To better understand sporadic ALS, we report epigenomic profiles, as measured by ATAC-seq, of motor neuron cultures derived from a diverse group of 380 ALS patients and 80 healthy controls. We find that chromatin accessibility is heavily influenced by sex, the iPSC cell type of origin, ancestry, and the inherent variance arising from sequencing. Once these covariates are corrected for, we are able to identify ALS-specific signals in the data. Additionally, we find that the ATAC-seq data is able to predict ALS disease progression rates with similar accuracy to methods based on biomarkers and clinical status. These results suggest that iPSC-derived motor neurons recapitulate important disease-relevant epigenomic changes.

Amyotrophic lateral sclerosis (ALS)¹ is a neurodegenerative disorder characterized by motor neuron loss. Its heritability has been estimated to be as high as 50%², but the known genetic factors account for less than 15% of cases. One possible explanation for the missing genetic component is that many diverse genetic causes lead to similar disruptions in pathways that are then exacerbated by non-genetic factors. Disease models based on induced pluripotent stem cell (iPSC) derived motor neurons generated from a broad cross-section of ALS patients may help identify such convergent, early effects. In this study, we examine the epigenomic profiles of more than five hundred cell cultures of iPSC-derived motor neurons (iMNs) generated from ALS

patients and healthy controls to test for the presence of genetically driven, disease-relevant changes in chromatin accessibility and dys-regulated transcriptional programs.

Epigenetics is an especially relevant level at which to look for genetically encoded ALS-specific impact in these cells. Changes in chromatin accessibility are generally attributed to the binding of pioneer transcription factors³, DNA methylation, chromatin remodeling complexes, and histone post translational modifications (PTMs)⁴. Previous research has implicated several of these mechanisms in ALS pathology. For example, post-mortem spinal cord tissue from ALS patients exhibited elevated levels of the DNA methyltransferases

¹Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. ²Cedars-Sinai Biomanufacturing Center, Cedars-Sinai Medical Center, Los Angeles, CA, USA. ³Advanced Clinical Biosystems Research Institute, Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. ⁴Center for Systems and Therapeutics, Gladstone Institutes, San Francisco, CA, USA. ⁵Taube/Koret Center for Neurodegenerative Disease, Gladstone Institutes, San Francisco, CA, USA. ⁶Departments of Neurology and Physiology, University of California, San Francisco, San Francisco, CA, USA. ⁷Brain Science Institute, Johns Hopkins University School of Medicine, Baltimore, MD, USA. ⁸Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD, USA. ⁹Department of Neurobiology and Behavior, University of California, Irvine, CA, USA. ¹⁰Department of Psychiatry and Human Behavior, University of California, Irvine, CA, USA. ¹¹Institute for Memory Impairments and Neurological Disorders, University of California, Irvine, CA, USA. ¹²Sue and Bill Gross Stem Cell Center, University of California, Irvine, CA, USA. ¹³The Board of Governors Regenerative Medicine Institute and Department of Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, CA, USA. *A list of authors and their affiliations appears at the end of the paper.

✉ e-mail: fraenkel-admin@mit.edu

DNMT1 and DNMT3A compared to controls⁵. Motor neurons expressing FUS and TDP43 mutants exhibited a loss of subunits of the neuronal Brg1/Brm Associated Factor chromatin remodeling complex⁶. Changes in the expression of the ALS genes *FUS*, *TDP43*, and *C9orf72* were found to be associated with changes in histone PTMs⁷. Histone deacetylase inhibitors have even been proposed as a potential therapeutic for ALS⁸. The identification of other ALS-specific epigenetic signatures will improve our understanding of early disease mechanisms and may suggest new therapeutic strategies. Because iPSCs undergo epigenetic reprogramming⁹, the environmental contributions to ALS are likely to have been erased. As such, iPSC-derived cells allow a direct examination of the impact of as yet uncharacterized genetic factors on the epigenome.

The main problem in the identification of epigenetic signatures associated with ALS pathology is the heterogeneity in the genetic and clinical manifestations of the disease¹⁰, and the scarcity of ALS patient-derived neuronal tissue. To address these problems, the Answer ALS consortium (AALS) is generating iPSC lines from the peripheral blood mononuclear cells (PBMCs) of over 800 ALS patients and 200 healthy controls that have been whole genome sequenced¹¹. The iPSC lines generated by AALS are differentiated into motor neurons (iMNs) and subjected to epigenomic, transcriptomic, and proteomic analysis. The advantage of iPSCs is that they can be generated from patient blood samples, grown in large quantities, and differentiated into disease-affected cell types¹². iPSC models of ALS have previously been used to characterize phenotypic patterns of neurodegeneration in mid-size cohorts of sporadic ALS patient iPSC-derived motor neurons derived from a population of Japanese ALS and control subjects¹³, and to construct disease-associated protein-protein interaction networks for ALS cases associated with the mutant C9orf72 hexanucleotide repeat expansion¹⁴. iPSCs are being used to model many other neurodegenerative diseases in smaller-scale studies and through large initiatives such as FOUNDIN-PD¹⁵, iNDI¹², among others^{12,16–25}.

Important technical challenges arise in studies of this scale, which necessarily have many sources of variation. The standard approach for analyzing omics data uses differential analysis with case status (ALS or healthy control) as the primary covariate. Such an approach is inappropriate in this setting. For example, sex imbalances in case/control groups can lead to false positive differential signals associated with the sex chromosomes. While sex is a covariate that can, in theory, be controlled for by meticulous study design or adjusted for in analysis, other covariates cannot be handled in these ways. Differentiated cell type composition, for example, was found to be the main driver of variation in AALS iMN gene expression, and it is not known until after the data are analyzed²⁶. The identification of robust ALS-associated signals requires a thorough understanding of sources of variation associated with sequencing, differentiation and clinical parameters.

In this study, we identify covariates that drive variation in the epigenomic profiles of 533 iPSC-derived motor neurons from ALS patients and healthy controls as measured by the Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq)²⁷. This is one of the largest bulk ATAC-seq datasets generated by a single consortium (by bases sequenced), and the largest bulk ATAC-seq dataset overall from cell cultures of different donors using a single differentiation-protocol. The size of this dataset combined with the consistency of the data generation protocols allows it to be used as a tool to both investigate ALS-specific epigenomic signals, and establish practices for analyzing other ATAC-seq datasets. As might be expected from a study of a diverse population, conducted over several years, parameters such as sex, cell type composition, and sequencing efficiency drive much of the overall variance. Initially, we do not find any changes in chromatin accessibility when comparing all familial and sporadic ALS cases against controls – a finding that is also consistent with the variable etiology and phenotypes of ALS and in line with our

previous study analyzing RNA-seq data in the identical patient groups. However, once these factors are accounted for, a strong differential signal is seen when stratifying by patients carrying the C9orf72 mutant hexanucleotide repeat expansion, a major risk factor for familial ALS. Surprisingly, we also find that the ATAC-seq data can be used to predict ALS progression rates at levels similar to clinical and neurofilament data. These results demonstrate that the chromatin accessibility of iPSC-derived motor neurons can reflect both genetic and clinical variation in ALS.

Results

ATAC-seq data were generated for 533 iPSC-derived motor neuron lines

ATAC-seq was conducted on 533 differentiated motor neuron lines from 460 unique donors (380 ALS patients and 80 healthy controls); 73 samples correspond to study design controls. The production of the motor neuron lines is described in detail in Baxi et al.¹¹ and illustrated in brief in Fig. 1a. Blood samples were collected from ALS patients and healthy controls. PBMCs, classified as either T-cells or non-T-cells (monocytes), were isolated from the blood samples and reprogrammed into iPSCs. The iPSCs were differentiated into motor neurons, and the resulting cell cultures were frozen and distributed across sequencing facilities. Variation between differentiation and sequencing batches was controlled for by including batch differentiation controls (BDCs) and batch technical controls (BTCs), and staining differentiated cell cultures with immunocytochemical (ICC) staining markers (Fig. 1b, SI Section 1).

Evaluation of ATAC-seq data quality

Overall, ATAC-seq-specific alignment QC metrics satisfied ENCODE guidelines (see Figures S1a–c, Methods)²⁸. Annotations of peaksets of individual samples did not exhibit significant heterogeneities (Figure S1d), and the consensus peak set contained 100,363 chromatin regions, of which approximately 80% were intronic/intergenic, and 10% were promoters/5' UTRs (Fig. 1c). To evaluate the quality of the ATAC-seq data as it pertains to motor neurons, we examined the chromatin accessibility of genes specific to motor neurons as done previously by Sahinyan et al.²⁹ Chromatin accessibility was assessed for the housekeeping gene, *GAPDH*, a set of spinal motor neuron-specific genes (*LHX3*, *ISL1*, *ISL2*, *MNX1*)^{30,31}, and as a negative control, the pluripotency marker *POU5F1*^{29,32}. As expected, out of the six genes tested, only the promoter/TSS chromatin region for *POU5F1* was not accessible (Fig. 1d, e, S1e). Data reproducibility was confirmed by the assessment of inter-sample correlations and comparisons to re-differentiated samples (SI Section 2).

Variance in most variably-accessible regions is driven by three sources of variation

To investigate the underlying factors contributing to variance in the data, we conducted principal component analysis (PCA) (Fig. 2a, Figure S3a) and found three major sources of variance: sex, iPSC cell type of origin (PBMC/T-cell or PBMC/non-T-cell), and sequencing instrument (SI Section 3). In fact, we found that applying UMAP to the top 100 most variably-accessible regions separated samples into four distinct clusters defined by sex and PBMC type (Fig. 2b). We used this UMAP representation to estimate the PBMC type for samples with missing PBMC labels in downstream analyses. Additionally, we noted that the BTC/BDC samples completely separated from the remainder of the female samples (Fig. 2a) and found that the separation is partially driven by genomic variants specific to the BTC/BDC samples; for example, a genomic structural variant characterized by a 2 kb deletion dramatically affects chromatin accessibility in BTC/BDC samples (SI Section 4). Overall, motif enrichment analysis of the most variably accessible regions revealed enrichment for neuronal transcription factors (SI Section 5).

