Title

Large-scale differentiation of iPSC-derived motor neurons from ALS and control subjects

Permalink

https://escholarship.org/uc/item/8b71x7n0

Journal

Neuron, 111(8)

ISSN

0896-6273

Authors

Workman, Michael J
Lim, Ryan G
Wu, Jie
et al.

Publication Date

2023-04-01

DOI

10.1016/j.neuron.2023.01.010

Published in final edited form as:
Neuron. 2023 April 19; 111(8): 1191–1204.e5. doi:10.1016/j.neuron.2023.01.010.

Large-scale differentiation of iPSC-derived motor neurons from ALS and control subjects

Michael J. Workman[1,17], **Ryan G. Lim**[2,17], **Jie Wu**[3,17], **Aaron Frank**[4], **Loren Ornelas**[4], **Lindsay Panther**[4], **Erick Galvez**[4], **Daniel Perez**[4], **Imara Meepe**[4], **Susan Lei**[4], **Viviana Valencia**[4], **Emilda Gomez**[4], **Chunyan Liu**[4], **Ruby Moran**[4], **Louis Pinedo**[4], **Stanislav Tsitkov**[5], **Ritchie Ho**[1,6,7,8], **Julia A. Kaye**[9,10,11],
Answer ALS Consortium,
Terri Thompson[12], **Jeffrey D. Rothstein**[13,14], **Steven Finkbeiner**[9,10,11], **Ernest Fraenkel**[5], **Dhruv Sareen**[1,4,*], **Leslie M. Thompson**[2,3,15,16,*], **Clive N. Svendsen**[1,4,18,*]

[1]The Board of Governors Regenerative Medicine Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA

[2]Institute for Memory Impairments and Neurological Disorders, University of California, Irvine, CA, USA

[3]Department of Biological Chemistry, University of California, Irvine, CA, USA

[4]Cedars-Sinai Biomanufacturing Center, Cedars-Sinai Medical Center, Los Angeles, CA, USA

[5]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

[*]Correspondence: dhruv.sareen@cshs.org (D.S.), lmthomps@uci.edu (L.M.T.), clive.svendsen@cshs.org (C.N.S.).

AUTHOR CONTRIBUTIONS

Conceptualization, J.D.R., S.F., E.F., D.S., L.M.T., and C.N.S.; data curation and project administration, L.O. and T.T.; software and formal analysis, M.J.W., R.G.L., J.W., S.T., R.H., and J.A.K.; funding acquisition, J.D.R., L.M.T., and C.N.S.; investigation, A.F., L. Panther, E. Galvez, D.P., I.M., S.L., V.V., E.Gomez, C.L., R.M., and L. Pinedo; supervision, S.F., E.F., D.S., L.M.T., and C.N.S.; visualization, M.J.W., R.G.L., J.W., and S.T.; writing – original draft, M.J.W. and C.N.S.; writing – review & editing, M.J.W., R.G.L., J.W., S.T., L.M.T., and C.N.S.

DECLARATION OF INTERESTS

US patent 10,221,395-B2 has been granted describing a method for reprogramming blood to induced pluripotent stem cells. Apart from this issued patent filing, the authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.neuron.2023.01.010.

CONSORTIA

Additional members of the Answer ALS Consortium are Robert Baloh, Maria G. Banuelos, Veronica Garcia, Ronald Holewenski, Oleg Karpov, Danica-Mae Manalo, Berhan Mandefro, Andrea Matlock, Rakhi Pandey, Niveda Sundararaman, Hannah Trost, Vineet Vaibhav, Jennifer E. Van Eyk, Vidya Venkatraman, Oliver Wang, Jonathan D. Glass, Arish Jamil, Naufa Amirani, Leandro Lima, Krishna Raja, Wesley Robinson, Reuben Thomas, Edward Vertudes, Stacia Wyman, Carla Agurto, Guillermo Cecchi, Barry Landin, Raquel Norel, Omar Ahmad, Emily G. Baxi, Aianna Cerezo, Alyssa N. Coyne, Lindsey Hayes, John W. Krakauer, Nicholas Maragakis, Elizabeth Mosmiller, Promit Roy, Steven Zeiler, Miriam Adam, Noura Albistami, Aneesh Donde, Tobias Ehrenberger, Nhan Huynh, Velina Kozareva, Alex Lenail, Jonathan Li, Natasha Leanna Patel-Murray, Yogindra Raghav, Divya Ramamoorthy, Karen Sachs, Brook T. Wassie, James Berry, Merit E. Cudkowicz, Alanna Farrar, Sara Thrower, Sarah Luppino, Lindsay Pothier, Alexander V. Sherman, Ervin Sinani, Prasha Vigneswaran, Hong Yu, Jay C. Beavers, Mary Bellard, Elizabeth Bruce, Senda Ajroud-Driss, Deniz Alibazoglu, Ben Joslin, Matthew B. Harms, Sarah Heintzman, Stephen Kolb, Carolyn Prina, Daragh Heitzman, Todd Morgan, Ricardo Miramontes, Jennifer Stocksdale, Keona Wang, Jennifer Jockel-Balsarotti, Elizabeth Karanja, Jesse Markway, Molly McCallum, Tim Miller, and Jennifer Roggenbuck.

[6]Center for Neural Science and Medicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA

[7]Department of Neurology, Cedars-Sinai Medical Center, Los Angeles, CA, USA

[8]Department of Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, CA, USA

[9]Center for Systems and Therapeutics, Gladstone Institutes, University of California, San Francisco, San Francisco, CA, USA

[10]Taube/Koret Center for Neurodegenerative Disease, Gladstone Institutes, University of California, San Francisco, San Francisco, CA, USA

[11]Departments of Neurology and Physiology, University of California, San Francisco, San Francisco, CA, USA

[12]On Point Scientific Inc., San Diego, CA, USA

[13]Brain Science Institute, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[14]Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[15]Department of Neurobiology and Behavior, University of California, Irvine, CA, USA

[16]Department of Psychiatry and Human Behavior and Sue and Bill Gross Stem Cell Center, University of California, Irvine, CA, USA

[17]These authors contributed equally

[18]Lead contact

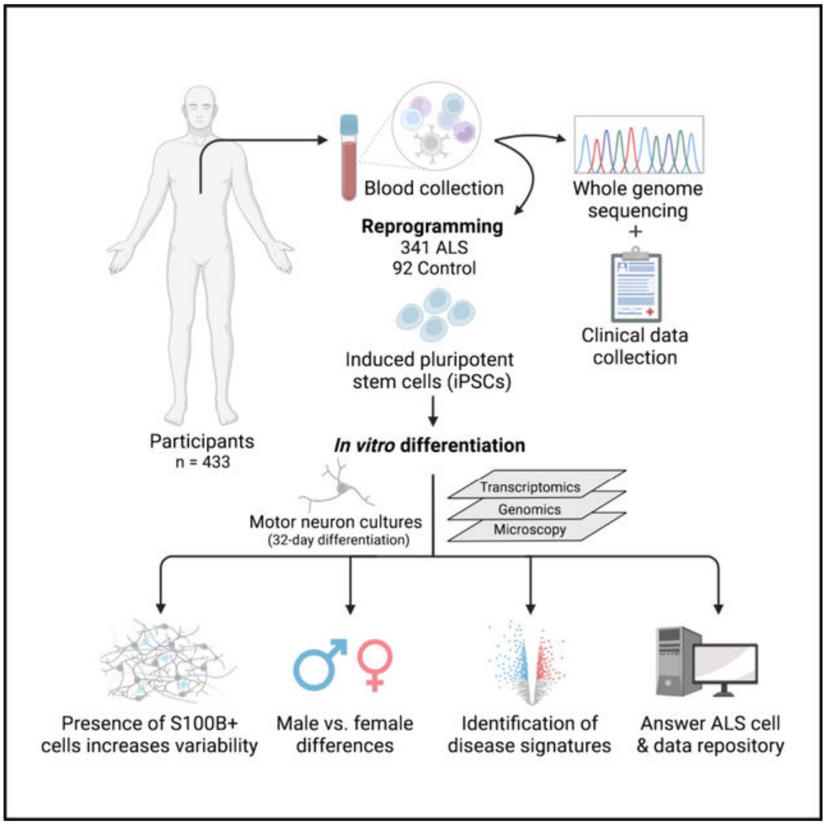
SUMMARY

Using induced pluripotent stem cells (iPSCs) to understand the mechanisms of neurological disease holds great promise; however, there is a lack of well-curated lines from a large array of participants. Answer ALS has generated over 1,000 iPSC lines from control and amyotrophic lateral sclerosis (ALS) patients along with clinical and whole-genome sequencing data. The current report summarizes cell marker and gene expression in motor neuron cultures derived from 92 healthy control and 341 ALS participants using a 32-day differentiation protocol. This is the largest set of iPSCs to be differentiated into motor neurons, and characterization suggests that cell composition and sex are significant sources of variability that need to be carefully controlled for in future studies. These data are reported as a resource for the scientific community that will utilize Answer ALS data for disease modeling using a wider array of omics being made available for these samples.

In brief

In this article, 433 iPSCs from control and ALS patients were differentiated into motor neurons and profiled by immunocytochemistry and RNA-seq. Results reveal sex and cell composition to be among the strongest drivers of gene expression variation, with a higher percentage of Islet1+ cells observed in male ALS patient-derived cultures.

Graphical Abstract



INTRODUCTION

The use of induced pluripotent stem cell (iPSC)-derived motor neurons to model genetic and sporadic forms of amyotrophic lateral sclerosis (ALS) is founded on the idea that the human-specific nature of these models will help to elucidate molecular mechanisms of disease and provide a paradigm for drug screening and discovery.[1] Although it may be possible to use a smaller number of patient lines when modeling genetic forms of ALS, where isogenic lines can be produced as controls,[2] this becomes more complicated when modeling sporadic ALS (SALS) with mixed etiologies. One study from Japan has shown that 32 ALS patients compared with 6 controls was sufficient to distinguish specific subsets of SALS patients.[3] However, differences in ethnic diversity between the USA and Japan may create a wider variation in iPSC biology that could increase the number of patients required to resolve additional disease-specific mechanisms.[4] In order to overcome and inform these issues fora diverse population of patients, a far larger set of iPSC lines and their differentiated motor neurons are required. Answer ALS (https://www.answerals.org) was established with the goal of making 1,000 iPSC lines from patients with all forms of ALS that have deep clinical data along with whole-genome sequencing (WGS) and multi-omics on differentiated motor neurons.[5]

In this study, 433 iPSC lines were differentiated into motor neurons using an optimized 32-day protocol selected based on extensive discussions with the scientific community.[6] Neuronal cell marker staining and gene expression patterns were analyzed in motor neuron cultures. Exploratory analyses and modeling with patient covariates were then used to reveal the major confounding variables and challenges associated with these large patient-derived iPSC differentiation studies as resources for the community. Interestingly, ALS participants generated significantly more motor neurons than control patients based on Islet 1 (ISL1) staining, but unsupervised principal component analysis (PCA) showed no clear overall separation of ALS versus control participants, and very few differentially expressed genes (DEGs) were detected. However, the correlation of principal components (PCs) with biological and technical covariates revealed high correlations of various PCs with the percentage of S100B+ cells. Additional subsets of genes were highly correlated with the staining data and other experimental and patient-specific covariates, such as iPSC origin (T cell versus non-T cell), *C9orf72* expansion, and a large number of residual genes that represent patient-to-patient variability likely driven by genetic differences. Interestingly, among the most prominent covariates was sex, which drove a transcriptomic signature that could completely separate males and females and was associated with a large number of differentially expressed autosomal and sex-linked genes.

RESULTS

Characterization of the main resource

The experimental design is summarized in Figure 1A and the concept and an overview of the Answer ALS project has been recently published.[5] For this study, peripheral blood mono-nuclear cells (PBMCs) were collected from 433 individuals (Table S1) enrolled in the Answer ALS clinical study and reprogrammed into iPSCs using established methods.[7] Within the ALS patient lines, there was the expected distribution of self-reported disease-causing mutations in genes such as *C9orf72* and superoxide dismutase 1 (*SOD1*), with the majority of cases being sporadic with no known disease-causing mutations (Figure 1B). WGS revealed similar numbers of pathogenic/likely pathogenic (P-LP) and *in silico* predicted damaging (IS-D) coding variants across the genomes of healthy control, SALS, and familial ALS (FALS) patients (Table 1). However, more subjects in the FALS group harbor P-LP (28.1%) and IS-D (40.6%) variants in 33 ALS-associated genes (see STAR Methods for genes) compared with other patient groups. Unexpectedly, several healthy control participants also carry P-LP (15.4%) or IS-D (9.6%) coding variants in one or more of the ALS genes, similar to the proportion of SALS patients carrying P-LP (14.6%) and IS-D (11.1%) ALS variants.