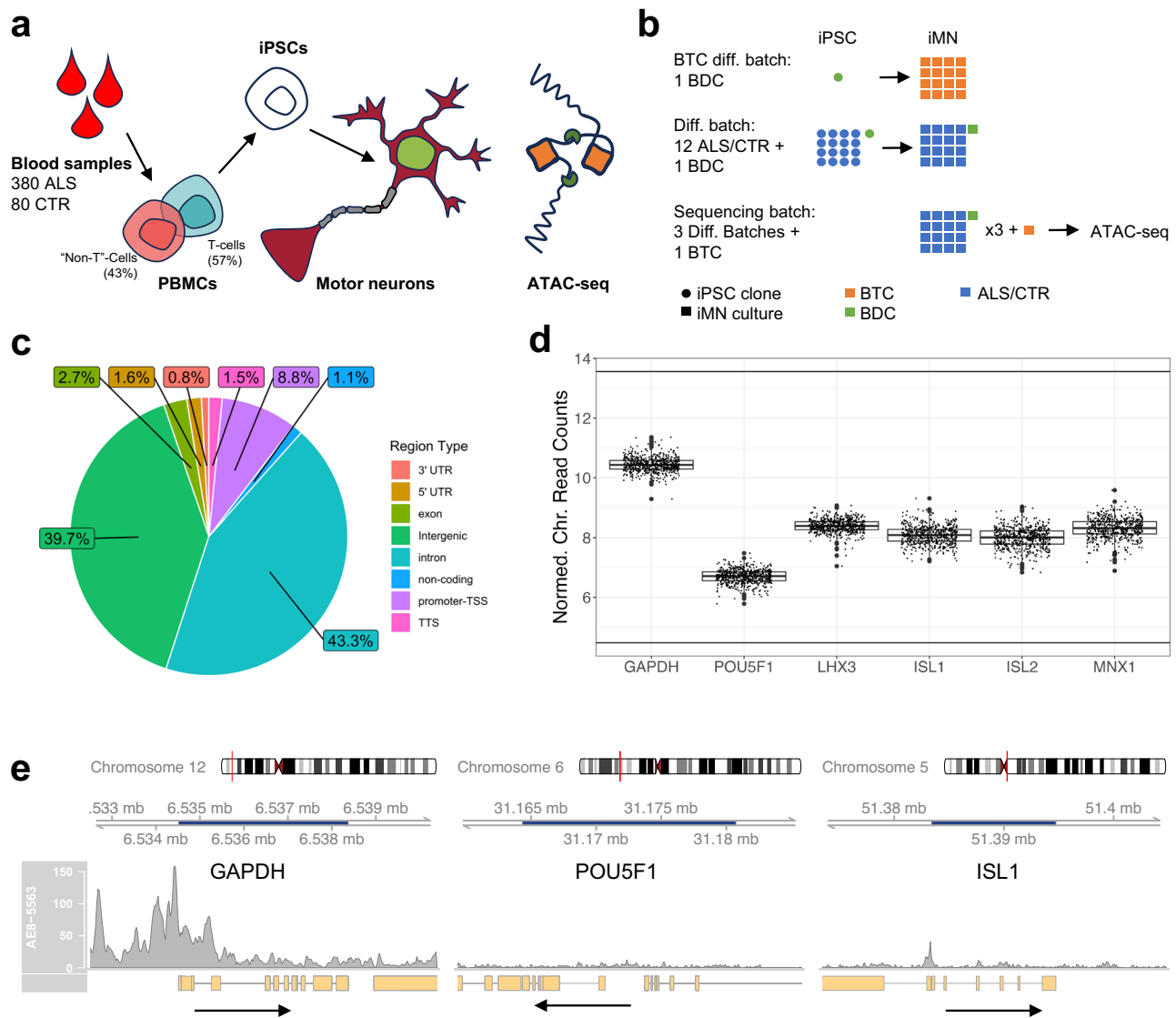


Fig. 1 | Answer ALS ATAC-seq data. **a** Overview of AALS data generation protocol. PBMCs from ALS patients and healthy controls are reprogrammed into iPSCs, which are in turn differentiated into motor neurons and sent for sequencing. **b** Overview of study design controls. Samples are divided into differentiation batches and sequencing batches. Each sequencing batch usually consists of three differentiation batches. A BDC is redifferentiated with each differentiation batch, and a BTC is resequenced with each sequencing batch. **c** Pie chart showing distribution of region annotations. **d** Normalized chromatin reads plotted for the promoter/TSS for the housekeeping gene, *GAPDH*, the pluripotency marker, *POU5F1*, and the spinal motor neuron-specific genes *LHX3*, *ISL1*, *ISL2*, and *MNX1* ($n = 533$). Boxplot boxes indicate the 25th, 50th (median), and 75th quartiles; boxplot whiskers extend 1.5 interquartile ranges from the median. **e** Raw read coverage plots spanning the gene bodies of *GAPDH*, *POU5F1*, and *ISL1*. Dark blue shading of the genome axis scale indicates the gene body location. All plots are drawn on the same scale. Arrows point in the direction of transcription.

To explore the effects of a more comprehensive set of covariates, we additionally fit the chromatin accessibility of each region to a set of 17 covariates using a linear mixed effects model (Fig. 2c). These covariates reflected the sequencing, differentiation, clinical, and demographic aspects of each sample. Differentiation batch explained the second-most variance across all regions, but with few regions explaining over 25% of the variance, indicating a small effect size. The association of several regions with the sequencing-associated covariates, Fraction of Reads in Peaks (FRiP) score and sequencer, could be explained by the normalization methods used and changes in raw read length (SI Section 6). The ICC staining markers contributed to the variation of the samples in a manner similar to that found in the gene expression data, with the most variance driven by *S100B* and *Nestin*; however, these markers had a small effect size overall, explaining less than 25% of the variance for all but 32 and 16 regions, respectively. As expected from the PCA analysis, the variability of

several regions was driven by sex, PBMC type, and sequencer. Interestingly, we also found a dependence on ancestry, which was not observed in gene expression²⁶.

In order to account for the variance contribution of the identified covariates and avoid false positive differential signals, we opted to include FRiP score, sex, PBMC type, and sequencer as covariates in all downstream differential analyses. Indeed, regressing out these covariates using a linear model removes the separations in plots of principal components of the most variably accessible regions (SI Section 3, Figures S3e, f). Notably, there was no global differential signal associated with ALS case status at a strict significance threshold (Bonferroni adj. p -value < 0.01), as might be expected for ALS, which is an extremely heterogeneous disease. However, there were extensive differential signals associated with the remaining covariates (Supplementary Data SI-5). As a final QC, we confirmed that motor neuron identity of the cell lines was not compromised by sex (Supplementary

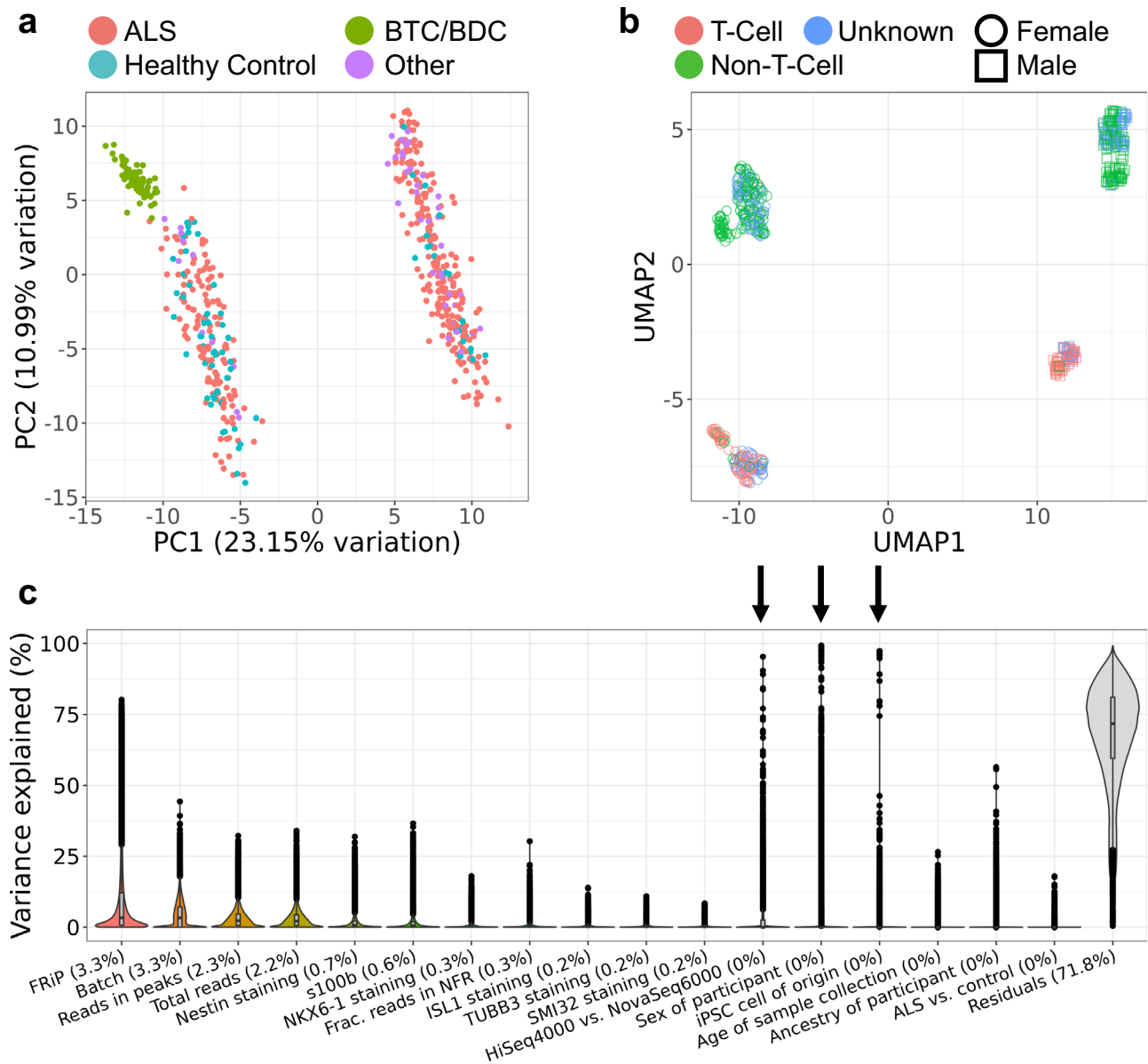


Fig. 2 | Drivers of variation in chromatin accessibility. **a** Biplot of PC1 and PC2 from principal component analysis on the top 500 most variably-accessible regions including all samples. “Other” refers to samples from individuals with non-ALS motor neuron disease and asymptomatic ALS. **b** UMAP applied to 100 most variably-accessible regions separates samples into clusters by sex and PBMC type. **c** Explained variance in chromatin accessibility by selected covariates across all

100,363 chromatin regions. Data was generated by fitting a linear mixed effects model to normalized chromatin reads for each chromatin region (see Methods). Percentages indicate median contribution to variance. Arrows indicate covariates that were found to drive variation in the PCA of the 500 most variably-accessible regions. Boxplot boxes indicate the 25th, 50th (median), and 75th quartiles; boxplot whiskers extend 1.5 interquartile ranges from the median.

Data S2) and PBMC type (Supplementary Data S3), as neither of these covariates were significantly associated with changes in the accessibility of the motor neuron-specific genes examined in Fig. 1d. Additionally, we searched for associations between the chromatin accessibilities of ALS gene-associated promoter regions and sample sex and PBMC type. Out of the 16 regions examined (Figure S6), only the promoter for the X chromosomal gene UBQLN2 was significantly associated with sex (B-H adj. p -value = $2e-8$, Supplementary Data S2).

Covariates associated with differentiation

It was interesting that unsupervised clustering separated samples along PBMC type (T-cell or non-T-cell; Fig. 2b); in the RNA-seq data, this was only observed when the gene set was constrained to four T-cell receptor associated genes²⁶. Examining DARs associated with PBMC type, we found that regions located next to the T-cell receptor delta

anti-sense 1 (*TRD-ASI*) gene were dramatically less accessible in T-cell derived cell lines as would be expected (Fig. 3a, b) due to T-cell receptor rearrangements. Beyond the *TRD-ASI* regions, there are 180 DARs associated with PBMC type (adj. p -value < 0.01, abs(log2FC) > 0.5); 25 of these regions are annotated as promoters. The top 5 significant promoter regions, other than those that correspond to pseudogenes or lincRNAs, are labeled in Fig. 3a. To confirm that the differential signal was not a sequencing artifact, we compared the accessibility of these regions to the gene expression for matched samples, and found high correlations (Fig. 3c–f). The existence of genes not associated with the T-Cell receptor loci whose promoter accessibility and expression are dependent on PBMC type illustrates a modest, but detectable, impact of epigenetic memory in our dataset: the chromatin accessibility and gene expression profiles of differentiated cell cultures depend on the initial PBMC type. To determine

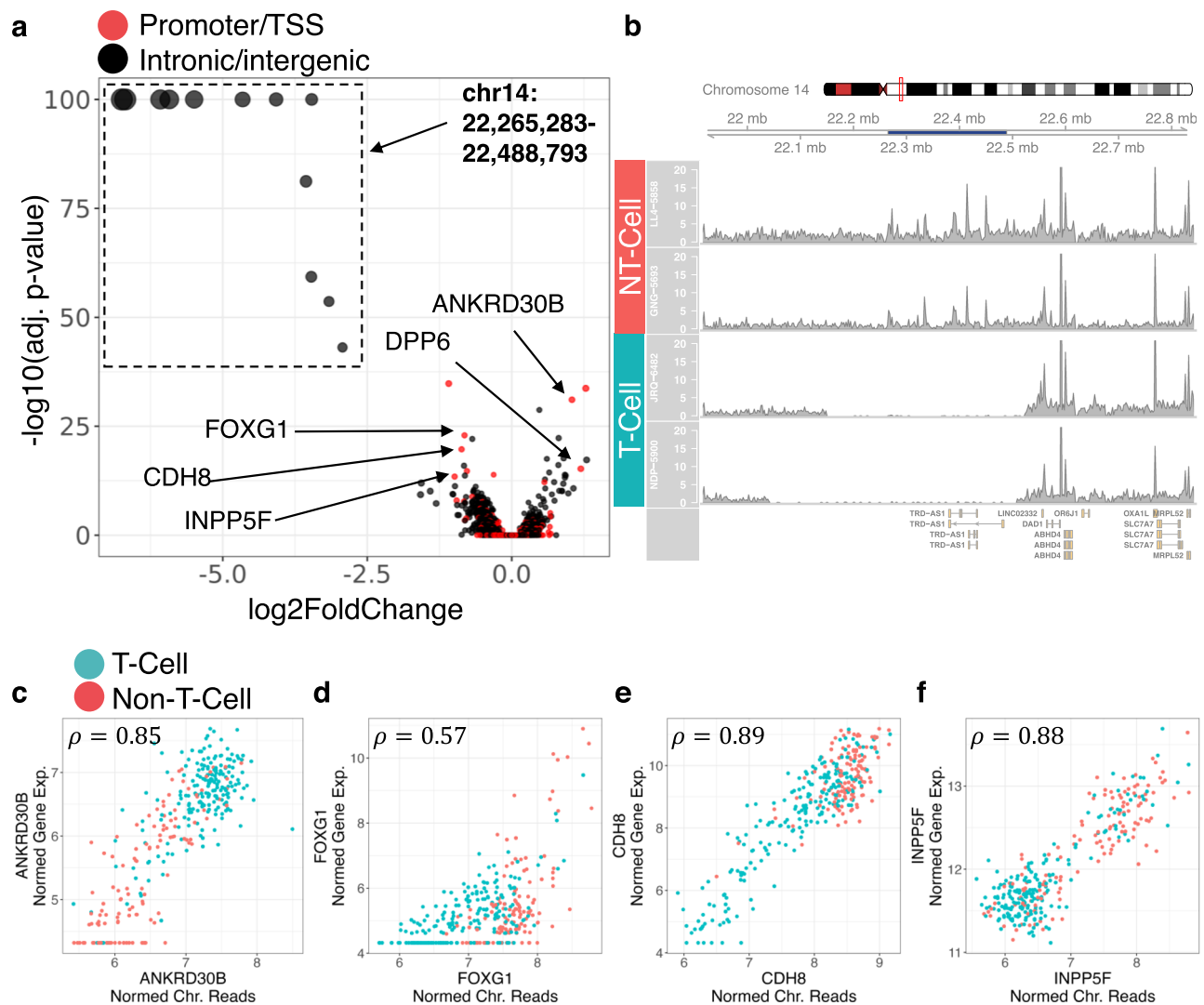


Fig. 3 | Differentiation-associated covariates. a Volcano plot for PBMC-associated differential signal. The p -values are calculated using a two-sided Wald test as implemented in DESeq2 and adjusted using a B-H correction. **b** Coverage plot spanning T-cell receptor genomic region for two T-cell-derived samples and two non-T-cell-derived samples. Dark blue region in genome axis scale spans the

regions in the box from **a**. **c–f** Plots of chromatin accessibility against gene expression for selected genes from **a**. Samples colored according to PBMC type. Pearson correlations between plotted chromatin accessibility and gene expression are indicated in top left corner.

whether the observed epigenetic memory could be attributed to differentiation bias resulting in different cell type distributions, we compared PBMC type to ICC staining data of the motor neuron cultures. There were no significant associations ($p < 0.01$) between PBMC type and the percent of cells that stained positive for S100B, Nestin, ISL1, NKX6.1, TUJ1, or SMI32. In general, chromatin accessibility was less correlated with ICC staining markers than gene expression (SI Section 7). This indicated that the observed epigenetic memory, which influences a small set of genes, cannot be attributed to differentiation bias.

Clinical and demographic covariates

The key question in the analysis of iPSC derived motor neurons is whether omic profiles capture clinical information about the individuals from whom the cells were derived. In Fig. 2c, we found that out of the four patient-specific covariates tested, only sex and ancestry drove variance in chromatin accessibility; the contributions of case status and age were negligible. The lack of an age-associated signal is not surprising, as iPSC-reprogrammed cells exhibit elongated telomeres, reduced oxidative stress, and a loss of senescence markers, all

hallmarks of younger cells^{33–35}. The ancestry signal was confirmed with differential analysis and revealed 47 DARs (adj. p -value < 0.01 , Figure S7a), of which 16 were more accessible in individuals of African ancestry. The top DAR by p -value was the promoter/TSS of *RNF135* (adj. p -value = $3e-14$, $\log_2\text{FC} = 0.8$) (Fig. 4a); despite the presence of six mismatches to the reference genome in the aligned reads (Figure S7b), subjects of African ancestry were found to have a higher accessibility of this promoter region than subjects of European ancestry. These mismatches are consistent with the genomics data (rsIDs: rs7221217, rs7221238, rs7219775, rs7221473, rs7225888, rs7211440)¹¹. Interestingly, the promoter/TSS of *RNF135* has previously been identified to exhibit hypomethylation in subjects of African ancestry, which is consistent with our observation of higher accessibility³⁶. The remaining two clinical covariates, sex and ALS status, are examined in the following sections.

Sex-associated DARs are not limited to sex chromosomes and reveal X-chromosome inactivation

There were 72 significant DARs associated with sex (adj. p -value < 0.01 , $\text{abs}(\log_2\text{FC}) > 1$); of these regions, 40 were Y-chromosomal, 22 were X-

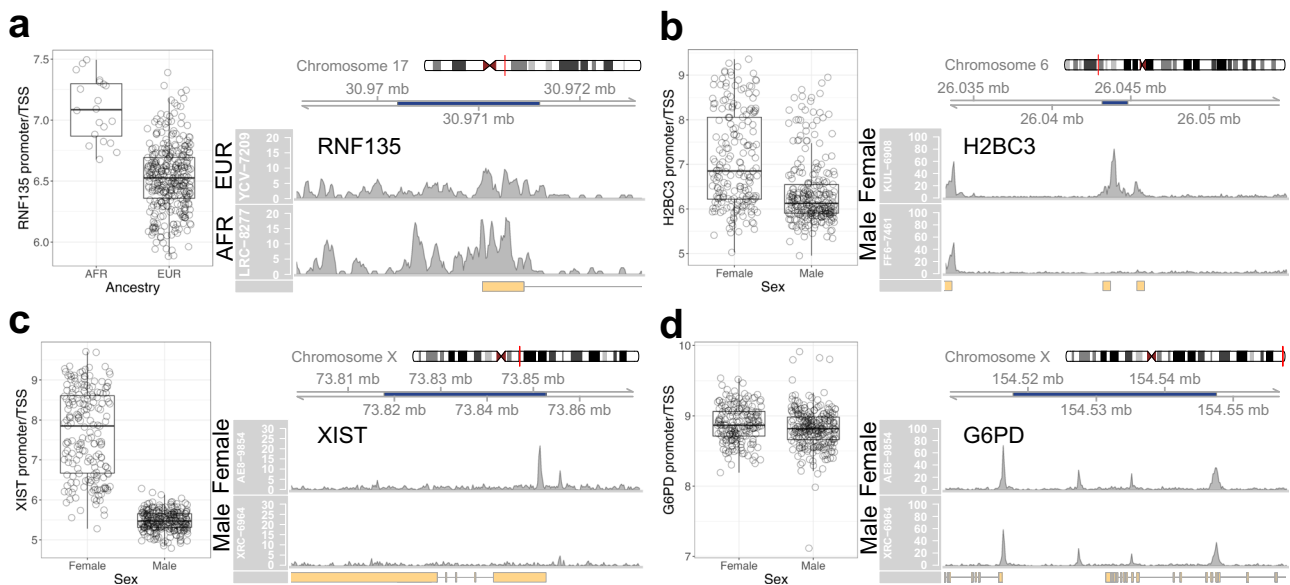


Fig. 4 | Clinical covariates. **a** (Left) Normalized chromatin read counts and (right) example coverage plot for individuals of African (AFR) and European (EUR) ancestry for the promoter/TSS of *RNF135* (adj. p -value = $3e-14$, $\log_2FC = -0.78$, $n_{AFR} = 20$, $n_{EUR} = 340$). **b** (Left) Normalized chromatin read counts and (right) example coverage plot for *H2BC3*, a sex-associated autosomal DAR (adj. p -value = $3e-17$, $\log_2FC = -1.16$, $n_{Female} = 194$, $n_{Male} = 262$). **c** (Left) Normalized chromatin read counts and (right) example coverage plot for *XIST*, a sex-associated DAR that

escapes X-inactivation (adj. p -value < $1e-278$, $\log_2FC = -4.00$, $n_{Female} = 194$, $n_{Male} = 262$). **d** (Left) Normalized read counts and (right) example coverage plot for *G6PD*, a housekeeping gene on chromosome X (adj. p -value = 0.02 , $\log_2FC = -0.05$, $n_{Female} = 194$, $n_{Male} = 262$). All boxplot boxes indicate the 25th, 50th (median), and 75th quartiles; boxplot whiskers extend 1.5 interquartile ranges from the median. All p -values are calculated using a two-sided Wald test as implemented in DESeq2 and adjusted using a B-H correction.

chromosomal, and 10 were autosomal (Figure S7c). The only promoter in the 10 autosomal significant DARs was that of *H2BC3*, a histone encoded on chromosome 6 (Fig. 4b). The most significant DAR on chromosome X corresponded to the promoter of *XIST*, a gene responsible for X chromosome inactivation (Fig. 4c). The most significant DAR on chromosome Y corresponded to the promoter of *ZFY* (Figure S7d). All of these genes were also differentially expressed; for example, the *XIST* promoter accessibility had a correlation of 0.94 with *XIST* gene expression, and the accessibility of the *H2BC3* region was most highly correlated with the expression of its corresponding gene (0.89), the histone *H3C3* (0.72), and the histone *H4C9* (0.46). The remaining autosomal regions were all most highly correlated with the expression of genes on the Y chromosome. Motif enrichment of the 10 significant autosomal DARs using HOMER³⁷ did not return any significantly enriched motifs. The inactivation of the X chromosome was confirmed by examining numbers of background reads (SI Section 8).

Identifying ALS-associated DARs

We conducted over two dozen differential analyses with the aim of identifying differential signals associated with ALS case status; the comparisons spanned metrics associated with disease progression, ALS subtypes, and medication intake (SI Section 9). One of the signals that emerged was the association of 1402 DARs with ALS case status at a FDR < 0.1. We note that this signal is only marginally significant, and there were no DARs associated with ALS case status at a FDR < 0.01, suggesting that even with the current sample size, it is difficult to detect such signals given the diversity of ALS. However, this does not mean that the ALS-associated signal should be overlooked; in fact, motif enrichment analysis of the 1402 differentially accessible regions against the remainder of the consensus peakset as a background revealed significant enrichment for the binding site of the NFY transcription factor (p -value $1e-15$). The disruption of NFY has previously been shown to cause neurodegeneration with a ubiquitin/p62 pathology³⁸.

We additionally identified a robust, but weak, differential signal associated with the number of years that passed between disease onset and PBMC sample collection, consisting of 2601 DARs (Supplementary Data S6, adj. p -value < 0.1). Interestingly, when compared against the GC content normalized consensus peakset, the genomic sequences for these chromatin regions were enriched for the binding sequence of Nrf1, a transcription factor involved in regulating cellular stress responses³⁹; its deletion in the mouse central nervous system has also been shown to cause motor neuron dysfunction⁴⁰. This observation suggests that the accessibility of certain chromatin regions is associated with a slower rate of disease progression. At the same time, it also appears to be inconsistent with the lack of a significant differential signal associated with total disease length, as measured by the number of years between death and onset. These two observations can be reconciled by noting that there are nearly twice as many subjects with a recorded age of onset and age of PBMC sample collection ($n = 337$) than those with a recorded age of onset and age of death ($n = 144$).