Control and ALS patient iPSCs were differentiated using a standardized 32-day direct induced motor neuron protocol in 6-well plates in 50 batches of 5–14 patients using a method based on dual SMAD inhibition and ventral patterning[6] (Figure S1A). In order to determine any drift within the differentiation protocol or confounding technical variability in the RNA sequencing (RNA-seq) analysis, we included a set of batch technical controls (BTCs—a control line differentiated in bulk, aliquoted, and included with each RNA-seq run as technical reference) and batch differentiation controls (BDCs—the same control line differentiated with each batch of patient lines) (Figures 1A and 1C). Replicate wells were dissociated and used for the RNA analysis and immunocytochemistry, which was performed and analyzed using high-content imaging. Differentiated motor neuron cultures were heterogeneous and consisted mainly of neurons based on staining for SMI-32/neurofilament heavy chain and tubulin beta-III (TUBB3) (Figure 1D). Although there were no observed GFAP-positive cells resembling reactive astrocytes, most cultures had cells with a flat morphology that were highly reactive for S100B, suggestive of an astrocyte-like phenotype[8] (Figure 1D). Positive staining for ISL1 suggested cultures contained putative motor neurons (Figure 1D). Surprisingly, ALS cultures had significantly more ISL1+ motor neurons than controls—17.0% ± 0.4% versus 14.1% ± 0.8% of total cells (mean ± SEM) (Figure 1E). Further quantification and separation into male and female lines showed that the higher percentage of ISL1 in ALS samples was primarily due to a significantly higher percentage of ISL1+ cells specifically in male ALS samples versus all other groups (Figure S1B). Additionally, the motor neuron progenitor marker NKX6–1 (Figure S1C) and the general neuronal marker TUBB3 (Figure S1D) were also similarly enriched in male ALS samples versus all other groups. Quantification of SMI-32, Nestin, and S100B staining showed no significant differences between ALS and control groups or between male and female participants (Figures S1E–S1G). Temporal analysis of a subset of lines (n = 8) revealed S100B+ cells to be present as early as day 18 and significantly more abundant at day 32 and day 46 in motor neuron cultures (Figure S1H).

T cell and non-T cell reprogramming leads to specific gene expression in motor neuron cultures

Reprogramming PBMCs into iPSCs was based on methods that favored both lymphoid T cells and myeloid non-T cells.[7,9] We initially attempted to avoid T cells; however, reprogramming non-T cells was less efficient, and hence, using solely non-T cells was not feasible for iPSC production at scale. As such, iPSC lines in the Answer ALS cohort were derived from both T cells and non-T cells (Figure 2A). Given the genetic DNA rearrangements that occur in T cell development, we first asked whether the original cell type from which the iPSCs were derived had any effect on gene expression in the differentiated motor neurons. Interestingly, T cell receptor genes *TRDC* and *TRGV9* were significantly upregulated, and *TRBC2* and *TRAC* were significantly downregulated in non-T versus T cell-derived samples (Figures 2B and 2C). PCA based on these four genes showed a clear separation of motor neuron samples along PC1 depending on the cell of origin (Figure 2D). Receiver operating characteristic (ROC) curve analysis using PC1 coordinates to classify samples revealed these 4 genes to be outstanding predictors of the original somatic cell of origin of each iPSC-derived motor neuron sample (Figure 2E). Interestingly, there were occasionally some differentiated motor neuron cultures with abnormally high levels of *POU5F1/OCT4* expression (Figure 2F), which were nearly all in cultures differentiated from T cell-derived iPSCs (Figure 2G). These were also clearly outliers on the global PCA generated using all expressed genes (Figure 2H) and were removed from the analyses.

Overall gene expression patterns and controlling for batch differences

Using PCA to determine overall gene expression patterns revealed that in the final set of patients, PC1 accounted for 44.6% of the variation in gene expression (top 500 highly variable genes [HVGs]) with PC2 representing 12.5% of the variation (Figure 3A). There was no clear separation of ALS patients from controls in either of these PCs or any separation of genetic and non-genetic forms of ALS. However, two distinct clusters occurred in the PCA completely separated along PC2, which were subsequently determined to be male and female participants. We next determined how consistent the differentiation method was by projecting the batch control samples onto the original PCA, which resulted in the BTCs clustered almost entirely in one location (Figure 3B, green) with a mean Spearman correlation coefficient of 0.98 (Figure 3C), suggesting very little technical variability in the RNA-seq assay due to sample prep, sequencing chemistry, and read processing. Also, most isogenic BDCs clustered in a similar region of the PCA (Figure 3B, salmon) with a mean Spearman correlation coefficient of 0.97 (Figure 3C), indicating high reproducibility of the method across 49 independent differentiations. Although there was overall individual patient variation spread across PC1 and PC2 (Figure 3B, gray), ingeneral, the anisogenic individual patient samples also correlated well with each other (Spearman correlation = 0.95; Figure 3C), suggesting that the differentiation protocol provides robust, high-quality, and consistent differentiations, even across patient cell lines. These results are well within the ENCODE recommended replicate concordance among isogenic and anisogenic samples.[10]

We next asked which genes contributed to variation in the BTCs and BDCs. In the BTCs, small nucleolar RNAs and mitochondrial genes were the most variable (Figure 3D), likely reflecting variation associated with measuring genes with large expression values. Gene ontology (GO) analysis of the HVGs in the BTCs (Table S2) showed significant enrichment for mitochondrial-associated GO terms (Figure S2A). This was similar for the BDCs; however, there was also high variation in collagens, cadherins, and contactins along with genes involved in Wnt and Notch signaling (Figure 3E; Table S3), suggesting that expression of these genes are susceptible to small differences in the differentiation parameters (e.g., plating density, exact media composition, and substrate differences between batches). Enriched GO terms in the BDC variable genes confirmed that extracellular matrix (ECM) and cell adhesion pathways were the dominant drivers of variation (Figure 3F). A heatmap of the top 20 HVGs in the BDCs revealed that a select few early batches and the very last differentiation batch drove much of the variation and were enriched in the expression of the top HVGs (Figure 3G). Interestingly, these highly variable batches coincided with the BDC samples having the highest percentage of S100B+ cells by staining (Figure S2B), suggesting that the high variability in these genes and pathways is likely attributable to differences in cell composition.

To establish whether the gene expression patterns represented in the PCA were reproducible for additional patients, 26 subjects were re-run with 13 control and 13 ALS samples split roughly equally between males and females. iPSCs were thawed and 32-day motor neurons were differentiated for the 26 samples in 3 sets of repeat batches. Raw counts of the first and second differentiation of subjects were combined and normalized. Unsupervised hierarchical clustering revealed that 9 of the repeat sample pairs clustered most closely together while the remaining pairs were more distant (Figure S2C). Multidimensional scaling revealed that many of the repeat samples clustered together as nearest neighbors (Figure S2D), and additional estimates of similarity by Spearman's correlation (Figure S2E) or simple error ratio estimate (SERE)[11] scores (Figure S2F) showed that gene expression for the intra-individual (self) repeats was significantly more correlated than inter-individuals (non-self). These results suggest that the repeat pairs were more similar than would have happened by chance; however, it was clear that for over half the cases there were stochastic factors that drove the same patient line to differentiate into a slightly different culture composition.

Main drivers of variance in motor neuron cultures

In order to determine the covariates contributing most to individual gene variation, we used variance partitioning[12] to estimate the percentage of gene expression that could be explained by technical and clinical variables associated with each sample. The differentiation batch was responsible for an average of 8.7% of the overall variance in gene expression—the highest percentage of all the covariates that were assessed (Figure 4A). The next highest contribution to variance was the number of S100B+ cells in each culture. Interestingly, analysis of the staining data revealed a high correlation between glial/progenitor markers (S100B and Nestin) and among neuronal markers (TUBB3, SMI-32, ISL1, and NKX6–1), as shown in the pairwise correlation matrix (Figure 4B). Feature selection of the top 500 HVGs was performed to identify the strongest drivers of sample clustering and find any PCs significantly correlating with various patient clinical and experimental covariates. Analysis of 12 covariates within the first 6 PCs revealed that the percentage of S100B+ cells in each culture was significantly correlated with PC1 (Figure 4C). Plotting PC1 and PC2 and coloring samples by percent S100B+ cells reveals that samples are largely ordered along PC1 by the level of S100B cells in each culture (Figure 4D). Surprisingly, there was no strong correlation of batch with any of the first 6 PCs despite this variable contributing significantly to gene expression variation. These findings suggest that gene variance associated with batch-to-batch variation is somewhat random and that the global clustering of samples is more dependent on cell composition and other participant variables.

We next examined genes that correlate with S100B staining and identified several in which the percent of S100B+ cells could explain over 50% of the variation in gene expression (Figure S3A). Interestingly, several genes including *COL5A2* were more highly correlated with %S100B+ staining than the expression of *S100B* itself (Figure S3B), suggesting that the RNA-protein relationship for this gene and many others is not always tightly correlated. In other cases, such as for ISL1, there was a more robust gene-protein relationship where staining explained variation in *ISL1* gene expression better than any other gene (Figures 4E and 4F). For other staining markers, NKX6–1 protein was also highly correlated with its own RNA, whereas Nestin protein staining was most correlated with genes such as *CHMP2B*, and TUBB3 staining with *EEF1A1* and *BDNF* gene expression (Figure S3C). Interestingly, *C9orf72* mutation status best-explained variance in *C9orf72* gene expression (Figure S3D), and patients with *C9orf72* hexanucleotide repeat expansion (HRE) had significantly less overall *C9orf72* gene expression (Figure 4G), supporting previous studies showing a loss of function due to the expanded repeats.[13–15] Comparing *C9orf72* patients with healthy controls or all other ALS patients revealed only a small number of DEGs (Figure S3E). The sample batch explained significant amounts of variation in ribosomal and small nucleolar RNA genes (Figure S3F) that overlapped with many of the HVGs present in the BTC and BDC samples. A large proportion of variation in several genes could be explained by sequencing depth (Figure S3G), and notably, sex could explain nearly all the variation in a number of genes, many being X- and Y-linked (Figure S3H). The disease status of the participant (ALS versus control) only explained an average of 0.22% of the variation in total gene expression (Figure 4A) and a maximum of approximately 15% of the variation of any single gene (Figure S3I). Lastly, correlating the age of the patients at iPSC sample collection with day 32 motor neuron gene expression showed very little association, with age at sample collection only able to explain a maximum of 5.7% of variation in any individual gene, the lowest of any of the covariates we tested (Figure S3J), supporting that age of the patient at iPSC production only minimally affects downstream gene expression of iPSC-derived tissues.

Gene expression related to sex separated males and females in motor neuron cultures

A large subset of the HVGs that separated motor neuron samples was associated with sex. Following the top 500 HVG feature selection, males and females were completely separated along PC2 (Figure 5A), suggesting that even with all the other variations from cell composition and batch differences, this covariate had a significant effect. Interestingly, even in the absence of *in vivo* hormonal signaling, iPSC-derived motor neurons have strong gender-specific gene expression patterns. To determine whether this sex effect was also observed in human tissue *in vivo*, we evaluated RNA-seq data from post-mortem human samples from ALS subjects using data from Prudencio et al.[16] and from the New York Genome Center. In brain samples,[16] the largest separation was based on the tissue profiled (frontal cortex versus cerebellum), with sex separating subjects along PC2 with 5% of the variance (Figure 5B). Sex also separated human post-mortem thoracic spinal cord samples along PC2, which accounted for 9% of the variance in gene expression (Figure 5C). In both cases, there was a clear separation of samples according to sex, validating that sex differences in neuronal gene expression are not simply an artifact of iPSC reprogramming and differentiation. The distinct clustering of male versus female samples appeared to be driven mainly by X- and Y-linked genes; however, PCA using only autosomal genes revealed weaker but still significant correlations between sex and PC clustering (Figure S4A). Plotting samples along the two highest correlated autosomal-only PCs (PC4 and PC7) revealed significantly more overlap between sexes (Figure S4B) compared with PCA including sex-linked genes (Figure 5A); however, differential gene expression testing between male and female samples uncovered 1,016 DEGs, the majority of which were autosomal (Figure 5D; Table S4). To confirm these findings were not simply a product of random sample noise, sample sex labels were randomly shuffled resulting in an average of 53% correctly labeled samples (Figure 5E) and only 4 DEGs (Figure 5F) across 467 permutations, suggesting that the large number of DEGs found in male-female comparison are true biological sex differences. *XIST* expression was significantly higher in female samples compared with males (Figure 5G), but female samples displayed higher variation in expression with several samples exhibiting low *XIST* levels, possibly reflecting issues with X chromosome inactivation (XCI). However, an analysis of XCI status confirmed that nearly all female samples display normal XCI with X-linked gene dosage comparable to males (Figure S4C), which is in line with previous reports assessing XCI in human iPSCs and their differentiated progeny.[17] By comparing the expression level of each X-linked gene in females with the average expression in male samples (Figure S4D), we identified a small number of female samples (n = 4) that display aberrant expression of X-linked genes and have an XX:XY expression ratio >1.2, indicating some erosion of XCI. Given the small number of samples that appear to be affected, defects in XCI are unlikely to be a major driver of the sex-dependent differences we observe in our dataset. As expected, XX:XY expression ratios were negatively correlated with *XIST* levels (Figure S4E), but surprisingly, several female samples maintained normal X-linked gene dosage despite expressing low levels of *XIST* (Figure S4F). Conversely, many Y-linked genes such as *ZFY* were significantly enriched in male samples (Figure 5G); however, the gene most often associated with sex determination, *SRY*,[18] was very lowly expressed and only detectable in 78 of 252 male samples (Figure S4G), suggesting alternative drivers of sexual dimorphism in motor neurons. Mapping of the DEGs between males and females shows they are distributed across all chromosomes with several interesting genes displaying sex-specific differences (Figure S4H). For example, the neurodevelopmental gene *DCX* and the Alzheimer's-associated *APOE* gene had higher expression in females (Figure 5H), whereas genes encoding cytoskeletal intermediate filaments such as *NEFM* and *DES* were higher in males (Figure 5I).

ALS gene signatures in motor neuron cultures

Given the significant sex differences in day 32 motor neurons and lack of a clear ALS versus control difference when using all participants, differential gene expression testing was run separately in males and females. Although very few DEGs were detected when using all participants or females only, 132 genes were downregulated and 220 genes were upregulated in male ALS versus male controls (Figures 5J and 5K; Table S5). No significant pathway enrichment was detected in the downregulated genes; however, inflammatory pathways related to TNF and NF-κB signaling were significantly enriched in the upregulated genes in male ALS samples (Figure 5L). Given the emergence of dysfunctional cryptic splicing as a potentially fundamental ALS disease mechanism, we also analyzed samples for cryptic exon (CE) inclusion in two commonly associated ALS genes: *STMN2*[19,20] and *UNC13A*.[21,22] Although we did not detect significant *STMN2* CE inclusion in any of the day 32 motor neuron cultures (data not shown), most samples did have detectable *UNC13A* CE inclusion (Figure 5M). However, the percent spliced in (PSI, ψ) of *UNC13A* CE was quite low at less than 0.1% PSI in all samples with no statistically significant differences observed between ALS and control or between males and females.

We also sought to identify any genes that correlate with ALS patient clinical data that could potentially be used as predictive classifiers. We therefore performed a random 80/20 split of the ALS patient samples into training (n = 264) and validation (n = 66) groups and performed variance partitioning in the training group using revised ALS Functional Rating Scale (ALSFRS-R) progression rate, age at symptom onset, and site of disease onset as covariates (Figure S5A). This allowed us to identify several genes in which a portion of the variation of expression in day 32 iPSC-derived motor neurons could be explained by the patient clinical data, including a number of genes related to ALSFRS-R progression rate (Figure S5B). Several of these genes such as *HSPBAP1* and *NUP188* were positively correlated with ALSFRS-R slope, with lower expression correlating with faster disease progression (Figure S5C). Interestingly, *HSPBAP1* has previously been observed to be downregulated in ALS patient motor cortex samples compared with healthy controls[23,24] and *NUP188* has been identified as a TDP-43 target.[25] Several genes were also negatively correlated with ALSFRS-R slope (Figure S5D). ALSFRS-R progression rates ranged from approximately −3.0 to + 0.8 in the training set (Figure S5E), with the median value of −0.54 used to delineate fast and slow progressors. Using the top 7 genes that most highly correlated with ALSFRS-R slope, we generated PCA plots of ALS samples in the training set (Figure S5F) and used ROC curves to estimate the performance of the model to separate fast and slow progressors along PC1. Using this approach, we obtained a borderline acceptable classifier model for predicting disease progression rates from day 32 motor neuron gene expression (AUC = 0.66, p = 3e–4; Figure S5G). The distribution of ALSFRS-R progression rates was similar in the validation set (Figure S5H) and patients were plotted along PC1 (Figure S5I) according to the expression of the 7 genes identified in the training group. Applying the classifier model to this set confirmed the significant, but borderline acceptable predictive ability of the model to separate fast and slow progressors using an independent cohort of patients (AUC = 0.68, p = 0.042; Figure S5J). We also analyzed several genes that have been previously identified by GWAS and expression quantitative trait loci (eQTL) studies to correlate with disease progression (*TTN*[26]) or disease onset (*ACSL5*[27] and *ZNF512B*[28]) and found no correlation in our dataset (Figure S5K).

DISCUSSION

This resource article describes the large-scale differentiation of motor neurons from ALS and control iPSC lines and the subsequent profiling by cell marker expression and bulk RNA-seq, with the goal of understanding whether there are biological or technical covariates that may be critical to subsequent data analytics and disease subtyping. Through extensive quality control measures, we showed three major sources that could account for variation related to the reprogramming process and/or differentiation to motor neurons: (1) differences between T cell and non-T cell reprogramming, (2) occasional aberrant persistent expression of *OCT4/POU5F1*, and (3) a small subset of cells expressing S100B. Further, we found that sex separated out the data into two distinct subgroups independent of disease status. Intriguingly, significantly more ISL1+ motor neurons were found in the ALS cases compared with controls, which when further analyzed was driven partly by the finding that males in general generated more motor neurons compared with females.

T cells versus non-T cells

Numerous methods exist to generate iPSCs from PBMCs. Isolating and expanding CD34-positive cells has been used by several groups[29] but requires *ex vivo* expansion of cells before reprogramming, which increases the possibility of proliferation-related genomic alterations. We previously developed a protocol for direct reprogramming of PBMCs without CD34 expansion.[7] Given that T cells have genetic rearrangements required for a diverse repertoire of T cell receptors, we initially attempted to avoid reprogramming T cells. However, to meet the demand for large-scale generation of iPSC lines for the Answer ALS program, we ultimately reprogrammed both non-T cells and T cells into iPSCs. This proved more practical, but we observed that several T cell receptor-related genes were differentially expressed in motor neuron cultures. One potential problem for iPSC differentiation is the persistence of *OCT4/POU5F1* or other pluripotency genes in differentiated cells. We screened for reprogramming factors and as expected, nearly all were downregulated following differentiation except for occasional lines that continued to highly express *OCT4*. Over 90% of high *OCT4*-expressing motor neuron cultures were from iPSCs derived from T cells, although mechanisms underlying this phenomenon are unclear. Many iPSC lines in the ALS disease modeling community have been fibroblast-derived. However, the improved cytogenetic stability of PBMC-derived versus fibroblast-derived iPSCs,[9] combined with their ease of collection and lack of expansion requirement, make PBMCs a preferable source of iPSCs at scale. Additionally, a patient's genetics typically override the differences caused by a somatic cell of origin, and differentiation propensities are not suspected to be influenced by whether the iPSC line was fibroblast or blood-derived[30] suggesting that the PBMC source of iPSCs in this dataset should not confound comparisons with other datasets using fibroblast-derived iPSCs.

Challenges with differentiation at scale

Differentiating a large number of iPSC lines to a specific neuronal phenotype is difficult at-scale and for practical reasons required batches of cells to be differentiated over an extended time period. When a single line was differentiated, split into reference vials, and used with each RNA-seq assay as a technical control, variation was very low, indicating the RNA-seq assay is very consistent. As a further control, the same patient line was differentiated with every batch (BDC) as an anchor to ensure the robustness of each round of differentiation. In general, the protocol was remarkably reliable with batch controls falling very closely together between each run. Interestingly, most genes that differed with multiple runs of the same participant line were related to ECM, cell adhesion, and Hox patterning. ECM has critical effects on neuronal differentiation,[31] and it is possible that subtle differences in plating density, substrates, and other media components led to changes in differentiation reflected in random changes in ECM gene expression and anterior-posterior patterning. In addition to the gene expression profiles, immunocytochemistry demonstrated that there was a stochastic appearance of larger flat cells in some of the batch controls that did not have neuronal morphology and were positive for S100B—a calcium-binding protein mainly concentrated in astrocytes and activated following brain injury or in neurodegenerative disease.[8] We found that the percentage of S100B+ cells was highly correlated with many collagen and ECM-related genes, suggesting the variable number of S100B cells may have contributed to the ECM changes and variation in batch control differentiations. The tightly associated expression of ECM genes such as collagens, syndecans, and matrilins with the percentage of S100B+ cells, combined with their large, flat morphology and lack of GFAP staining, suggests that these cells may more closely resemble non-neuronal glial or fibroblast-like cells.

S100B expression and sex are major covariates

PCA of all samples to identify clusters that correlated with any of the covariates showed two variables that displayed high correlation values (>0.7) and statistically significant correlations with various PCs. The first was percent S100B, which was highly correlated with PC1, and the second was sex along PC2, which could completely separate male and female samples. S100B was primarily found in non-neuronal cells in the cultures, indicating that controlling this aspect of differentiation could significantly decrease variation. This may be accomplished in future runs by cell sorting to eliminate S100B+ cells before the final analysis. The sex differences were interesting given that, in this model, cells are not exposed to exogenous sex hormones, so any changes must come from the endogenous differences mediated by X and Y chromosomes. Indeed, many gene expression differences were associated with genes on the sex chromosomes; however, most DEGs were autosomal, suggesting a wider association of sex differences across the genome. It is known that the incidence of ALS is higher in men (average male:female ratio ≈ 1.3–1.5) and the onset of disease is earlier in men than women.[32,33] In addition, male rodents with a SOD1 mutation have earlier onset and faster progression than females.[34,35] Differences in onset and progression are intriguingly not affected by modifying hormonal interactions,[34] suggesting there are intrinsic male-female differences that lead to sexually dimorphic ALS disease presentation that are unrelated to hormonal signaling. Despite differences in incidence, prevalence rates are similar in men and women if examining ALS patients who had not undergone a tracheostomy, but men have a significantly higher prevalence rate if patients who had undergone tracheostomy are included.[36] Correlation of sex to motor phenotypes interestingly reveals an association of flail arm and respiratory phenotypes with males and bulbar phenotype with females, but the bulbar association is only through an interaction with age.[37] A question in the iPSC field relates to whether variability in random X inactivation in female iPSC lines could confound data interpretation.[38–40] Here, the data using iPSC-derived motor neurons show a significant separation between males and females that does not appear to reflect simple variability in gene silencing due to reprogramming given the proper X-linked gene dosage observed in most female samples and the concordance with human post-mortem spinal cord and brain. The data instead reflect widespread inherent gene expression differences between male and female neuronal cells and tissues.[41,42]

Conclusions

Several papers have described cellular phenotypes in iPSC-derived motor neurons from ALS patients, although most have focused on specific ALS-causing mutations such as *C9orf72* and *SOD1*.[1,43] Although a significant increase in motor neurons observed in ALS samples could suggest that overproduction may occur in the disease, this finding could also be the result of unbalanced groups (92 control and 341 ALS). More control lines are needed to further establish this observation. There were clearly no strong global differences between ALS and control, suggesting that using this protocol, ALS and control individuals cannot be easily separated solely by assessing transcriptomic signatures. Furthermore, even when patients are grouped by *C9orf72* HRE, a genetic mutation that explains the largest number of sporadic and familial ALS cases, only a small number of dysregulated genes are detected, highlighting the difficulties in discerning gene expression changes and ALS disease signatures attributable to this mutation. The striking sex differences in our data suggest that males and females may need to be analyzed separately, requiring twice as many subjects for equivalently powered studies. Taken together, it is likely that the inability to detect clear genome-wide clusters among ALS and control individuals using a single assay reflects the issues the field has had in identifying effective therapeutics. Furthermore, ALS is a complex disease with different ages of onset, progression rates, and distinctions between upper or lower motor neurons that may be selectively involved.[44] Although the bulk transcriptomic analysis did not show a clear separation between ALS and control, we have previously shown using single-cell sequencing that motor neurons within complex mixed cultures do show specific gene sets enriched in ALS even when not accounting for sex differences.[45] For bulk data, clearly cell heterogeneity, patient sex, and likely genetic effects need to be accounted for in order to identify subtle disease signatures. Interestingly, independently analyzing males and females revealed several dysregulated genes and pathways specifically in the male ALS patient-derived cells. Given no significant differences were observed in percent S100B+ cells between ALS and control or males and females, the percentage of these cells in each culture is unlikely to be driving the differences. Samples from the same cultures analyzed in this study were processed and frozen for proteomics and ATAC-seq. Ultimately, the immunocytochemistry and RNA-seq data will need to be integrated with the full set of multi-omics data as well as clinical information using machine learning and network-based approaches. This combination may reveal more complex relationships between RNA expression levels, genomics, proteomics, and epigenomics that reveal disease-specific patterns and subgroups of individuals. We have systematically explored the current transcriptomics data for factors contributing to sample variability, which is likely confounding the identification of disease-relevant gene expression changes. Elucidation of these variables should allow the wider scientific community to better utilize our data and is a first step in the overarching goals of integrating all the Answer ALS data for disease modeling and therapy development.