C9orf72 TSS is differentially accessible in ALS patients with *C9orf72* hexanucleotide repeat expansion

We hypothesized that the heterogeneity of ALS might obscure disease-relevant signals. To test this hypothesis, we compared ALS cases that were known to harbor the mutant *C9orf72* hexanucleotide repeat expansion ($n = 31$: C9+) to verified C9- negative cases and healthy controls ($n = 116$: HC). For both the C9+/C9- and C9+/HC comparisons, the *C9orf72* TSS had significantly lower normalized chromatin reads (adj. p -value = $1e-50$) in C9+ ALS patients, with a consistent \log_2FC of -0.6 (30% decrease in accessibility) (Fig. 5a). The differential signal extends over 6 raw read lengths away from the repeat expansion, indicating that the signal is not a mapping artifact (Figure S8a). Interestingly, the inclusion of the FRiP score as a covariate in the differential analysis improved the adjusted p -value from $1e-20$ to $1e-50$, but did not influence the \log_2FC . The observation of lower chromatin reads is consistent with the hypothesis of haploinsufficiency of the *C9orf72*

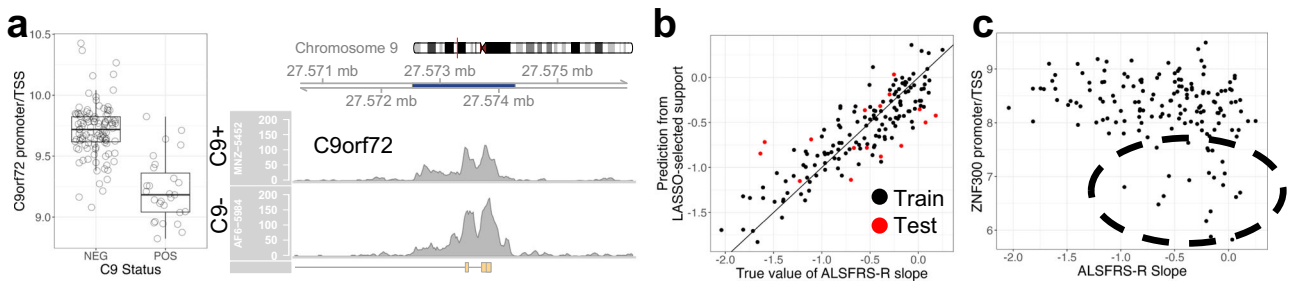


Fig. 5 | AALS ATAC-seq ALS signals. **a** (Left) Normalized chromatin read counts ($n_{POS} = 27$, $n_{NEG} = 112$) and (right) example coverage plot for the promoter/TSS of *C9orf72*; the coverage plot corresponds to an ALS case with a repeat expansion length of 274 (C9+) and an ALS case without the *C9orf72* mutation (C9-). **b** Prediction vs true value of ALSFRS-R slope; samples used to train the classifier are

black ($n_{train} = 140$), and samples used to test the classifier are red ($n_{test} = 16$). Black line is a reference line of slope 1 and intercept 0. **c** Normalized chromatin read counts for the promoter of *ZNF300*, a chromatin region associated with ALSFRS-R slope from the classifier; a set of samples that appear to both have slower disease progressions and low accessibility are circled.

transcript contributing to neurodegeneration and agrees with our previous report showing reduced *C9orf72* transcript levels^{26,41}. We did not observe any significant dependence between C9 repeat length and chromatin reads that could not be explained by C9 status.

ATAC-seq signals predict ALSFRS-R slope

Rates of disease progression in ALS are highly variable, with the time from first symptoms until death ranging from months to decades. One of the most widely used measures of the rate of disease progression is the linear slope of the ALSFRS-R score across time. To explore whether chromatin accessibility contains information related to progression, we sought to predict ALSFRS-R slopes from ATAC-seq.

ALSFRS-R slope was not significantly associated with the accessibility of any one chromatin region. This is not necessarily surprising, as any genetic component to the rate of progression is likely to be multifactorial in nature. To take this into account, we used linear regression with a LASSO penalty to search for a small set of regions that were the most predictive (see Methods, Figures S8b, c). LASSO linear regression identifies a set of predictive features (predictors) in an underdetermined system by penalizing the size of regression coefficients; the design of the penalty term allows for variable selection. The features identified by this regression approach are not necessarily exhaustive; only one of a set of highly correlated features may appear in the final set of predictors.

To construct the predictor, we used a set of 156 filtered samples (see Methods, Figures S8b, c, Supplementary Data S7) that were split into a training data set (140 samples, Supplementary Data S7) and an out-of-sample testing data set (16 samples, Supplementary Data S7). We ran LASSO linear regression with ten-fold cross-validation on the training set to select the regularization parameter (Figure S8d). As the set of selected features (in this case, chromatin regions) depends on how samples are split into validation folds, we reran the feature selection step 1000 times, each time randomly reassigning the samples in the training data to different validation folds. The predictors fit from each run were evaluated according to their performance on the out of sample test data set (16 samples). At least one region was returned in 889 of the 1000 runs and there were 24 regions that appeared across over half of the runs (Supplementary Data S8). Across these 889 runs, the predicted model consisted of 24.2 ± 0.3 (mean \pm s.e.) regions and resulted in a mean training root mean squared error (RMSE) of 0.239 ± 0.001 and R^2 of 0.774 ± 0.003 (mean \pm s.e.). The out of sample testing data set RMSE and R^2 were 0.468 ± 0.001 and 0.244 ± 0.003 , respectively (mean \pm s.e.) (results for one run in Fig. 5b). Surprisingly, performance on the held-out test data are on par with the models that have attempted to predict ALSFRS-R slope from clinical data; a model using neurofilament concentrations at diagnosis exhibited RMSEs of 0.4 and 0.9 in validation cohorts⁴²,

and models based on clinical metadata returned root mean squared deviations of 0.54^{43} .

We examined the 24 regions that were returned in over half of the 1000 cross validation reruns (Supplementary Data S8). While the genes in these regions are not significantly enriched for any single biological process (likely due to the limited size of the geneset), several are of particular interest. For example, *CHCHD2* is a mitochondrial protein definitively associated with Parkinson's disease and an interaction partner of the ALS protein *CHCHD10*⁴⁴. *CHCHD2* gene expression has also been found to be significantly reduced in post-mortem brains of individuals with Parkinson's disease, and mutations within the protein have been reported to promote alpha-synuclein aggregation^{45,46}. *OTF* and *LMX1A* are both transcription factors required for the development of dopaminergic neurons⁴⁷; *LMX1A* is associated with Parkinson's disease, and previous experiments have shown that the reduction of *LMX1A* and the closely related *LMX1B* negatively impacts dopaminergic neuron survival by increasing oxidative stress and generating mitochondrial DNA damage⁴⁸. The protein encoded by *FGF1*, also known as aFGF, is known to be a neuroprotective and neuroregenerative factor, and its application has been shown to protect cortical neuron cultures against oxygen glucose deprivation-induced cell damage⁴⁹.

Another transcription factor in the feature set is *ZNF300*; it appears that a subset of ALS cases with slower progression rates also exhibit lower *ZNF300* gene promoter/TSS chromatin accessibility (Fig. 5c). Additionally, the chromatin accessibility of this gene has a correlation of 0.89 with its gene expression. Previous work has shown that *ZNF300* is a transcription repressor that localizes to the nucleus and is expressed in the heart, skeletal muscle, and brain⁵⁰. It is also associated with NF- κ B pathway activation and MAPK/ERK signaling⁵¹. Interestingly, two other genes appearing in the geneset, *TRAF3IP2* and *IRF7*, which are both involved in the inflammatory response, are closely related to the NF- κ B pathway; *TRAF3IP2*, which encodes the Act1 protein, activates NF- κ B⁵² and *IRF7* can form a transcriptional complex enhanceosome with NF- κ B⁵³. The NF- κ B pathway has been suggested to play a role in ALS disease progression, suggesting that the lower chromatin accessibility of the *ZNF300* TSS potentially indicates a protective mechanism against ALS progression⁵⁴.

ATAC-RNA co-expression analysis reveals putative enhancers for ALS genes

ATAC-seq data can be integrated with RNA-seq data to functionally characterize chromatin regions and identify cis-regulatory elements⁵⁵. We found that the expression of 13,261 genes is significantly associated with the accessibility of 24,810 chromatin regions (adj. p -value < 0.01) located within 250 kb of gene transcription start sites (Fig. 6a, b, Supplementary Data S9) (window width chosen to match Corces et al.⁵⁵ and the default settings for integrative analysis in ArchR⁵⁶). Peaks in

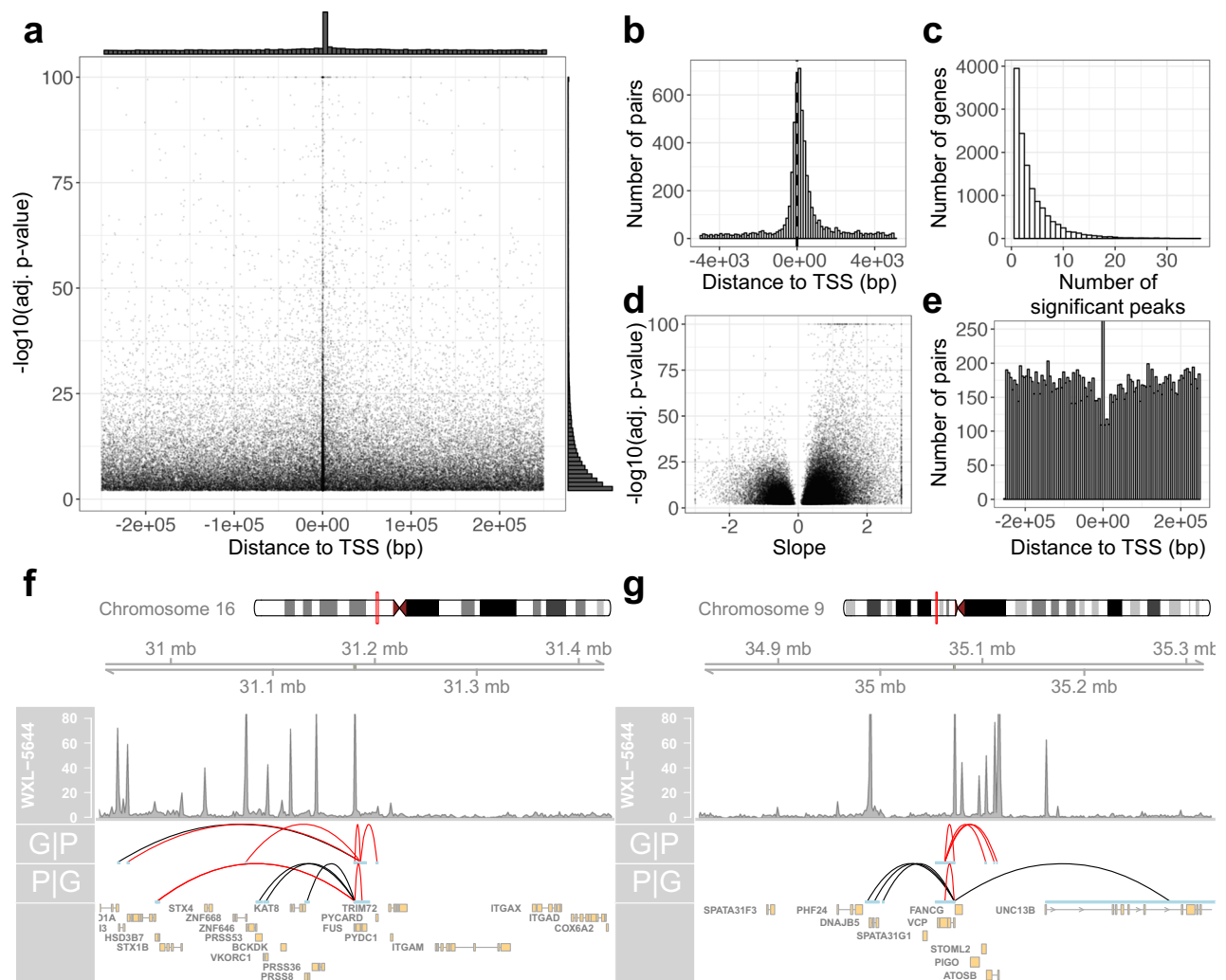


Fig. 6 | ATAC/RNA co-expression analyses. a Bonferroni-adjusted p -value and distance between peak center and gene transcription start site for all significant peak-gene pairs, with marginal histograms along the top and right axes. Each p -value is calculated using a two-sided Student's t test. **b** Same as top marginal histogram in (a) but zoomed in to ± 5 kb. **c** Histogram for the number of peaks that each gene is significantly associated with. **d** Volcano plot of the slope and p -value of all significant peak-gene pairs. Each p -value is calculated using a two-sided

Student's t -test. **e** Histogram of distance between peak center and gene transcription start site for all significant peak-gene pairs with a negative slope. Coverage plots centered around the *FUS* gene (**f**) and *VCP* gene (**g**). Arcs in the G|P row link the *FUS/VCP* transcription start site to peaks that are significantly associated with *FUS/VCP* gene expression, respectively. Arcs in the P|G row indicate correlations of the *FUS/VCP* promoter peak accessibility and gene expression. Red – positive correlation; black – negative correlation.

significant peak-gene pairs were concentrated near gene transcription start sites, but otherwise uniformly distributed across the examined 500 kb window (Fig. 6b). This geneset included 89 ALS associated genes (from the ALSod³⁷), whose expression was significantly associated with the accessibility of 422 chromatin regions, indicating the utility of this dataset to study disease relevant pathways. For example, *EPHA4*, a gene whose expression has been reported to modify ALS disease progression⁵⁸, was associated with the accessibility of a peak 20 kb upstream (adj. p -value $1e-42$).