STAR★METHODS

RESOURCE AVAILABILITY

Lead contact—Further information and requests for resources should be directed to the lead contact, Clive Svendsen (clive.svendsen@cshs.org).

Materials availability—The iPSC lines used in this study can be searched and selected through the Answer ALS Data Portal (https://dataportal.answerals.org) which links to the cell line catalog at the Cedars-Sinai Biomanufacturing Center (https://biomanufacturing.cedars-sinai.org) for order fulfillment.

Data and code availability—A large portion of the data used in this study is currently publicly available through the Answer ALS Data Portal (https://dataportal.answerals.org) following approval of a Data Use Agreement (DUA) form. At the time of writing, some sample files are in process of being added to the data portal but have not been formally released. However, all data and code used in this study will be made available by reasonable request to the lead contact following DUA approval.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Clinical data and blood samples were collected from patients at eight neuromuscular clinics across the USA (Cedars-Sinai Medical Center, Johns Hopkins, Massachusetts General Hospital, Ohio State, Emory University, Washington University, Northwestern University, and Texas Neurology) as part of the Answer ALS program. Subjects were allocated to experimental groups based upon clinical diagnosis of ALS. No statistical tests were used for sample size estimation as this was a unique study with no prior knowledge to perform a power analysis on. The study was approved by each site’s Institutional Review Board and patients provided written informed consent.

METHOD DETAILS

iPSC generation and quality control—Patient enrollment, clinical data collection, blood sampling, and reprogramming and QC of iPSC lines has been described previously.[5,7,9] Briefly, 5×10^6 PBMCs isolated from the buffy coat of patient blood samples were nucleofected with plasmids pEP4 E02S ET2K,[46] pCXLE-hOCT3/4-shp53-F,[47] pCXLE-hUL,[47] pCXLE-hSK,[47] and pCXWB-EBNA1[48] using the Amaxa Human T cell Nucleofector Kit and Nucleofector 2D Device. Following nucleofection, cells were resuspended in either T cell media (X-VIVO 10 media supplemented with 30U/mL IL-2 and 5 μl/well Dynabeads Human T-activator CD3/CD28) or non-T cell medium (αMEM supplemented with 10%heat inactivated FBS, 10 ng/ml IL-3, 10 ng/ml IL-6, 10 ng/ml G-CSF, and 10 ng/ml GM-CSF), plated onto mitomycin treated mouse embryonic fibroblasts, and placed in a 37°C incubator. Two days after nucleofection, 2 mL/well of Primate ES Cell Medium supplemented with 5 ng/ml bFGF was added to the wells. Stem cell-like colonies, typically appearing between days 25–32, were mechanically isolated and plated onto Matrigel-coated plates in mTeSR1. Quality control of iPSC lines was carried out as reported[5,7] and consisted of karyotyping, pluripotency testing, EBNA-related gene analysis, and short tandem repeat (STR) confirmation of cell identity.

TCRB and TCRG T cell clonality assay—Total genomic DNA was isolated from iPSC lines using a MagMAX™ DNA Multi-Sample Ultra 2.0 Kit (Applied Biosystems). TCRB and TCRG T cell clonality testing was carried out using an IdentiClone TCRB + TCRG T Cell Clonality Assay Gel Detection kit (Invivo-scribe). PCR products amplifying regions of gene rearrangement and translocation in the TCR-αβ and TCR-γδ loci were analyzed using 6% TBE gel electrophoresis with gel red staining. iPSC lines with any detectable bands were considered T cell derived.

Differentiation of iPSCs to motor neurons—iPSCs in batches of up to 14 patient cell lines were thawed and cultured for 2–3 weeks before passaging for differentiation. Differentiation to motor neurons was performed using an optimized 32-day protocol[6] summarized in Figure S1A and detailed in Baxietal.[5] Briefly, iPSCs were dissociated with Accutase and plated at 5×10^5 cells/well in Matrigel-coated 6 well plates in mTeSR1. Twenty-four hours later, at the start of differentiation (Day 0), mTeSR1 was replaced with Stage 1 media (1:1 mixture of IMDM:F12 supplemented with 1% NEAA, 2% B27 supplement, 1% N2 supplement, 1% PSA, 200 nM LDN193189, 10 mM SB431542, and 3 mM CHIR99021), exchanged daily for 6 days. On day 6, cells were dissociated with Accutase and re-plated at 7.5×10^5 cells/well in Matrigel-coated 6 well plates using Stage 2 media with ROCKi (Stage 1 media supplemented with 0.1 uM ATRA, 1 uM SAG, and 10 uM ROCKi/Y-27632). Twenty-four hours later, media was replaced with Stage 2 media without ROCKi. Stage 2 media was then replaced every other day until day 12. Starting on day 12, Stage 3 media (1:1 mixture of IMDM:F12 supplemented with 1% NEAA, 2% B27 supplement, 1% N2 supplement, 1% PSA, 0.1 uM Compound E, 2.5 uM DAPT, 0.1 uM db-cAMP,0.5 uM ATRA, 0.1 uM SAG,200 ng/ml ascorbic acid, 10 ng/ml BDNF, and 10 ng/ml GDNF) was exchanged every other day until day 32.

RNA isolation, sequencing, and quality control—RNA isolation, QC, preprocessing, and data analysis was performed as previously described.[5] Briefly, total RNA was isolated from each sample using the Qiagen RNeasy mini kit. RNA samples for each subject were entered into an electronic tracking system and processed at the University of California, Irvine GHTF. RNA was QCed using an Agilent Bioanalyzer and quantified by Nanodrop. RNA quality is measured as RIN values (RNA Integrity Number), and 260/280 and 260/230 ratios to evaluate any potential contamination. Only samples with RIN >8 were used for library prep and sequencing. Library prep processing was initiated with total RNA of 1 μg using a Ribo-Zero Gold rRNA depletion and Truseq Stranded total RNA kit. RNA was chemically fragmented and subjected to reverse transcription, end repair, phosphorylation, A-tailing, ligation of barcoded sequencing adapters, and enrichment of adapter-ligated cDNAs. RNA-seq libraries were titrated by qPCR (Kapa), normalized according to size (Agilent Bioanalyzer 2100 High Sensitivity chip). Each cDNA library was then subjected to Illumina (Novaseq 6000) paired end (PE), 100 cycle sequencing to obtain approximately 50–65M PE reads. Fastq were subject to QC and reads with quality scores (>Q20) were collected and further analyzed. Reads were mapped to the GRCh38 reference genome using Hisat2 (v.2.2.1), QCed, and underwent normalization and transformation before further exploratory and differential expression analysis. Samples with suspected mislabeling of sex labels based on PCA clustering and Y-chromosome gene expression, and samples with abnormal expression of OCT4 are indicated in Table S1 and were largely excluded from analyses unless otherwise indicated.

Immunocytochemistry—For cell staining, a replicate plate of day 32 motor neurons was washed with phosphate buffered saline (PBS) and fixed with 4% paraformaldehyde for 10 min at room temperature. After fixing, cells were washed with PBS and blocked for 1 hour at room temperature with 5% normal donkey serum and 0.2% Triton X-100 in PBS. After blocking, each well was incubated with primary antibody for 1 hour at room temperature. Following primary incubation, cells were washed with 0.1% Triton X-100 in PBS and then stained with secondary antibodies for 1 hour at room temperature in the dark. Following secondary incubation, each well was washed with 0.1% Triton X-100 in PBS. DAPI solution was then added for 3 minutes at room temperature to stain nuclei. Cells were then washed again with PBS and stored at 4 °C until image acquisition using an ImageXpress Micro XLS Widefield High-Content Analysis System (Molecular Devices) with 64 regions of interest captured per well. Primary and secondary antibodies and dilutions are listed in the key resources table. ALS-Control and Male-Female comparisons of percentage of cells staining positive for each marker were analyzed in Graphpad Prism using one-way ANOVA with Tukey's HSD post-hoc analysis.

Figure design and visualization—Graphical illustrations depicting study overview and cell differentiation were designed and prepared with Biorender.com and Adobe Illustrator. Graphs, heatmaps, and other plots were generated with R and GraphPad Prism and assembled in Adobe Illustrator.

QUANTIFICATION AND STATISTICAL ANALYSIS

RNA-seq data processing and differential expression testing—Raw RNA-seq reads were mapped to GRCh38 (hg38) human reference genome using Hisat2. Lowly expressed genes with less than 1 average raw count per sample were filtered from the dataset and raw counts were normalized and transformed using the *varianceStabilizingTransformation* pipeline from the R package *DESeq2*. In order to extract disease signals in the presence of strong confounders, we used either stratification and regression approaches. For stratification, samples were subset to males or females only followed by two group comparisons (ALS vs control) using DESeq2. Although lower statistical power, robust comparison was still possible owing to the large cohort of samples. Statistical analyses were performed in *R* and differentially expressed genes were detected for each covariate using false discovery rate or Bonferroni adjustment for multiple testing correction.

Whole genome sequencing and coding variant analysis—Whole genome sequencing on patient PBMCs was performed at the New York Genome Center (NYGC) as previously described.[5] Paired-end reads were aligned to GRCh38 reference and processed following the GATK best practices workflow. Participants with available WGS data were assessed for ClinVar/InterVar/Harms-annotated pathogenic or likely pathogenic (P-LP) coding variants in all genes or specifically in ALS associated genes. A similar assessment of *in silico* predicted damaging (IS-D) variants was also performed. *In silico* predictions were based on previous results generated using SIFT, PolyPhen2, LRT prediction, MutationTaster, MutationAssessor, FATHMM, and dbNSFP prediction tools.[5] ALS-associated genes include: *ALS2, ANG, ANXA11, ATXN2, C21orf2, C9orf72, CAMTA1, CCNF, CHCHD10, DAO, DCTN1, FIG4, FUS, HNRNPA1, HNRNPA2B1, KIF5A, MATR3, MOBP, NEK1, OPTN,* PFN1, *SCFD1, SETX, SOD1, SQSTM1, TAF15, TARDBP, TBK1, TUBA4A, UBQLN2, UNC13A, VAPB*, and *VCP*. Differences between ALS and healthy control populations were determined using a Wilcoxon rank-sum test for total number of coding variants per individual, and a Pearson's chi-squared test for proportion of individuals carrying P-LP and IS-D ALS gene coding variants.

Dimensionality reduction and sample clustering—Principal component analysis (PCA) was performed using *prcomp* function in R with default settings. Following variance stabilizing transformation in DESeq2, the top 500 highly variable genes (HVGs), were used as input for PCA and clustering of samples. All genes, minus those with a mean raw count of <1 per sample, were used as input in PCA for identifying high *OCT4* expressing outlier samples.

Batch control sample analysis—PCA for the batch control samples was performed using the *predict.prcomp* function from the stats package in R to estimate the PC coordinates of the batch control samples within the original principal component space generated using all patient samples without batch controls. For Spearman's rank coefficient of correlation, samples were first normalized using *DESeq2* variance stabilizing transformation and analyzed using the *correlate* function from the *stats* package in R with method = "spearman". To identify highly variable genes (HVGs) in batch control samples, counts were normalized by library size using *DESeq2* and a regression line was fit on the average estimates of gene expression and dispersion. Genes were then ranked by deviation from the fit and a chi-squared test with Bonferroni correction was used to calculate significance of deviation from the fit. Gene ontology (GO) enrichment of HVGs was conducted using DAVID Bioinformatics Resource. To avoid any biases from the large number of BTCs and BDCs contained in the dataset, batch control samples were removed from analysis for global sample clustering, variance partitioning, and differential gene expression testing.

Repeat sample analysis—To assess repeatability of individual patient differentiations, 26 subjects were differentiated a second time to day 32 motor neurons starting from frozen iPSCs and reprofiled. Samples were chosen from multiple previous batches and were evenly split between ALS vs. control and male vs. female groups. Raw RNA-seq counts from the first and second rounds of differentiation were normalized with *DESeq2* and Euclidean distance between samples was calculated using the *dist* function from the stats package in R. Multidimensional scaling of the distance matrix was performed using the *cmdscale* function from the stats package in R and visualized with the *ggplot* package. Spearman's rank coefficient of correlation was calculated using the *correlate* function and Simple Error Ratio Estimates[11] were used to confirm repeatability of biological replicates and were calculated in R.

Gene expression variance partitioning—Fraction of gene expression variation explained by each covariate was estimated by fitting a linear mixed model for each gene following the *variancePartition* package framework in R. Raw counts were first filtered of low expressors (<1 average count/sample) and normalized with *DESeq2* variance stabilizing transformation. For model formula design, categorical variables (batch, sex, disease status, iPSC cell of origin, and C9orf72 carrier status) were modeled as random effect. Continuous variables (sequencing depth, staining data, and iPSC patient age) were modeled as fixed effect. iPSC patient age explained the least amount of gene variation of any of the covariates and was not included in the final variance partitioning model formula design.

Correlation of principal component clustering to covariates—Principal components were generated in R and eigenvectors were correlated against covariates through the *eigencorplot* function from the *PCAtools* package in R. The test statistic is computed with *PCAtools* and is based on the Pearson's product moment correlation coefficient and follows a t-distribution with Bonferroni adjustment for multiple testing correction.

Human post-mortem RNA-seq analysis—For human post-mortem brain, the raw RNA-seq count matrix and meta data were downloaded from GEO (GSE67196). PCA was generated using the PlotPCA function in the *DESeq2* R package. For human thoracic spinal cord samples, raw reads were obtained from NYGC and were QCed and pseudo aligned to GRCh38 (hg38) human reference transcriptome using GENCODE annotation. Transcript level expression was quantified using Kallisto and summarized to gene level using R package *tximport*. PCA was generated using *PlotPCA* function and visualized with *ggplot2*.

X chromosome inactivation analysis—X chromosome inactivation status (XCI) was assessed in female samples essentially as described.[17] Briefly, X-linked gene expression values were normalized and the male median of each gene was determined. Non-expressed and highly variable genes were excluded and XX:XY expression ratios were calculated for each X-linked gene by dividing expression in individual female samples by the male median. These values were plotted along the X chromosome coordinates as a moving average with a window of 50 genes. The median value of all X-linked genes was used as the average XX:XY expression ratio for the female samples.

Quantification of cryptic exon inclusion—Alignment was performed and BAM files were generated with STAR aligner and indexed using samtools. Regtool was used to generate junction files and intron clustering was done using LeafCutter. Novel intron inclusion (cryptic exon) of *STMN2* and UNC13A was determined from the intron count file. Percent spliced in (PSI, y) values were reported by LeafCutter.

Classifier model for ALS progression rate—To identify genes predictive of ALS patient clinical data, ALS samples were randomly assigned to training and validation groups using the *sample* function in R using an 80/20 split, respectively. Variance partitioning was performed as described previously on the training group samples only, with site of onset modeled as random effect and age at symptom onset and ALSFRS-R disease progression slope modeled as fixed effects. The top 7 genes which ALSFRS-R slope explains the largest fraction of variation in expression were used to cluster samples by PCA. The median ALSFRS-R progression slope of −0.54 was used to separate patients into fast and slow progressors and ROC curves were used to determine the ability to separate fast and slow progressing ALS patients along the first principal component. The ROC model was then applied to the independent validation set which was withheld from the original analysis, to determine the accuracy of the model to classify patients into fast and slow progressing.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

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, and Bari Lipp Foundation. This work was also made possible, in part, through access to the following: the Genomics Research and Technology Hub (formerly Genomics High-Throughput Facility) Shared Resource of the Cancer Center Support Grant (P30CA-062203), the Single Cell Analysis Core shared resource of Complexity, Cooperation and Community in Cancer (U54CA217378), the Genomics-Bioinformatics Core of the Skin Biology Resource Based Center @ UCI (P30AR075047) at the University of California, Irvine and NIH shared instrumentation grants 1S10RR025496–01, 1S10OD010794–01, and 1S10OD021718–01. We gratefully acknowledge Dr. Soshana Svendsen for editing of the manuscript. Lastly, we thank all the study participants for their contribution and support of the Answer ALS program.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

REFERENCES

1. Giacomelli E, Vahsen BF, Calder EL, Xu Y, Scaber J, Gray E, Dafinca R, Talbot K, and Studer L (2022). Human stem cell models of neurodegeneration: from basic science of amyotrophic lateral sclerosis to clinical translation. Cell Stem Cell 29, 11–35. 10.1016/j.stem.2021.12.008. [PubMed: 34995492]
2. Shi Y, Lin S, Staats KA, Li Y, Chang WH, Hung ST, Hendricks E, Linares GR, Wang Y, Son EY, et al. (2018). Haploinsufficiency leads to neurodegeneration in C9ORF72 ALS/FTD human induced motor neurons. Nat. Med. 24, 313–325. 10.1038/nm.4490. [PubMed: 29400714]
3. Fujimori K, Ishikawa M, Otomo A, Atsuta N, Nakamura R, Akiyama T, Hadano S, Aoki M, Saya H, Sobue G, and Okano H (2018). Modeling sporadic ALS in iPSC-derived motor neurons identifies a potential therapeutic agent. Nat. Med. 24, 1579–1589. 10.1038/s41591-018-0140-5. [PubMed: 30127392]
4. Chang EA, Tomov ML, Suhr ST, Luo J, Olmsted ZT, Paluh JL, and Cibelli J (2015). Derivation of ethnically diverse human induced pluripotent stem cell lines. Sci. Rep. 5, 15234. 10.1038/srep15234. [PubMed: 26482195]
5. Baxi EG, Thompson T, Li J, Kaye JA, Lim RG, Wu J, Ramamoorthy D, Lima L, Vaibhav V, Matlock A, et al. (2022). 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. 10.1038/s41593-021-01006-0. [PubMed: 35115730]
6. Sances S, Bruijn LI, Chandran S, Eggan K, Ho R, Klim JR, Livesey MR, Lowry E, Macklis JD, Rushton D, et al. (2016). Modeling ALS with motor neurons derived from human induced pluripotent stem cells. Nat. Neurosci. 19, 542–553. 10.1038/nn.4273. [PubMed: 27021939]
7. Toombs J, Panther L, Ornelas L, Liu C, Gomez E, Martín-Ibáñez R, Cox SR, Ritchie SJ, Harris SE, Taylor A, et al. (2020). Generation of twenty four induced pluripotent stem cell lines from twenty four members of the Lothian Birth Cohort 1936. Stem Cell Res. 46, 101851. 10.1016/j.scr.2020.101851. [PubMed: 32450543]
8. Michetti F, D’Ambrosi N, Toesca A, Puglisi MA, Serrano A, Marchese E, Corvino V, and Geloso MC (2019). The S100B story: from biomarker to active factor in neural injury. J. Neurochem. 148, 168–187. 10.1111/jnc.14574. [PubMed: 30144068]
9. Panther L, Ornelas L, Jones MR, Gross AR, Gomez E, Liu C, Berman B, Svendsen CN, and Sareen D (2021). Generation of iPSC lines with high cytogenetic stability from peripheral blood mononuclear cells (PBMCs). 10.1101/2021.09.27.462082.
10. ENCODE Consortium (2017). ENCODE experimental guidelines for ENCODE3 RNA-seq. https://www.encodeproject.org/about/experiment-guidelines/.
11. Schulze SK, Kanwar R, Gölzenleuchter M, Therneau TM, and Beutler AS (2012). SERE: single-parameter quality control and sample comparison for RNA-Seq. BMC Genomics 13, 524. 10.1186/1471-2164-13-524. [PubMed: 23033915]
12. Hoffman GE, and Schadt EE (2016). variancePartition: interpreting drivers of variation in complex gene expression studies. BMC Bioinformatics 17, 483. 10.1186/s12859-016-1323-z. [PubMed: 27884101]
13. DeJesus-Hernandez M, Mackenzie IR, Boeve BF, Boxer AL, Baker M, Rutherford NJ, Nicholson AM, Finch NA, Flynn H, Adamson J, et al. (2011). Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. Neuron 72, 245–256. 10.1016/j.neuron.2011.09.011. [PubMed: 21944778]
14. Donnelly CJ, Zhang PW, Pham JT, Haeusler AR, Mistry NA, Vidensky S, Daley EL, Poth EM, Hoover B, Fines DM, et al. (2013). RNA toxicity from the ALS/FTD C9ORF72 expansion is mitigated by antisense intervention. Neuron 80, 415–428. 10.1016/jneuron.2013.10.015. [PubMed: 24139042]

15. Waite AJ, Bäumer D, East S, Neal J, Morris HR, Ansorge O, and Blake DJ (2014). Reduced C9orf72 protein levels in frontal cortex of amyotrophic lateral sclerosis and frontotemporal degeneration brain with the C9ORF72 hexanucleotide repeat expansion. Neurobiol. Aging 35, 1779.e5–1779.e13. 10.1016/j.neurobiolaging.2014.01.016.

16. Prudencio M, Belzil VV, Batra R, Ross CA, Gendron TF, Pregent LJ, Murray ME, Overstreet KK, Piazza-Johnston AE, Desaro P, et al. (2015). Distinct brain transcriptome profiles in C9orf72-associated and sporadic ALS. Nat. Neurosci. 18, 1175–1182. 10.1038/nn.4065. [PubMed: 26192745]

17. Bar S, Seaton LR, Weissbein U, Eldar-Geva T, and Benvenisty N (2019). Global characterization of X chromosome inactivation in human pluripotent stem cells. Cell Rep. 27, 20–29.e3. 10.1016/j.celrep.2019.03.019. [PubMed: 30943402]

18. Koopman P, Gubbay J, Vivian N, Goodfellow P, and Lovell-Badge R (1991). Male development of chromosomally female mice transgenic for Sry. Nature 351, 117–121. 10.1038/351117a0. [PubMed: 2030730]

19. Melamed Z, López-Erauskin J, Baughn MW, Zhang O, Drenner K, Sun Y, Freyermuth F, McMahon MA, Beccari MS, Artates JW, et al. (2019). Premature polyadenylation-mediated loss of stathmin-2 is a hallmark of TDP-43-dependent neurodegeneration. Nat. Neurosci. 22, 180–190. 10.1038/s41593-018-0293-z. [PubMed: 30643298]

20. Klim JR, Williams LA, Limone F, Guerra San Juan I, Davis-Dusenbery BN, Mordes DA, Burberry A, Steinbaugh MJ, Gamage KK, Kirchner R, et al. (2019). ALS-implicated protein TDP-43 sustains levels of STMN2, a mediator of motor neuron growth and repair. Nat. Neurosci. 22, 167–179. 10.1038/s41593-018-0300-4. [PubMed: 30643292]

21. Ma XR, Prudencio M, Koike Y, Vatsavayai SC, Kim G, Harbinski F, Briner A, Rodriguez CM, Guo C, Akiyama T, et al. (2022). TDP-43 represses cryptic exon inclusion in the FTD-ALS gene UNC13A. Nature 603, 124–130. 10.1038/s41586-022-04424-7. [PubMed: 35197626]

22. Brown AL, Wilkins OG, Keuss MJ, Hill SE, Zanovello M, Lee WC, Bampton A, Lee FCY, Masino L, Qi YA, et al. (2022). TDP-43 loss and ALS-risk SNPs drive mis-splicing and depletion of UNC13A. Nature 603, 131–137. 10.1038/s41586-022-04436-3. [PubMed: 35197628]

23. Saris CG, Horvath S, van Vught PW, van Es MA, Blauw HM, Fuller TF, Langfelder P, DeYoung J, Wokke JH, Veldink JH, et al. (2009). Weighted gene co-expression network analysis of the peripheral blood from amyotrophic lateral sclerosis patients. BMC Genomics 10, 405. 10.1186/1471-2164-10-405. [PubMed: 19712483]

24. Wang XS, Simmons Z, Liu W, Boyer PJ, and Connor JR (2006). Differential expression of genes in amyotrophic lateral sclerosis revealed by profiling the post mortem cortex. Amyotroph. Lateral Scler. 7, 201–210. 10.1080/17482960600947689. [PubMed: 17127558]

25. Roczniak-Ferguson A, and Ferguson SM (2019). Pleiotropic requirements for human TDP-43 in the regulation of cell and organelle homeostasis. Life Sci. Alliance 2, 2. 10.26508/lsa.201900358.