Most genes were associated with multiple peaks (Fig. 6c) and one third of peak-gene pairs exhibited an inverse association between gene expression and peak accessibility (Fig. 6d). As is true for all peak-gene pairs, the pairs with inverse associations of accessibility and expression were also concentrated near transcription start sites. However, the inversely correlated pairs had a noticeable dip in density extending roughly 5 kb upstream and 15 kb downstream (Fig. 6e). We hypothesized that these inverse associations could be attributed to binding of transcriptional repressors. The genomic sequences of the peaks in these associations that are also proximal to a TSS (<2.5 kb) were enriched for the binding motifs of the YY1 (p -value $1e-12$) and KLF14

(p -value $1e-10$) transcription factors (see Methods), both of which have been reported to exhibit repressive transcriptional activity^{59,60}. The most significant peak-gene pairs with such a negative effect size involved the *HOXB5* and *HOXB4* genes, likely due to their regulatory activity; in tumor samples from the Cancer Genome Atlas, similarly strong correlations between accessibility and expression were found at the *HOXB* locus as well⁵⁵.

While many transcription factors are thought of as activators, several are known to be capable of either activating or repressing genes (citations including Dawson et al.⁶¹). We found that nearly 20% of all significant peaks were associated with both increased expression of some genes and decreased expression of others. In fact, this is the case for over 20% of all significant peaks (Supplementary Data S10). The top peaks that exhibit this behavior are located near promoters of genes responsible for transcriptional regulation. For example, the promoter for *HOXB9* exhibited a highly significant positive effect size with the expression of *HOXB9* (adj. p -value $1e-121$), but a highly significant negative effect size with the expression of *HOXB5* (adj. p -value $1e-88$). Other examples of peaks include the promoter for *CHD4*⁶², a gene encoding a member of a chromatin remodeling complex, and the

promoter for *PCDHA4*⁶³, a gene which lies in the tightly regulated protocadherin A gene cluster. We also find a similar behavior, albeit to a less significant extent, when we examine ALS genes. For example, the chromatin accessibility of the *FUS* TSS is positively correlated with *FUS* expression (adj. *p*-value $1e-5$), but negatively (and more significantly) correlated with the expression of *PRSS8* (Fig. 6f). Overall, it appears that the behavior of this subset of peaks can be attributed to global changes in transcription, of which the peaks in question are usually not the root cause.

Another example demonstrating how global patterns in transcription affect the results of epigenomic/transcriptomic integration can be found by analyzing links between genes and their promoters. In most cases, genes are most strongly correlated with the nearest peak. But there is another pattern that is found in one quarter of the genes. These genes are most strongly correlated with distal peaks, while their promoter accessibility is most strongly correlated with an entirely different gene. For example, the expression of the ALS gene *VCP* was much less significantly associated with the accessibility of its promoter (adj. *p*-value $2e-3$) than a distal element 42 kb upstream (adj. *p*-value $7e-15$) (Fig. 6g). At the same time, the accessibility of the *VCP* promoter was significantly associated with the expression of *DNAJB5* (adj. *p*-value $2e-13$) and *UNC13B* (adj. *p*-value $7e-8$). These phenomena highlight the complexity of the mechanisms involved in regulating gene transcription. More broadly, these analyses represent a bridgehead, showing the utility of these data in integrating the epigenomic and transcriptomic profiles of iPSC-derived motor neurons to characterize gene regulatory networks active in ALS and beyond.

Discussion

With over 5 trillion bases sequenced, the ATAC-seq data presented in this work is the largest ATAC-seq dataset generated for iPSC-derived motor neurons to date, and it is one of the largest ATAC-seq datasets generated by a single consortium overall. The consistency of the biological sample being produced (i.e., motor neuron cultures from hundreds of individuals rather than mixed tissues from the same individual) makes this dataset amenable to revealing insights that extend beyond ALS disease-associated signatures to other covariates, such as sex and the iPSC tissue of origin.

To assemble a dataset of this size, it is necessary to conduct a study which spans years. Over this period, the goal of maintaining consistent data generation methods can come into conflict with facility changes, instrumentation modernizations, and other unavoidable events. This can in turn influence downstream processing results. We showed an example of this in the analysis of the sequencer-associated differential signal. In general, the best way to monitor these changes is to examine batch-specific QC metrics. In future studies, efforts should be focused not only on surpassing a specific set of QC metrics, such as those defined by ENCODE, but also to ensure that the final quality control metrics have minimal batch to batch variance.

Chromatin read counts are susceptible to influence by genetic variants and the resulting mismapping. The best example of this is the apparent difference in chromatin accessibility between the line used as a BTC/BDC and other control lines. We determined that the difference is due to a 2 kb deletion, which we confirmed through comparisons with the genomics data. This signal arises despite a read mapping rate above 97.5% across almost all samples. We benefited from the fact that full-genome sequences were available for each sample in our study. In the future, we recommend that differential accessibility signals are verified against genomic data of the same sample in inter-individual comparisons, or that raw reads are aligned to individualized genomes when those data are available. The latter approach has previously been shown to alter peak calls in ChIP-seq data and alignment in RNA-seq data⁶⁴. When genomics data is not available, inspection of raw reads

aligned to differential peaks could reveal SNPs; genome coverage visualization software such as Gviz⁶⁵ provides a streamlined approach for this.

Several studies have examined the question of which normalization approach is superior for the analysis of ATAC-seq data. Our analysis was based solely in the framework provided by DESeq2, and we decided to use the default DESeq2 geometric median of ratios algorithm to estimate normalization factors. We found that it performed similarly to normalizing by reads in peaks (RiP). Notably, it outperformed normalization by total reads, which failed to identify the *C9orf72* TSS DAR in C9+/C9- ALS comparisons and to generate separation by PBMC type in PCA. However, RiP normalization alone is imperfect; indeed, the covariates total reads, reads in peaks, and FRiP score are closely related, and we observed a strong dependence of certain chromatin regions on FRiP score even after RiP normalization. These strong dependences are a concern as they will lead to false positive results if co-accessibility analyses, such as WGCNA, are used. Future work could focus on the analysis of alternative methods for normalization, which could include data from other omics modalities as validation.

The concept of epigenetic memory describes the phenomenon wherein iPSC-derived cells retain epigenetic characteristics of the cell type from which the iPSC clone was dedifferentiated⁶⁶. In the analysis of this data, we observed that the chromatin accessibility of several regions was significantly associated with the PBMC type. A fraction of the affected regions corresponded to T-cell receptor loci, where genomic TCR rearrangements prevented reads from mapping. These DARs are therefore not reflective of epigenetic memory⁶⁷. At the same time, the differential signal at several other chromatin regions could not be explained by mapping artifacts. For example, the TSS for *FOXG1* is significantly more accessible in non-T-cell derived samples than it is in T-cell derived samples, while the TSS for *ANKRD30B* exhibits the opposite dependence on PBMC type. *FOXG1* is a neurodevelopmental factor that functions as a transcriptional repressor, promotes neurogenesis, and inhibits gliogenesis; mutations in the gene are associated with Rett's syndrome⁶⁸. *ANKRD30B* is a gene that is expressed in the brain; it has recently been found to be differentially methylated in subsets of patients with Alzheimer's disease⁶⁹ and Williams Syndrome⁷⁰.

The ATAC-seq data exhibited significant correlations with ICC staining markers, but these signals were not as strong as those found in the RNA-seq data, and they mainly corresponded to intronic/intergenic chromatin regions. This was surprising because the epigenome is responsible for establishing cellular phenotypes⁷¹. There are a few possible explanations for the weaker signal compared to the RNA-seq data. First, there is more biological noise in the transduction of an epigenomic signal into a proteomic signal, as it requires both transcription and translation. Another possible explanation for the discrepancy is that the expression of ICC markers is a response to an external stimulus, which could induce changes in the cellular populations of transcription factors. For example, S100B can be released by damaged cells⁷². Finally, it is conceivable that the same ICC staining markers can stain multiple cell types, all of which have unique epigenomic signatures, resulting in low correlations with chromatin accessibility.

We identified both ancestry- and sex-associated differential signals in the ATAC-seq data. The observation of ancestry-specific differential accessibility in iPSC-derived motor neurons highlights the need to explore whether ancestry may play a role in motor neuron function and survival.

The sex associated DARs spanned both the autosomal and sex chromosomes, a finding which is consistent with previous work. For example, sex has been found to influence autosomal chromatin accessibility in immune cells in an age-dependent manner⁷³. We showed that the differential signal associated with chromosome X was

mostly driven by background reads from the inactivated chromosome X, with the exception of 22 DARs that clearly escaped inactivation and include the promoter/TSS for *XIST*. It was important to establish X chromosome dosage compensation in these iPSC lines, as its erosion has been found to be a limitation in iPSC-based disease modeling^{74,75}. While low passage iPSCs retain X inactivation, longer culture times lead to gradual re-activation of the inactivated X chromosome that is not reversed by differentiation⁷⁶. Hallmarks of eroded dosage compensation include decreased *XIST* gene expression and a loss of H3K27me3 marks, and can lead to the remodeling of the iPSC proteome^{76,77}. It is also interesting to note that in gene expression data, the number of sex-associated differentially expressed genes (78 genes, $\text{abs}(\log_2\text{FC}) > 1$, $\text{adj. } p\text{-value} < 0.01$) was much higher than in the ATAC-seq data; similar to the observations with ICC staining markers, this suggests that there is an additional level of regulation governing the expression of these genes that is not apparent at the epigenome level. Overall, the fact that there is still sex-based variant gene expression and chromatin accessibility at the level of the motor neuron cultures raises questions regarding whether sex may impact motor neuron function and survival.

The strongest ALS-specific signal revealed by the ATAC-seq data was a difference in chromatin accessibility at the *C9orf72* TSS for C9+ALS cases. This finding is perhaps not surprising. There is strong evidence that ALS is not one disease, but several different diseases that culminate in the same clinical phenotype of upper and lower motor neuron degeneration¹. There are more than 25 known independent genetic causes of ALS, which collectively explain less than 15% of cases. Thus, it is likely that the variability among ALS patients may be greater than any common “ALS signature.” In addition, due to epigenetic reprogramming, iPSCs are likely to best represent early phases of disease. Thus, these samples may reflect the diverse early causes of the disease and not later stages of cell death that may be common to more patients. The iPSC data in this study, therefore, are best used to explore how genomic factors beyond the known disease-causing loci contribute to the high heritability of ALS.

Several studies have attempted to use clinical data and other biomarkers to group ALS patients and predict ALS disease progression. The Prize4Life challenge crowdsourced machine learning models to predict ALS disease progression from clinical data, identifying time from disease onset, ALSFRS, forced vital capacity, and blood pressure among the top predictors⁴³. Semi-supervised machine learning models applied to clinical data of Italian ALS patients was found to separate ALS patients according to the Chio criteria⁷⁸. The distribution of T cell populations in the CSF of ALS patients was found to be associated with ALS disease progression⁷⁹. In this study, we show that the ATAC-seq data of iPSC-derived motor neurons from ALS patients has a predictive power for disease progression rate that is on-par or better than predictions from clinical data or blood-based biomarkers.

We seek to answer a different question than these prior studies. We asked whether iPSC-based models retain clinically relevant signals. On the one hand, the high heritability of ALS suggests that they should. On the other hand, iPSC-derived motor neurons are expected to exhibit the characteristics of a ‘younger’ cell, with reversed senescence due to reprogramming³³. Are iPSC-derived neurons too ‘young’? Our results suggest that ALS-relevant genomic influences emerge very early in this system. It remains an open question how early such signals might emerge in patients, but some studies of presymptomatic C9orf72 mutant repeat expansion carriers suggest that some effects may be detectable early in life^{80,81}.

Much more work will be needed to identify how the genetic variants influence disease progression. By their nature, the machine learning models we used only return a subset of the relevant features and cannot determine which correlated features are causal, only which have the strongest predictive value in a particular dataset. Batch

effects also hinder the interpretability of signals associated with disease progression, especially when already small batch sizes are further halved by the study design requirements to match the numbers of ALS cases and controls. Nevertheless, it is interesting that the chromatin regions returned by the ALSFRS-R slope predictor are associated with neurodegenerative diseases. For example, *LMX1A* and *CHCHD2* have been previously associated with Parkinson’s disease. We also identify the potentially protective role of decreased accessibility at the *ZNF300* promoter/TSS. Finally, it is likely that ALSFRS-R slope may not be the best signal to try and predict, as it is a sum of scores reflecting deficits in extremely diverse symptoms and may mask important variability among patients with the same overall score. Future work will need to use more sophisticated analyses of clinical states and will need to integrate other omic signals. Overall, these results suggest that there is value in initiating iPSC model-based efforts geared towards studying disease progression rates, rather than case/control differences; these models could serve as a complement to existing biomarkers to explore the molecular basis of disease.

In this study, we examined the epigenomic profiles of one of the largest sets of iPSC derived motor neurons generated to date. Whereas these cell lines were generated to identify ALS-specific disease signatures, we showed that the epigenomic analysis of the iMN cultures could be used to gain insights that extend beyond the disease. We found that chromatin accessibility measurements were influenced by clinical covariates, such as sex and ancestry, differentiation-associated covariates, such as the iPSC cell type of origin, and sequencing-associated covariates, such as FRiP score and the read length used. Importantly, as this data is used by a wider audience, these covariates must be factored into any differential and co-accessibility analyses to avoid false-positive signals that are associated with ALS case status.

We described two ALS signals in this study. The first one was a decrease in the chromatin accessibility of the *C9orf72* promoter/TSS in samples exhibiting the mutant hexanucleotide repeat expansion. This supports the hypothesis of haploinsufficiency for the *C9orf72* transcript contributing to disease⁴¹. Additionally, we found that the epigenomics data could be used to construct a predictor of ALSFRS-R slope, and identified the downregulation of the *ZNF300* gene as having a potentially protective effect.

Overall, this paper underscores the value of conducting large-scale investigations of iPSC-derived cells for the study of ALS. After carefully compensating for sources of variance, these data reveal some of the complex interplay between chromatin accessibility, genetics, and disease subtypes, including, surprisingly, an association between epigenomic signals and the rate of disease progression. The expanding multi-omic data from Answer ALS and other efforts raises the prospect that integration of these data with other omics modalities and ALS omics datasets will uncover new directions in ALS research and help identify novel therapies.

Methods

Generation of iPSC motor neuron cultures

The iPSC cells were differentiated into motor neurons according to the direct iPSC cell-derived motor neuron (diMNs) protocol, which comprises three main stages (see Baxi et al.¹¹ for detailed procedure). Briefly, the Cedars-Sinai Biomanufacturing Center reprogrammed PBMCs using a non-integrating episomal plasmid method and differentiated iPSCs into motor neurons using a directed differentiation protocol comprising 3 stages. In Stage 1 iPSCs were plated in 6-well plates at a density of $5E + 05$ cells per well, and neural induction and hindbrain specification of iPSCs was achieved by dual inhibition of the SMAD and GSK3 β pathways for 6 days. In Stage 2, precursors were replated in fresh 6-well plates at a density of $7.5E + 05$ cells per well, and specification of spinal motor neuron precursors was achieved by addition of Shh agonists and retinoic acid for an additional 6 days.

Finally, in Stage 3 the precursors matured for recipe in Table. On Day 32 cells were collected and pelleted for subsequent shipping to 'omics sites. Representative wells of each cell line were fixed at the end of the differentiation and immunostained for markers of motor neuron identity (TUBB3, ISL1, SMI32, NKX6.1, s100B).

ATAC-seq experimental methods and quality control

As we wrote in Baxi et al.¹¹, ATAC-seq sample prep, sequencing and peak generation were carried out by Diagenode Inc. as further described⁸². Briefly, cells were lysed in ATAC-seq resuspension buffer (RSB; 10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl₂, and protease inhibitors) with a mixture of detergents (0.1% Tween-20, 0.1% NP-40, and 0.01% digitonin) on ice for 5 min. The lysis reaction was washed out with additional ATAC-RSB containing 0.1% Tween-20 and inverted to mix. Then 50,000 nuclei were collected and centrifuged at 450 × *g*. for 5 min at 4 °C. The pellet was resuspended in 50 μl of transposition mixture (25 μl of 2× Illumina Tagment DNA buffer, 2.5 μl of Illumina Tagment DNA enzyme, 16.5 μl of phosphate-buffered saline, 0.5 μl of 1% digitonin, 0.5 μl of 10% Tween-20 and 5 μl of water). The transposition reaction was incubated at 37 °C for 30 min followed by DNA purification. An initial PCR amplification was performed on the tagmented DNA using Nextera indexing primers (Illumina). Real-time (RT)-qPCR was run with a fraction of the tagmented DNA to determine the number of additional PCR cycles needed, and a final PCR amplification was performed. Size selection was done using AMPure XP beads (Beckman Coulter) to remove small, unwanted fragments (<100 bp). The final libraries were sequenced using the Illumina HiSeq 4000 (PE, 75-nt kit) and NextSeq 6000 (PE, 50-nt kit) platforms. All samples passed QC checks that included morphological evaluation of nuclei, fluorescence-based electrophoresis of libraries to assess size distribution and RT-qPCR to assess the enrichment of open chromatin sites.

ATAC-seq read alignment and peak calling

ATAC-seq data was processed using the ENCODE-DCC ATAC-seq pipeline v1.7.1. Reads were aligned to GRCh38 genome build using Bowtie2. The quality of the sequencing was assessed using FastQC. Samples had 62.0±0.8 (s.e.) million total reads after mitochondrial filtering and deduplication, with 90% of samples having over 40 million reads (Figure S1a). The average sample FRiP score was 0.242±0.002 (s.e.), with 77% of samples having a FRiP score higher than 0.2 (Figure S1b). The average transcription start site enrichment (TSSE) was 14.2±0.1 (s.e.), with 99% of samples having a TSSE greater than 7 (Figure S1c). All samples had a distinct nucleosome free-region in fragment length distribution plots. We identified open chromatin regions separately for each sample using the peak-calling software MACS2 and determined differentially open sites using DESeq2 (adj. *p*-value < 0.1).

Generation of consensus peakset and raw counts matrix

After peak calling, using the R package DiffBind⁸³, a consensus peakset was constructed by retaining peaks that were open in at least 10% of samples; it consisted of 100,363 variable-width chromatin regions. These regions were annotated using HOMER³⁷. Chromatin read counts from the consensus peakset were normalized using the DESeq2⁸⁴ *vst* function. A parametric fit was used for the dispersion estimate, and the default DESeq2 geometric median of ratios was used for the scaling factor estimate.

Evaluation of processed data quality

Coverage plots and read pileups were generated using the R package Gviz⁶⁵ (Fig. 1e). Sample-wise Pearson correlations within the BTC, BDC, and inter-individual groups (Figure S1d) were calculated using the columns of the normalized counts matrix. Outlying samples were identified using complete link hierarchical clustering with Euclidean

distance on the columns of the correlation matrix (heatmap for BDCs shown in Figure S1f).

Replication cohort analysis

22 samples were redifferentiated, sequenced, and compared to the initial cohort. To compare samples in the replication cohort to the initial cohort, DiffBind⁸³ was run on the 44 total samples from the initial and replication cohorts to generate a new consensus peakset and raw counts matrix. The raw counts matrix was normalized using the DESeq2 *vst* function as before. After normalization, an inter-sample correlation matrix was constructed by calculation Pearson correlations between individual samples. Samples were clustered by applying complete link hierarchical clustering on the Euclidean distance between the columns of this correlation matrix.

PCA/UMAP analysis

PCA was conducted on the top 500 most variably accessible regions using the R package PCAtools⁸⁵. The UMAP representation on the 100 most variably accessible regions was generated using the R package UMAP⁸⁶. For downstream analysis, the UMAP representation was used to estimate the PBMC-type identity of samples with missing data. PCA in Fig. 2a was conducted while including BDC/BTC samples, and PCA in Figures S2c-g excluded BDC/BTC samples.

Genomics data analysis

Genomics vcf files were obtained from Baxi et al.¹¹ and analyzed using Pysam⁸⁷.

Fitting linear mixed model

Each row (chromatin region) of the normalized read counts matrix was fit to a set of 16 covariates with a linear mixed model using the R package variancePartition⁸⁸. Discrete variables (differentiation batch, sequencer, sex, case status, PBMC type, ancestry) were modeled with random effects, and continuous variables were modeled with fixed effects.

Differential analysis for generating volcano plots

Differentially accessible regions (DARs) associated with covariates were identified by running DESeq2 on the raw counts matrix using a parametric dispersion fit and the default size parameter estimates while including FRiP score, sequencer, sex, PBMC type, and ALS case status as covariates. The *p*-values were adjusted for multiple hypothesis testing (labeled throughout the manuscript as “adj. *p*-value”) using the Benjamini-Hochberg (B-H) correction, unless noted otherwise; notably, *p*-values in ATAC-RNA co-expression analyses were corrected for multiple tests using the Bonferroni method. Adjusted *p*-values and log₂FC values for each covariate were used for volcano plots.

Motif enrichment analysis

All known motif enrichment analyses were conducted using the HOMER findMotifsGenome.pl function with the ‘-nomotif’ flag. In all analyses, target peaks were first narrowed to ±100bp of the peak summit. Background sequences normalized for GC content were generated by HOMER unless otherwise specified. Motif enrichment analysis of the promoter-proximal peaks with an inverse association with gene expression (Fig. 6e) was conducted against a background of all significant promoter-proximal peaks.

Simulations to identify the role of sequencer in influencing differential accessibility results

The effect of raw read length on differential accessibility was evaluated on a subset of 20 samples (10 with 75 bp HiSeq4000 reads, and 10 with 50 bp NovaSeq6000 reads). Raw reads were trimmed down to 50 bp and 36 bp lengths, and realigned to the reference genome.

To generate a consensus peakset and raw counts matrix, Diffbind was run on a set of 50 samples: 10 75 bp HiSeq4000 samples, 10 trimmed 50 bp HiSeq4000 samples, 10 trimmed 36 bp HiSeq4000 samples, 10 50 bp NovaSeq6000 samples, and 10 trimmed 36 bp NovaSeq6000 samples. The same set of samples were used for each trimming. The raw counts were analyzed using DESeq2 *vst* normalization and differential analysis.

The differential signal in this test set recapitulated the signal from analyzing all samples, with 75 bp HiSeq4000 samples exhibiting higher measured accessibility across most DARs (Figure S3a). Following this confirmation, the 75 bp reads were trimmed down to 50 bp, and the analysis was repeated; the most significant DARs had lost their significance (Figure S3b). We found that the measured chromatin accessibility of the most significant DARs fell with decreasing read length (Figure S3c).

Correlations with gene expression

There were 335 samples that had both gene expression and chromatin accessibility reported. Generation of raw counts for gene expression as described in Workman et al.²⁶. Raw counts from gene expression data were normalized using the DESeq2 *vst* function and the default DESeq2 parameters (parametric fit for the dispersion estimate, geometric median of ratios for size factor estimate).

Estimating reads not in peaks in X chromosome

The total number of reads that mapped to the X chromosome was estimated from the mitochondrial filtered, deduplicated BAM files using Samtools⁸⁹. The raw reads in peaks from the X chromosome was estimated using the Diffbind-output raw counts matrix. The number of reads not in peaks (RniP) on the X chromosome were estimated by subtracting these two quantities. The X chromosome RniP were normalized by dividing by the total number of reads to generate Figure S4e.