26. Watanabe H, Atsuta N, Hirakawa A, Nakamura R, Nakatochi M, Ishigaki S, Iida A, Ikegawa S, Kubo M, Yokoi D, et al. (2016). A rapid functional decline type of amyotrophic lateral sclerosis is linked to low expression of TTN. J. Neurol. Neurosurg. Psychiatry 87, 851–858. 10.1136/jnnp-2015-311541. [PubMed: 26746183]

27. Nakamura R, Misawa K, Tohnai G, Nakatochi M, Furuhashi S, Atsuta N, Hayashi N, Yokoi D, Watanabe H, Watanabe H, et al. (2020). A multi-ethnic meta-analysis identifies novel genes, including ACSL5, associated with amyotrophic lateral sclerosis. Commun. Biol. 3, 526. 10.1038/s42003-020-01251-2. [PubMed: 32968195]

28. Iida A, Takahashi A, Kubo M, Saito S, Hosono N, Ohnishi Y, Kiyotani K, Mushiroda T, Nakajima M, Ozaki K, et al. (2011). A functional variant in ZNF512B is associated with susceptibility to amyotrophic lateral sclerosis in Japanese. Hum. Mol. Genet. 20, 3684–3692. 10.1093/hmg/ddr268. [PubMed: 21665992]

29. Mack AA, Kroboth S, Rajesh D, and Wang WB (2011). Generation of induced pluripotent stem cells from CD34+ cells across blood drawn from multiple donors with non-integrating episomal vectors. PLoS One 6, e27956. 10.1371/journal.pone.0027956. [PubMed: 22132178]

30. Kyttälä A, Moraghebi R, Valensisi C, Kettunen J, Andrus C, Pasumarthy KK, Nakanishi M, Nishimura K, Ohtaka M, Weltner J, et al. (2016). Genetic variability overrides the impact of

parental cell type and determines iPSC differentiation potential. Stem Cell Rep. 6, 200–212. 10.1016/j.stemcr.2015.12.009.

31. Smith LR, Cho S, and Discher DE (2018). Stem cell differentiation is regulated by extracellular matrix mechanics. Physiology (Bethesda) 33, 16–25. 10.1152/physiol.00026.2017. [PubMed: 29212889]

32. McCombe PA, and Henderson RD (2010). Effects of gender in amyotrophic lateral sclerosis. Gend. Med. 7, 557–570. 10.1016/j.genm.2010.11.010. [PubMed: 21195356]

33. Manjaly ZR, Scott KM, Abhinav K, Wijesekera L, Ganesalingam J, Goldstein LH, Janssen A, Dougherty A, Willey E, Stanton BR, et al. (2010). The sex ratio in amyotrophic lateral sclerosis: A population based study. Amyotroph. Lateral Scler. 11, 439–442. 10.3109/17482961003610853. [PubMed: 20225930]

34. Suzuki M, Tork C, Shelley B, McHugh J, Wallace K, Klein SM, Lindstrom MJ, and Svendsen CN (2007). Sexual dimorphism in disease onset and progression of a rat model of ALS. Amyotroph. Lateral Scler. 8, 20–25. 10.1080/17482960600982447. [PubMed: 17364431]

35. Cacabelos D, Ramírez-Núñez O, Granado-Serrano AB, Torres P, Ayala V, Moiseeva V, Povedano M, Ferrer I, Pamplona R, Portero-Otin M, and Boada J (2016). Early and gender-specific differences in spinal cord mitochondrial function and oxidative stress markers in a mouse model of ALS. Acta Neuropathol. Commun. 4, 3. 10.1186/s40478-015-0271-6. [PubMed: 26757991]

36. Chió A, Mora G, Moglia C, Manera U, Canosa A, Cammarosano S, Ilardi A, Bertuzzo D, Bersano E, Cugnasco P, et al. (2017). Secular trends of amyotrophic lateral sclerosis: the Piemonte and Valle d'aosta register. JAMA Neurol. 74, 1097–1104. 10.1001/jama-neurol.2017.1387. [PubMed: 28692730]

37. Chió A, Moglia C, Canosa A, Manera U, D'Ovidio F, Vasta R, Grassano M, Brunetti M, Barberis M, Corrado L, et al. (2020). ALS phenotype is influenced by age, sex, and genetics: A population-based study. Neurology 94, e802–e810. 10.1212/WNL.0000000000008869. [PubMed: 31907290]

38. Janiszewski A, Talon I, Chappell J, Collombet S, Song J, De Geest N, To SK, Bervoets G, Marin-Bejar O, Provenzano C, et al. (2019). Dynamic reversal of random X-chromosome inactivation during iPSC reprogramming. Genome Res. 29, 1659–1672. 10.1101/gr.249706.119. [PubMed: 31515287]

39. Brenes AJ, Yoshikawa H, Bensaddek D, Mirauta B, Seaton D, Hukelmann JL, Jiang H, Stegle O, and Lamond AI (2021). Erosion of human X chromosome inactivation causes major remodeling of the iPSC proteome. Cell Rep. 35, 109032. 10.1016/j.celrep.2021.109032. [PubMed: 33910018]

40. Tchieu J, Kuoy E, Chin MH, Trinh H, Patterson M, Sherman SP, Aimiuwu O, Lindgren A, Hakimian S, Zack JA, et al. (2010). Female human iPSCs retain an inactive X chromosome. Cell Stem Cell 7, 329–342. 10.1016/j.stem.2010.06.024. [PubMed: 20727844]

41. Vawter MP, Evans S, Choudary P, Tomita H, Meador-Woodruff J, Molnar M, Li J, Lopez JF, Myers R, Cox D, et al. (2004). Gender-specific gene expression in post-mortem human brain: localization to sex chromosomes. Neuropsychopharmacology 29, 373–384. 10.1038/sj.npp.1300337. [PubMed: 14583743]

42. Tukiainen T, Villani AC, Yen A, Rivas MA, Marshall JL, Satija R, Aguirre M, Gauthier L, Fleharty M, Kirby A, et al. (2017). Landscape of X chromosome inactivation across human tissues. Nature 550, 244–248. 10.1038/nature24265. [PubMed: 29022598]

43. Okano H, and Morimoto S (2022). iPSC-based disease modeling and drug discovery in cardinal neurodegenerative disorders. Cell Stem Cell 29, 189–208. 10.1016/j.stem.2022.01.007. [PubMed: 35120619]

44. Goyal NA, Berry JD, Windebank A, Staff NP, Maragakis NJ, van den Berg LH, Genge A, Miller R, Baloh RH, Kern R, et al. (2020). Addressing heterogeneity in amyotrophic lateral sclerosis CLINICAL TRIALS. Muscle Nerve 62, 156–166. 10.1002/mus.26801. [PubMed: 31899540]

45. Ho R, Workman MJ, Mathkar P, Wu K, Kim KJ, O'Rourke JG, Kellogg M, Montel V, Banuelos MG, Arogundade OA, et al. (2021). Cross-comparison of human iPSC motor neuron models of familial and sporadic ALS reveals early and convergent transcriptomic disease signatures. Cell Syst. 12, 159–175.e9. 10.1016/j.cels.2020.10.010. [PubMed: 33382996]

46. Yu J, Hu K, Smuga-Otto K, Tian S, Stewart R, Slukvin II, and Thomson JA (2009). Human induced pluripotent stem cells free of vector and transgene sequences. Science 324, 797–801. 10.1126/science.1172482. [PubMed: 19325077]

47. Okita K, Matsumura Y, Sato Y, Okada A, Morizane A, Okamoto S, Hong H, Nakagawa M, Tanabe K, Tezuka K, et al. (2011). A more efficient method to generate integration-free human iPS cells. Nat. Methods 8, 409–412. 10.1038/nmeth.1591. [PubMed: 21460823]

48. Okita K, Yamakawa T, Matsumura Y, Sato Y, Amano N, Watanabe A, Goshima N, and Yamanaka S (2013). An efficient nonviral method to generate integration-free human-induced pluripotent stem cells from cord blood and peripheral blood cells. Stem Cells 31, 458–466. 10.1002/stem.1293. [PubMed: 23193063]

Highlights

- Large-scale generation and differentiation of control and ALS iPSCs to motor neurons
- Patient sex and cell culture composition are significant sources of variation
- Male ALS cultures have more motor neurons and enrichment of stress-related pathways
- Deposition of data and cell lines into public repository with additional multi-omics

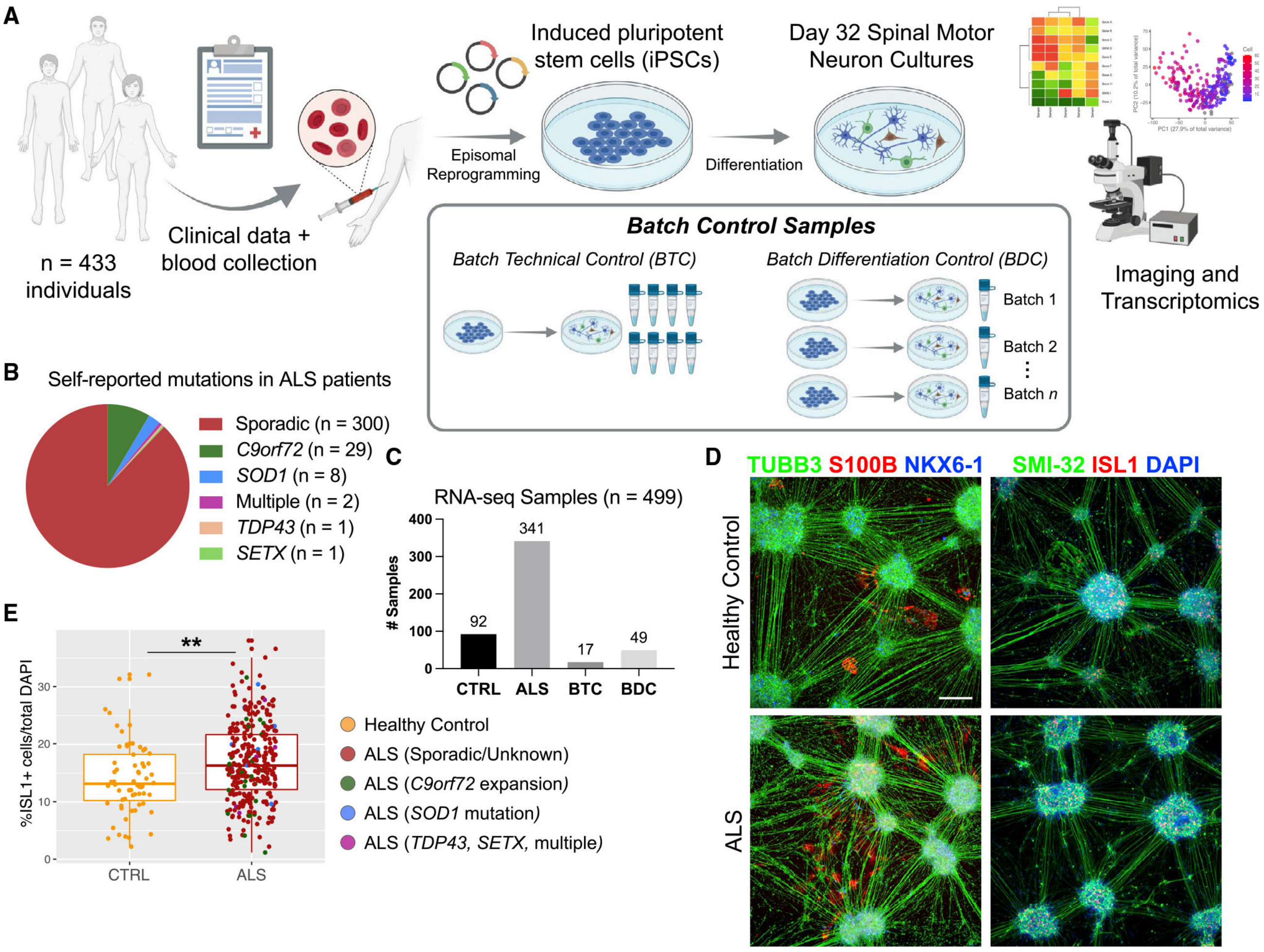


Figure 1. Differentiation and characterization of human iPSC-derived motor neurons

(A) Schematic overview of sample collection, reprogramming, and differentiation of motor neurons from 433 human subjects.