LASSO linear regression

The predictor was constructed on a set of filtered ALS cases. In total, there were 242 ALS samples with a recorded ALSFRS-R slope. The normalized chromatin read counts matrix was first subset to chromatin regions within 2 kb of a TSS and samples with a recorded ALSFRS-R slope. An inter-sample correlation matrix was constructed using this subset counts matrix. Euclidean distance complete-link hierarchical clustering on this correlation matrix revealed a set of 12 samples that were poorly correlated with other samples; these samples were excluded from further analysis (Figure S7b). In the remaining sample set, six samples were identified as having an outlying ALSFRS-R slope and excluded as well (Figure S7c). Outliers were defined in the traditional way as samples that exhibited an ALSFRS-R slope that was more than 1.5 interquartile ranges away from the first and third quartiles. To mitigate the effects of batch-to-batch variation, samples that came from batches with less than 4 representatives in the sample set were removed; this caused the removal of 68 samples. Next, samples were randomly divided into a 90/10 split for a training set (140 samples) and testing set (16 samples). LASSO linear regression was conducted using the R package glmnet⁹⁰. After each cross-validation variable-selection run, final predictors were constructed by fitting ALSFRS-R slopes of the training samples to the set of selected regions using ordinary multiple linear regression. Training RMSE and R² values were calculated for this model. This fit model was then used to predict the ALSFRS-R slopes of the out-of-sample testing data. The performance of the fit was evaluated according to the RMSE and R² values. In evaluating the reported R² value, the total variation was calculated relative to the mean of the training data. Otherwise, the squared correlation between predicted and actual value of the ALSFRS-R slope for the out of sample testing data is slightly higher: 0.275+/-0.002 (mean +/- s.e.).

ATAC-seq and RNA-seq co-expression analyses

ATAC-seq and RNA-seq co-expression was conducted by repurposing the tensorQTL package⁹¹, primarily due to its fast computations enabled on a GPU and its ability to incorporate covariates into the analysis. For matched samples, the FRiP score, sequencer, case status, sex, and PBM type were included as covariates. We regressed gene expression on chromatin accessibility using the *cis.map_nominal* function to get a list of nominal *p*-values for associations of each gene to all chromatin regions within 250 kb of its TSS. The Bonferroni method was used to correct for multiple hypothesis testing.

Ethical approval processes and donor consent for the use of cells

The Answer ALS program¹¹ collected the clinical data and derived all the cell lines used in this study. Answer ALS was approved by local institutional review boards, and all participants provided written informed consent. Consent was uniform across all sites and included an agreement to share data broadly for medical research. Subsequent use of these cells by the authors was done in accordance with local institutional review boards.

Reporting summary

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

Data availability

Data used in the preparation of this article were obtained from the ANSWER ALS Data Portal (AALS-01184). For up-to-date information on the study, visit <https://dataportal.answerals.org>. All data is available through the Answer ALS Data Portal following approval of a Data Use Agreement (DUA) form. Sample IDs used in this analysis can be found in Supplementary Data S11.

References

- van Es, M. A. et al. Amyotrophic lateral sclerosis. *Lancet* **390**, 2084–2098 (2017).
- Ryan, M., Heverin, M., McLaughlin, R. L. & Hardiman, O. Lifetime risk and heritability of amyotrophic lateral sclerosis. *JAMA Neurol.* **76**, 1367 (2019).
- Soufi, A. et al. Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* **161**, 555–568 (2015).
- Zhang, L., Liu, Y., Lu, Y. & Wang, G. Targeting epigenetics as a promising therapeutic strategy for treatment of neurodegenerative diseases. *Biochem. Pharmacol.* **206**, 115295 (2022).
- Chestnut, B. A. et al. Epigenetic regulation of motor neuron cell death through DNA methylation. *J. Neurosci.* **31**, 16619–16636 (2011).
- Tibshirani, M. et al. Dysregulation of chromatin remodelling complexes in amyotrophic lateral sclerosis. *Hum. Mol. Genet.* **26**, 4142–4152 (2017).
- Cobos, S. N., Bennett, S. A. & Torrente, M. P. The impact of histone post-translational modifications in neurodegenerative diseases. *Biochim. et. Biophys. Acta (BBA) - Mol. Basis Dis.* **1865**, 1982–1991 (2019).
- Klingl, Y. E., Pakravan, D. & Van Den Bosch, L. Opportunities for histone deacetylase inhibition in amyotrophic lateral sclerosis. *Br. J. Pharm.* **178**, 1353–1372 (2021).
- Brix, J., Zhou, Y. & Luo, Y. The epigenetic reprogramming roadmap in generation of iPSCs from somatic cells. *J. Genet. Genom.* **42**, 661–670 (2015).
- Goutman, S. A. et al. Recent advances in the diagnosis and prognosis of amyotrophic lateral sclerosis. *Lancet Neurol.* **21**, 480–493 (2022).
- Baxi, E. G. et al. Answer ALS, a large-scale resource for sporadic and familial ALS combining clinical and multi-omics data from induced pluripotent cell lines. *Nat. Neurosci.* **25**, 226–237 (2022).
- Ramos, D. M., Skarnes, W. C., Singleton, A. B., Cookson, M. R. & Ward, M. E. Tackling neurodegenerative diseases with genomic

- engineering: A new stem cell initiative from the NIH. *Neuron* **109**, 1080–1083 (2021).
13. Fujimori, K. et al. Modeling sporadic ALS in iPSC-derived motor neurons identifies a potential therapeutic agent. *Nat. Med.* **24**, 1579–1589 (2018).
 14. Li, J. et al. An integrated multi-omic analysis of iPSC-derived motor neurons from C9ORF72 ALS patients. *iScience* **24**, 103221 (2021).
 15. Bressan, E. et al. The Foundational Data Initiative for Parkinson Disease: Enabling efficient translation from genetic maps to mechanism. *Cell Genom.* **3**, 100261 (2023).
 16. Douvaras, P. & Fossati, V. Generation and isolation of oligodendrocyte progenitor cells from human pluripotent stem cells. *Nat. Protoc.* **10**, 1143–1154 (2015).
 17. Kwart, D. et al. A large panel of isogenic APP and PSEN1 mutant human iPSC neurons reveals shared endosomal abnormalities mediated by APP β -CTFs, not A β . *Neuron* **104**, 256–270.e5 (2019).
 18. Streeter, I. et al. The human-induced pluripotent stem cell initiative—data resources for cellular genetics. *Nucleic Acids Res* **45**, D691–D697 (2017).
 19. Karch, C. M. et al. A comprehensive resource for induced pluripotent stem cells from patients with primary Tauopathies. *Stem Cell Rep.* **13**, 939–955 (2019).
 20. The HD iPSC Consortium Induced pluripotent stem cells from patients with Huntington’s disease show CAG-repeat-expansion-associated phenotypes. *Cell Stem Cell* **11**, 264–278 (2012).
 21. Kaye, J., Reisine, T. & Finkbeiner, S. Huntington’s disease iPSC models—using human patient cells to understand the pathology caused by expanded CAG repeats. *Fac. Rev.* **11**, <https://doi.org/10.12703/r/11-16> (2022).
 22. Johns, A. E. & Maragakis, N. J. Exploring motor neuron diseases using iPSC platforms. *Stem Cells* **40**, 2–13 (2022).
 23. Bowles, K. R. et al. ELAVL4, splicing, and glutamatergic dysfunction precede neuron loss in MAPT mutation cerebral organoids. *Cell* **184**, 4547–4563.e17 (2021).
 24. Kiskinis, E. et al. Pathways disrupted in human ALS motor neurons Identified through Genetic Correction of Mutant SOD1. *Cell Stem Cell* **14**, 781–795 (2014).
 25. Maor-Nof, M. et al. p53 is a central regulator driving neurodegeneration caused by C9orf72 poly(PR). *Cell* **184**, 689–708.e20 (2021).
 26. Workman, M. J. et al. Large-scale differentiation of iPSC-derived motor neurons from ALS and control subjects. *Neuron* **111**, 1191–1204.e5 (2023).
 27. Buenostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A method for assaying chromatin accessibility genome-wide. *CP Mol. Biol.* **109**, 21.29.1–21.29.9 (2015).
 28. The ENCODE Project Consortium An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
 29. Sahinyan, K. et al. Application of ATAC-Seq for genome-wide analysis of the chromatin state at single myofiber resolution. *eLife* **11**, e72792 (2022).
 30. Stifani, N. Motor neurons and the generation of spinal motor neuron diversity. *Front. Cell. Neurosci.* **8**, 1–22 (2014).
 31. Liao, E. S. et al. Single-cell transcriptomic analysis reveals diversity within mammalian spinal motor neurons. *Nat. Commun.* **14**, 46 (2023).
 32. Sekine, K. et al. Robust detection of undifferentiated iPSC among differentiated cells. *Sci. Rep.* **10**, 10293 (2020).
 33. Lapasset, L. et al. Rejuvenating senescent and centenarian human cells by reprogramming through the pluripotent state. *Genes Dev.* **25**, 2248–2253 (2011).
 34. Soria-Valles, C. & López-Otín, C. iPSCs: On the road to reprogramming aging. *Trends Mol. Med.* **22**, 713–724 (2016).
 35. Miller, J. D. et al. Human iPSC-based modeling of late-onset disease via progerin-induced aging. *Cell Stem Cell* **13**, 691–705 (2013).
 36. Park, C. S. et al. Hepatocyte gene expression and DNA methylation as ancestry-dependent mechanisms in African Americans. *npj Genom. Med.* **4**, 29 (2019).
 37. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
 38. Yamanaka, T. et al. NF-Y inactivation causes atypical neurodegeneration characterized by ubiquitin and p62 accumulation and endoplasmic reticulum disorganization. *Nat. Commun.* **5**, 3354 (2014).
 39. Ruvkun, G. & Lehrbach, N. Regulation and functions of the ER-associated Nrf1 transcription factor. *Cold Spring Harb. Perspect. Biol.* **15**, a041266 (2023).
 40. Kobayashi, A. et al. Central nervous system-specific deletion of transcription factor Nrf1 causes progressive motor neuronal dysfunction: CNS-specific deletion of transcription factor Nrf1. *Genes Cells* **16**, 692–703 (2011).
 41. Shi, Y. et al. Haploinsufficiency leads to neurodegeneration in C9ORF72 ALS/FTD human-induced motor neurons. *Nat. Med* **24**, 313–325 (2018).
 42. Witzel, S. et al. Neurofilament light and heterogeneity of disease progression in amyotrophic lateral sclerosis: development and validation of a prediction model to improve interventional trials. *Transl. Neurodegener.* **10**, 31 (2021).
 43. Küffner, R. et al. Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. *Nat. Biotechnol.* **33**, 51–57 (2015).
 44. Burstein, S. R. et al. In vitro and in vivo studies of the ALS-FTLD protein CHCHD10 reveal novel mitochondrial topology and protein interactions. *Hum. Mol. Genet.* **27**, 160–177 (2018).
 45. Liu, X. et al. Reduced erythrocytic CHCHD2 mRNA is associated with brain pathology of Parkinson’s disease. *Acta Neuropathol. Commun.* **9**, 37 (2021).
 46. Ikeda, A. et al. Mutations in CHCHD2 cause α -synuclein aggregation. *Hum. Mol. Genet.* **28**, 3895–3911 (2019).
 47. Wasel, O. & Freeman, J. L. Chemical and genetic Zebrafish models to define mechanisms of and treatments for dopaminergic neurodegeneration. *IJMS* **21**, 5981 (2020).
 48. Doucet-Beaupré, H. et al. Lmx1a and Lmx1b regulate mitochondrial functions and survival of adult midbrain dopaminergic neurons. *Proc. Natl. Acad. Sci. USA.* **113**, (2016).
 49. Tsai, M. J. et al. Acidic FGF promotes neurite outgrowth of cortical neurons and improves neuroprotective effect in a cerebral ischemic rat model. *Neuroscience* **305**, 238–247 (2015).
 50. Gou, D. et al. Identification and functional analysis of a novel human KRAB/C2H2 zinc finger gene ZNF300. *Biochim. et. Biophys. Acta (BBA) - Gene Struct. Expr.* **1676**, 203–209 (2004).
 51. Sobocińska, J., Molenda, S., Machnik, M. & Oleksiewicz, U. KRAB-ZFP transcriptional regulators acting as oncogenes and tumor suppressors: an overview. *IJMS* **22**, 2212 (2021).
 52. Li, X. et al. Act1, an NF- κ B-activating protein. *Proc. Natl Acad. Sci. USA.* **97**, 10489–10493 (2000).
 53. Ning, S., Pagano, J. S. & Barber, G. N. IRF7: activation, regulation, modification and function. *Genes Immun.* **12**, 399–414 (2011).
 54. Källstig, E., McCabe, B. D. & Schneider, B. L. The Links between ALS and NF- κ B. *IJMS* **22**, 3875 (2021).
 55. Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).
 56. Granja, J. M. et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
 57. Abel, O., Powell, J. F., Andersen, P. M. & Al-Chalabi, A. ALSod: A user-friendly online bioinformatics tool for amyotrophic lateral sclerosis genetics. *Hum. Mutat.* **33**, 1345–1351 (2012).