(B) Summary of self-reported genetic mutations in ALS patient cohort.

(C) Breakdown of bulk RNA-seq samples used for analysis. Batch technical controls (BTCs) and batch differentiation controls (BDCs) were used to assess technical noise and variation in the differentiation protocol.

(D) Representative immunofluorescent images of day 32 iPSC-derived spinal motor neuron cultures from control and ALS patients. Scale bar, 200 μm.

(E) Quantification of percent ISL1+ cells of total DAPI-stained cells (**p < 0.01, unpaired two-tailed t test. Box represents interquartile range (IQR), line indicates median, and whiskers denote +/− 1.5*IQR).

See also Figure S1.

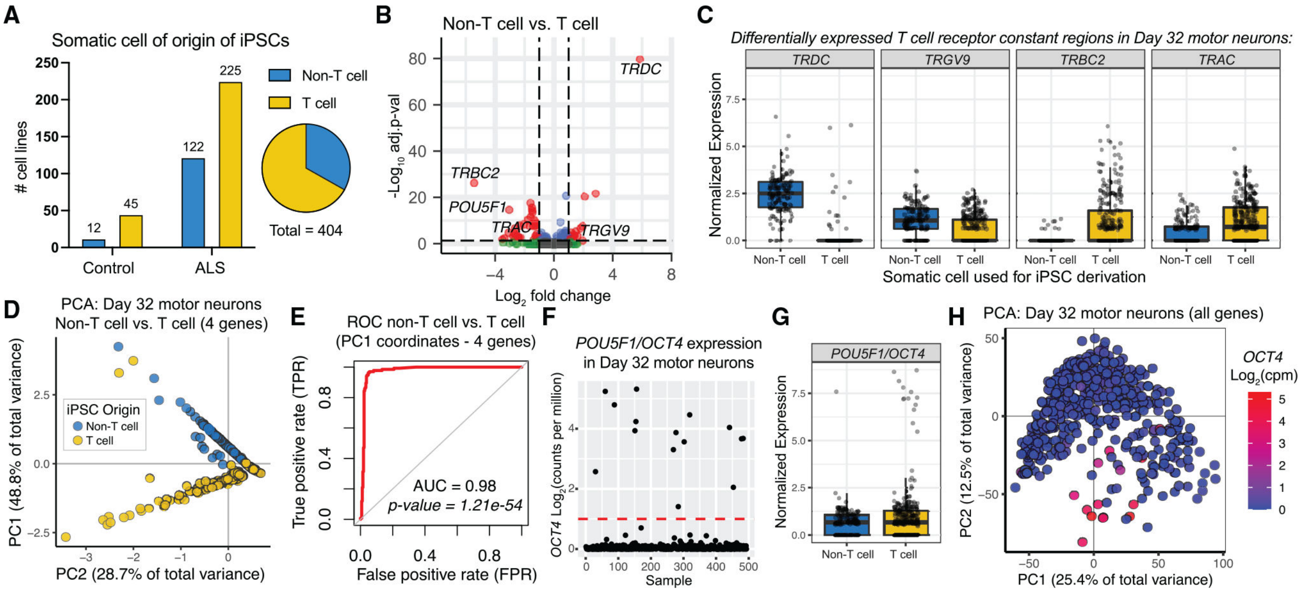


Figure 2. Somatic cell used for iPSC reprogramming detectable in day 32 motor neurons

(A) Breakdown of known somatic cells used for iPSC reprogramming.

(B) Volcano plot of differentially expressed genes in day 32 motor neurons differentiated from non-T cell-derived iPSCs versus day 32 motor neurons differentiated from T cell-derived iPSCs (*p < 0.05, DESeq2 Wald test with Bonferonni correction).

(C) Normalized expression of the top differentially expressed genes (Box represents interquartile range (IQR), line indicates median, and whiskers denote +/− 1.5*IQR).

(D) PCA of day 32 motor neurons using the 4 genes shown in Figure 2C and colored by the type of somatic cell from which the iPSCs were derived.

(E) Classification of sample origin using a receiver operating characteristic (ROC) curve based on the PC1 coordinates in Figure 2D (AUC = 0.98, p = 1.21e54, Mann-Whitney U test).

(F) Several day 32 motor neuron samples displayed aberrant expression of the pluripotency factor *OCT4/POU5F1* at more than 100 times the interquartile range of the dataset (red dashed line).

(G) The majority of *OCT4* outliers were from samples originally derived from T cells.

(H) High *OCT4*-expressing samples are mainly clustered as outliers in global principal component analysis.

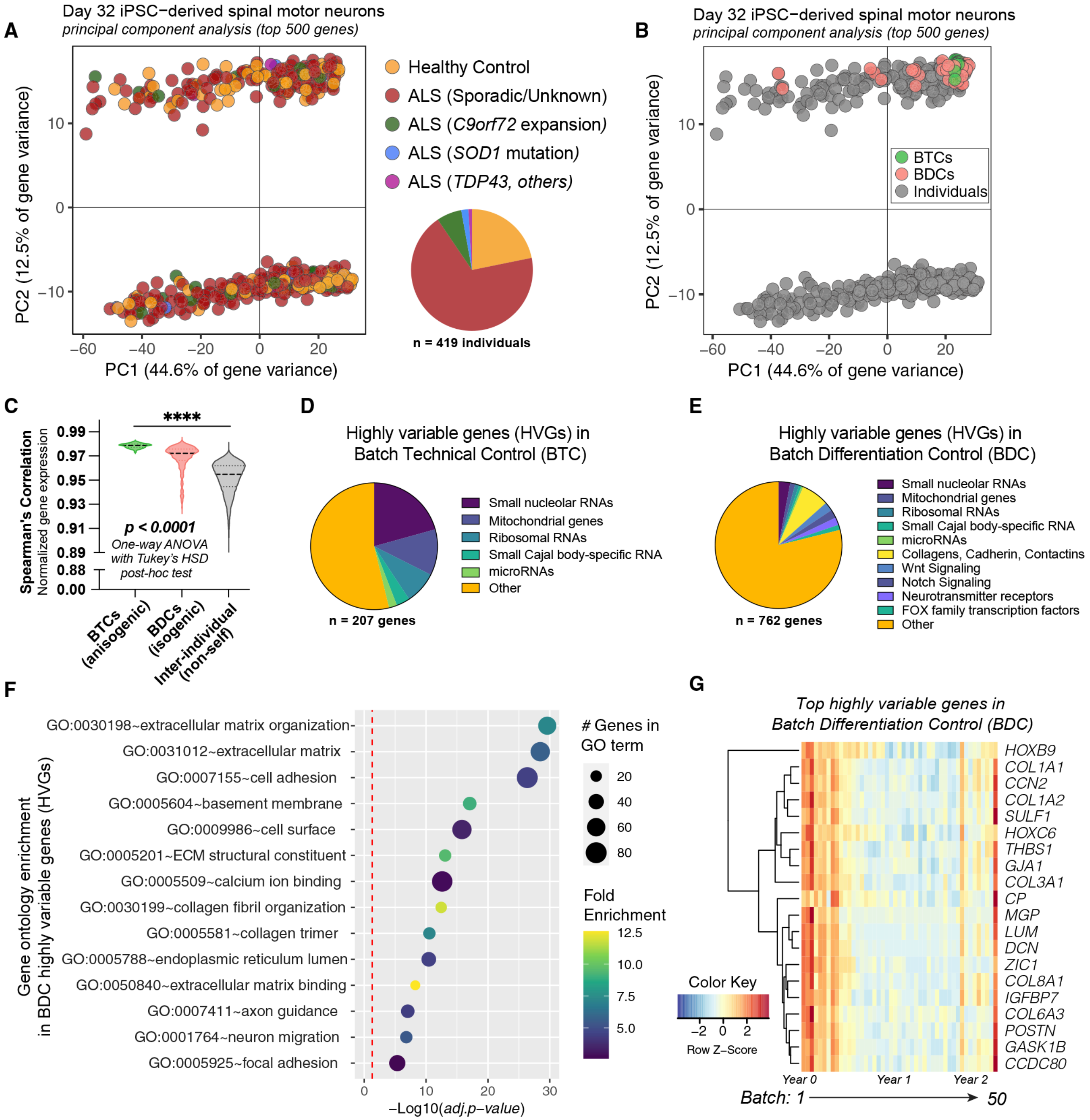


Figure 3. Analysis of batch control samples identifies HVGs associated with technical noise and protocol variation

(A) Clustering of samples by principal component analysis (PCA) with data normalization and removal of BTC, BDC, and outlier samples.

(B) Projection of BTC and BDC samples along PC1 and PC2 in the global clustering of samples shown in Figure 3A.

(C) Spearman correlation between samples (****p < 0.0001, one-way ANOVA with Tukey's HSD post hoc test).

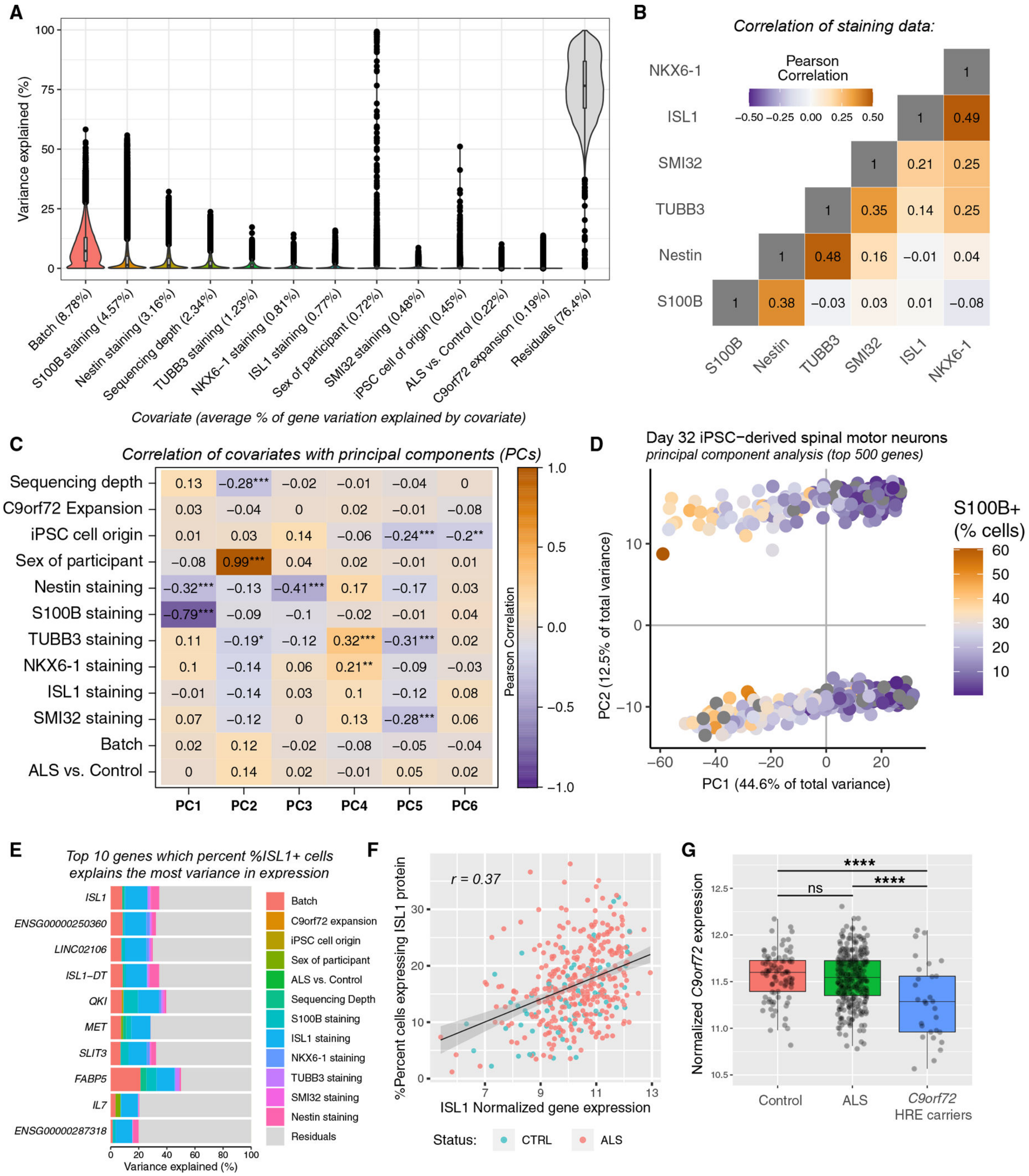
(D) Grouping of HVGs identified in BTCs.