58. Van Hoecke, A. et al. EPHA4 is a disease modifier of amyotrophic lateral sclerosis in animal models and in humans. *Nat. Med.* **18**, 1418–1422 (2012).
59. Gordon, S., Akopyan, G., Garban, H. & Bonavida, B. Transcription factor YY1: structure, function, and therapeutic implications in cancer biology. *Oncogene* **25**, 1125–1142 (2006).
60. Truty, M. J., Lomber, G., Fernandez-Zapico, M. E. & Urrutia, R. Silencing of the Transforming Growth Factor- β (TGF β) Receptor II by Krüppel-like Factor 14 underscores the importance of a negative feedback mechanism in TGF β signaling. *J. Biol. Chem.* **284**, 6291–6300 (2009).
61. Dawson, S. J., Liu, Y.-Z., Rodell, B., Mörröy, T. & Latchman, D. S. The ability of POU family transcription factors to activate or repress gene expression is dependent on the spacing and context of their specific response elements. *Biochem. J.* **314**, 439–443 (1996).
62. Perissi, V., Jepsen, K., Glass, C. K. & Rosenfeld, M. G. Deconstructing repression: evolving models of co-repressor action. *Nat. Rev. Genet.* **11**, 109–123 (2010).
63. Chen, W. V. & Maniatis, T. Clustered protocadherins. *Development* **140**, 3297–3302 (2013).
64. Groza, C., Kwan, T., Soranzo, N., Pastinen, T. & Bourque, G. Personalized and graph genomes reveal missing signal in epigenomic data. *Genome Biol.* **21**, 124 (2020).
65. Hahne, F. & Ivanek, R. Visualizing genomic data using Gviz and bioconductor. in *Statistical Genomics* (eds. Mathé, E. & Davis, S.) vol. 1418 335–351 (Springer New York, New York, NY, 2016).
66. Kim, K. et al. Epigenetic memory in induced pluripotent stem cells. *Nature* **467**, 285–290 (2010).
67. Matsumoto, T. et al. Functional neurons generated from T cell-derived induced pluripotent stem cells for neurological disease modeling. *Stem Cell Rep.* **6**, 422–435 (2016).
68. Patriarchi, T. et al. Imbalance of excitatory/inhibitory synaptic protein expression in iPSC-derived neurons from FOXG1+/- patients and in foxg1+/- mice. *Eur. J. Hum. Genet.* **24**, 871–880 (2016).
69. Semick, S. A. et al. Integrated DNA methylation and gene expression profiling across multiple brain regions implicate novel genes in Alzheimer's disease. *Acta Neuropathol.* **137**, 557–569 (2019).
70. Kimura, R. et al. Integrated DNA methylation analysis reveals a potential role for ANKRD30B in Williams syndrome. *Neuropsychopharmacology* **45**, 1627–1636 (2020).
71. Chen, T. & Dent, S. Y. R. Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nat. Rev. Genet.* **15**, 93–106 (2014).
72. Kleindienst, A., Hesse, F., Bullock, M. R. & Buchfelder, M. The neurotrophic protein S100B: value as a marker of brain damage and possible therapeutic implications. in *Progress Brain Res.* vol. 161 317–325 (Elsevier, 2007).
73. Márquez, E. J. et al. Sexual-dimorphism in human immune system aging. *Nat. Commun.* **11**, 751 (2020).
74. Sahakyan, A., Plath, K. & Rougeulle, C. Regulation of X-chromosome dosage compensation in human: mechanisms and model systems. *Philos. Trans. R. Soc. B* **372**, 20160363 (2017).
75. Sahakyan, A., Yang, Y. & Plath, K. The role of Xist in X-Chromosome dosage compensation. *Trends Cell Biol.* **28**, 999–1013 (2018).
76. Mekhoubad, S. et al. Erosion of dosage compensation impacts human iPSC disease modeling. *Cell Stem Cell* **10**, 595–609 (2012).
77. Brenes, A. J. et al. Erosion of human X chromosome inactivation causes major remodeling of the iPSC proteome. *Cell Rep.* **35**, 109032 (2021).
78. Faghri, F. et al. Identifying and predicting amyotrophic lateral sclerosis clinical subgroups: a population-based machine-learning study. *Lancet Digit. Health* **4**, e359–e369 (2022).
79. Yazdani, S. et al. T cell responses at diagnosis of amyotrophic lateral sclerosis predict disease progression. *Nat. Commun.* **13**, 6733 (2022).
80. Bertrand, A. et al. Early cognitive, structural, and microstructural changes in presymptomatic C9orf72 Carriers Younger Than 40 Years. *JAMA Neurol.* **75**, 236 (2018).
81. Gossink, F. et al. Early life involvement in C9orf72 repeat expansion carriers. *J. Neurol. Neurosurg. Psychiatry* **93**, 93–100 (2022).
82. Milani, P. et al. Cell freezing protocol suitable for ATAC-Seq on motor neurons derived from human induced pluripotent stem cells. *Sci. Rep.* **6**, 25474 (2016).
83. Ross-Innes, C. S. et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–393 (2012).
84. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
85. Blighe, K. & Lun, A. PCAtools: PCAtools: Everything Principal Components Analysis. R package version 2.12.0.
86. Konopka, T. umap: Uniform Manifold Approximation and Projection. R package version 0.2.10.0. (2023).
87. Pysam. <https://github.com/pysam-developers/pysam>.
88. Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinforma.* **17**, 483 (2016).
89. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Giga-Science* **10**, gjab008 (2021).
90. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Soft.* **33**, 1–5 (2010).
91. Taylor-Weiner, A. et al. Scaling computational genomics to millions of individuals with GPUs. *Genome Biol.* **20**, 228 (2019).

Acknowledgements

Support for the Answer ALS program was kindly provided by the Robert Packard Center for ALS Research at Johns Hopkins, Travelers Insurance, ALS Finding a Cure Foundation, Stay Strong Vs. ALS, Answer ALS Foundation, Microsoft, Caterpillar Foundation, American Airlines, Team Gleason, National Institutes of Health, Fishman Family Foundation, Aviators Against ALS, AbbVie Foundation, Chan Zuckerberg Initiative, ALS Association, National Football League, F. Prime, M. Armstrong, Bruce Edwards Foundation, The Judith and Jean Pape Adams Charitable Foundation, Muscular Dystrophy Association, Les Turner ALS Foundation, PGA Tour, Gates Ventures, and Bari Lipp Foundation. This work was also supported in part by grants from the National Institutes of Health (1RF1AG057331) and the MIT-GSK Gertrude B. Elion Research Fellowship Program for Drug Discovery and Disease (40009014). We would like to thank Dr. Michael J. Workman, Dr. Jie Wu, Dr. Zhuoxing Wu, Dr. Ryan G. Lim, Dr. Vikas Bansal, Connie New, and Matthew Leventhal for providing insightful comments on the manuscript.

Author contributions

S.T. and E.F. wrote the manuscript. S.T., K.V., V.K. and A.D. conducted the analysis. A.F. and S.L. produced and performed iPSC cell differentiation analysis. The Answer ALS consortium provided cell lines and clinical data used in this study and assisted with analysis. L.M.T., C.N.S., S.T., V.K. and E.F. edited the manuscript. E.F., C.N.S., L.M.T., D.S., J.D.R., J.E.V.E. and S.F. supervised the study.

Competing interests

The authors declare no competing interests.

Inclusion and Ethics Statement

We advocate for research practices that embrace inclusivity and diversity.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-47758-8>.

Correspondence and requests for materials should be addressed to Ernest Fraenkel.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024

the Answer ALS Consortium

Michael J. Workman¹³, Ryan G. Lim¹³, Jie Wu⁹, Zhuoxing Wu⁹, Loren Ornelas², Lindsay Panther², Erick Galvez², Daniel Perez², Imara Meepe², Viviana Valencia², Emilda Gomez², Chunyan Liu², Ruby Moran², Louis Pinedo², Richie Ho¹³, Julia A. Kaye^{4,5,6}, Terri Thompson¹⁴, Dillon Shear¹⁴, Robert Baloh¹³, Maria G. Banelos¹³, Veronica Garcia¹³, Ronald Holewinski³, Oleg Karpov³, Danica-Mae Manalo³, Berhan Mandefro¹³, Andrea Matlock³, Rakhi Pandey³, Niveda Sundararaman³, Hannah Trost¹³, Vineet Vaibhav³, Vidya Venkatraman³, Oliver Wang³, Jonathan D. Glass¹⁵, Arish Jamil¹⁵, Naufa Amirani^{4,5}, Leandro Lima^{4,5}, Krishna Raja^{4,5}, Wesley Robinson^{4,5}, Reuben Thomas^{4,5}, Edward Vertudes^{4,5}, Stacia Wyman^{4,5}, Carla Agurto¹⁶, Guillermo Cecchi¹⁶, Raquel Norel¹⁶, Omar Ahmad^{7,8}, Emily G. Baxi^{7,8}, Aianna Cerezo², Alyssa N. Coyne^{1,2}, Lindsey Hayes², John W. Krakauer², Nicholas Maragakis², Elizabeth Mosmiller², Promit Roy², Steven Zeiler², Miriam Adam¹, Noura Albistami^{4,5}, Aneesh Donde¹, Tobias Ehrenberger¹, Nhan Huynh¹, Velina Kozareva¹, Connie New¹, Alex Lenail¹, Jonathan Li¹, Natasha Leanna Patel-Murray¹, Yogindra Raghav¹, Divya Ramamoorthy¹, Egun Im¹, Karen Sachs¹, Brook T. Wassie¹, James Berry¹⁷, Merit E. Cudkowicz¹⁷, Alanna Farrar¹⁷, Sara Thrower¹⁷, Sarah Luppino¹⁷, Lindsay Pothier¹⁷, Alexander V. Sherman¹⁷, Ervin Sinani¹⁷, Prasha Vigneswaran¹⁷, Hong Yu¹⁷, Jay C. Beavers¹⁸, Mary Bellard¹⁹, Elizabeth Bruce¹⁹, Senda Ajroud-Driss²⁰, Deniz Alibazoglu²⁰, Ben Joslin²⁰, Matthew B. Harms²¹, Sarah Heintzman²¹, Stephen Kolb²¹, Carolyn Prina²¹, Daragh Heitzman²², Todd Morgan²², Ricardo Miramontes¹¹, Jennifer Stocksdale²³, Keona Wang²³, Jennifer Jockel-Balsarotti²⁴, Elizabeth Karanja²⁴, Jesse Markway²⁴, Molly McCallum²⁴, Tim Miller²⁴ & Jennifer Roggenbuck²⁵

¹⁴On Point Scientific Inc, San Diego, CA, USA. ¹⁵Department of Neurology, Emory University, Atlanta, GA, USA. ¹⁶Computational Biology Center, IBM T.J. Watson Research Center, Yorktown Heights, NY, USA. ¹⁷Department of Neurology, Healey Center, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. ¹⁸Microsoft Research, Microsoft Corporation, Redmond, WA, USA. ¹⁹Microsoft University Relations, Microsoft Corporation, Redmond, WA, USA. ²⁰Department of Neurology, Northwestern University, Chicago, IL, USA. ²¹Department of Neurology and Genetics, Ohio State University Wexner Medical Center, Columbus, OH, USA. ²²Texas Neurology, Dallas, TX, USA. ²³Department of Psychiatry and Human Behavior and Sue and Bill Gross Stem Cell Center, University of California, Irvine, CA, USA. ²⁴Department of Neurology, Washington University, St. Louis, MO, USA. ²⁵Zofia Consulting, Reston, VA, USA.