(E) In addition to small RNA molecules, HVGs in BDC samples include genes associated with the extracellular matrix, signaling, neurotransmitter receptor expression, and various other genes.

(F) Gene ontology (GO) enrichment of HVGs in BDC samples.

(G) Heatmap of top 20 HVGs in BDC samples ordered by batch.

See also Figure S2.



	S100B	Nestin	TUBB3	SMI32	ISL1	NKX6-1
NKX6-1						1
ISL1					1	0.49
SMI32				1	0.21	0.25
TUBB3			1	0.35	0.14	0.25
Nestin		1	0.48	0.16	−0.01	0.04
S100B	1	0.38	−0.03	0.03	0.01	−0.08

	PC1	PC2	PC3	PC4	PC5	PC6
Sequencing depth	0.13	−0.28***	−0.02	−0.01	−0.04	0
C9orf72 Expansion	0.03	−0.04	0	0.02	−0.01	−0.08
iPSC cell origin	0.01	0.03	0.14	−0.06	−0.24***	−0.2**
Sex of participant	−0.08	0.99***	0.04	0.02	−0.01	0.01
Nestin staining	−0.32***	−0.13	−0.41***	0.17	−0.17	0.03
S100B staining	−0.79***	−0.09	−0.1	−0.02	−0.01	0.04
TUBB3 staining	0.11	−0.19*	−0.12	0.32***	−0.31***	0.02
NKX6-1 staining	0.1	−0.14	0.06	0.21**	−0.09	−0.03
ISL1 staining	−0.01	−0.14	0.03	0.1	−0.12	0.08
SMI32 staining	0.07	−0.12	0	0.13	−0.28***	0.06
Batch	0.02	0.12	−0.02	−0.08	−0.05	−0.04
ALS vs. Control	0	0.14	0.02	−0.01	0.05	0.02

Figure 4. Correlating RNA-seq data with covariates reveals S100B and sex as leading descriptors of variance in the transcriptomic data

(A) Results of a linear mixed model used to estimate the proportion of variation in each gene attributable to technical and biological variables included in the metadata.

(B) Pearson correlation analysis of the staining data.

(C) Correlation of principal components to sample covariates (*p < 0.05, **p < 0.01, ***p < 0.001, two-tailed t test with Bonferroni correction).

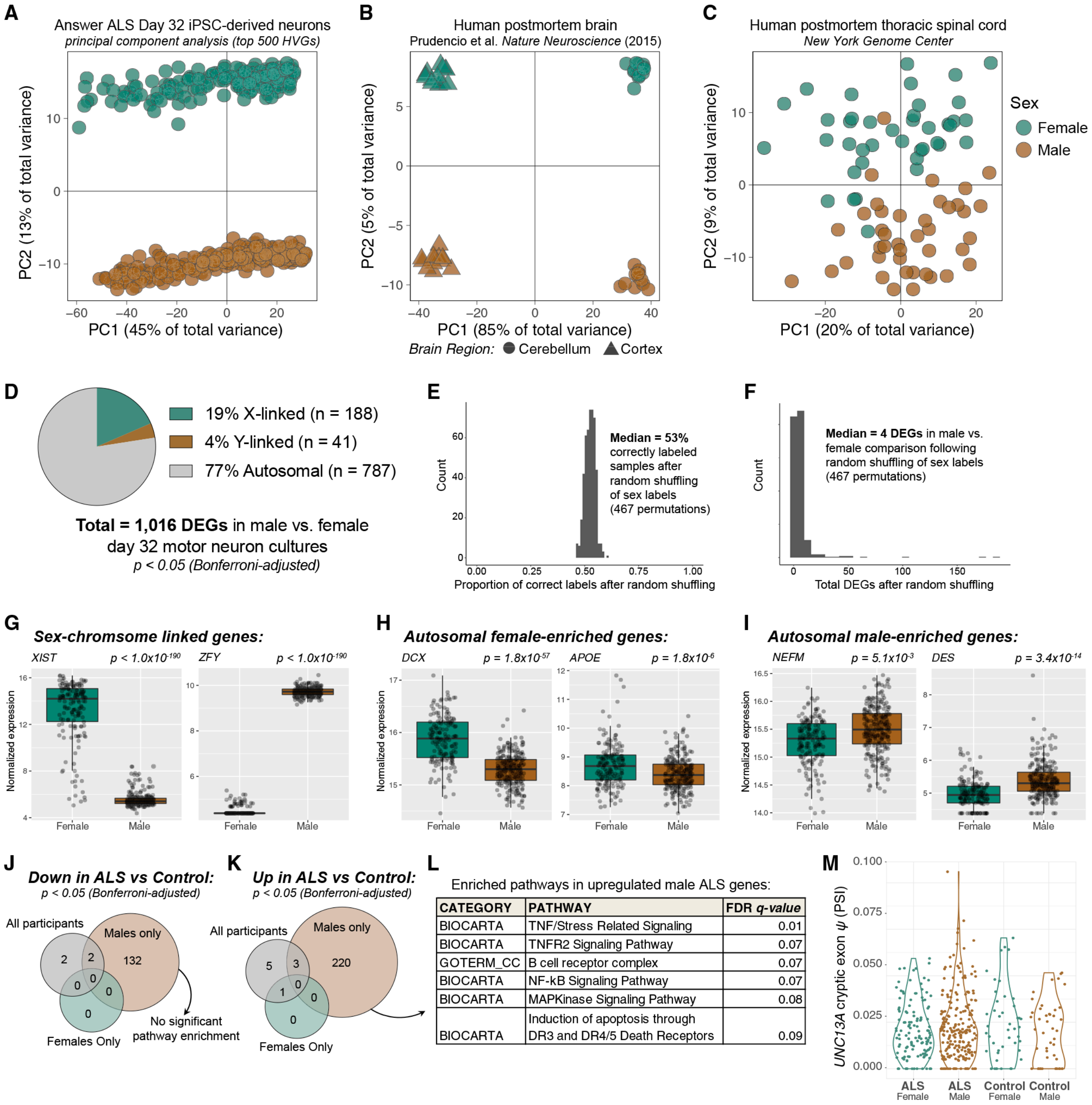
(D) Coloring samples in PCA by the percentage of cells expressing S100B.

(E) Top 10 genes in which percent ISL1+ cells explain the most amount of gene variation.

(F) Correlation of *ISL1* gene expression to ISL1 staining data.

(G) *C9orf72* gene expression in healthy control, *C9orf72* hexanucleotide repeat expansion carriers, and all other ALS subjects (****p < 0.0001, one-way ANOVA with Tukey's HSD post hoc test. Box represents interquartile range (IQR), line indicates median, and whiskers denote +/− 1.5*IQR).

See also Figure S3.



Enriched pathways in upregulated male ALS genes:

CATEGORY	PATHWAY	FDR *q-value*
BIOCARTA	TNF/Stress Related Signaling	0.01
BIOCARTA	TNFR2 Signaling Pathway	0.07
GOTERM_CC	B cell receptor complex	0.07
BIOCARTA	NF-kB Signaling Pathway	0.07
BIOCARTA	MAPKinase Signaling Pathway	0.08
BIOCARTA	Induction of apoptosis through DR3 and DR4/5 Death Receptors	0.09

Figure 5. *In vitro* and *in vivo* sex differences in neuronal cultures and tissues

(A) PCA using the top 500 most variable genes separates male and female day 32 motor neuron samples along PC2.

(B and C) Male-female differences in human post-mortem (B) brain and (C) spinal cord.

(D) Differential gene expression analysis reveals 1,016 genes differentially expressed between male and female samples (p < 0.05, DESeq2 Wald test with Bonferroni correction).

(E) Histogram showing the proportion of correctly labeled samples following random shuffling of sex labels in 467 permutations.

(F) Histogram of the number of DEGs from 467 reshuffles of the sex class label.

(G) X- and Y-linked genes show striking male-female differences in expression. (Box represents interquartile range (IQR), line indicates median, and whiskers denote +/− 1.5*IQR.)

(H and I) Many autosomal genes show enrichment specifically in (H) female or in (I) male samples (DESeq2 Wald test with Bonferonni correction).

(J and K) Analysis of differentially expressed genes (DEGs) between ALS and control reveals subsets of genes that are (J) downregulated or (K) upregulated specifically in male ALS versus male control.

(L) Pathway enrichment of the upregulated DEGs in male ALS samples uncovers pathways related to TNF and NF-κB signaling (Fisher's exact test with Benjamini-Hochberg false discovery rate [FDR] correction, FDR < 0.1).

(M) *UNC13A* cryptic exon expression shown as percent spliced in (PSI, ψ) in day 32 motor neuron cultures.

See also Figures S4 and S5.

Table 1.

Summary of pathogenic, likely pathogenic (P-LP) and *in silico* predicted damaging (IS-D) coding variants

Variant annotation	Variant type	Healthy control (n = 52)	Sporadic ALS (SALS) (n = 253)	SALS versus control p value	Familial ALS (FALS) (n = 32)	FALS versus control p value
ClinVar/InterVar/ Harms pathogenic-likely pathogenic (P-LP) n = 1,194 total detected variants	P-LP variants (all genes)	avg. = 35.6 P-LP variants/person	avg. = 36.1 P-LP variants/person	0.45 (n.s.)	avg. = 37.4 P-LP variants/ person	0.13 (n.s.)
	P-LP variants (ALS genes)	15.4% harbor P-LP ALS variant (n = 8/52)	14.6% harbor P-LP ALS variant (n = 37/253)	0.89 (n.s.)	28.1% harbor P-LP ALS variant (n = 9/32)	0.15 (n.s.)
	P-LP (ALS genes) and/or C9orf72 repeat expansion	15.4% harbor P-LP ALS variant or C9orf72 (n = 8/52)	19.0% harbor P-LP ALS variant or C9orf72 (n = 48/253)	0.54 (n.s.)	68.8% harbor P-LP ALS variant or C9orf72 (n = 22/32)	****p < 0.0001
In silico predicted damaging (IS-D) n = 13,561 total detected variants	IS-D variants (all genes)	avg. = 111 IS-D variants/person	avg. = 113 IS-D variants/person	0.32 (n.s.)	avg. = 110 IS-D variants/person	0.74 (n.s.)
	IS-D variants (ALS genes)	9.6% harbor IS-D ALS variant (n = 5/52)	11.1% harbor IS-D ALS variant (n = 28/285)	0.76 (n.s.)	40.6% harbor IS-D ALS variant (n = 13/32)	***p < 0.001
	IS-D (ALS genes) and/or C9orf72 repeat expansion	9.6% harbor IS-D ALS variant or C9orf72 (n = 5/52)	15.4% harbor IS-D ALS variant or C9orf72 (n = 39/285)	0.28 (n.s.)	75.0% harbor IS-D ALS variant or C9orf72 (n = 24/32)	****p < 0.0001

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Mouse anti-SMI-32/Neurofilament H (1:1000)	Biolegend	Cat#801701; RRID:AB_2564642
Goat anti-Human Islet-1/ISL1 (1:250)	R&D Systems	Cat#AF1837; RRID:AB_2126324
Mouse anti-NKX6.1 (1:1000)	DSHB	Cat#F55A10-s; RRID:AB_532378
Goat anti-NKX6.1	R&D Systems	Cat#AF5857; RRID:AB_1857045
Rabbit anti-TUBB3/Tubulin Beta-III (1:1000)	Abnova	Cat#PAB7874; RRID:AB_1716633
Rabbit anti-Nestin/NES (1:1000)	Sigma-Aldrich	Cat#ABD69; RRID:AB_2744681
Mouse anti-S100B (1:250)	Sigma-Aldrich	Cat#S2532; RRID:AB_477499
Donkey anti-Rabbit IgG Alexa Fluor™ 488 (1:1000)	ThermoFisher Scientific	Cat#A-21206; RRID:AB_2535792
Donkey anti-Mouse IgG Alexa Fluor™ 568 (1:1000)	ThermoFisher Scientific	Cat#A-10037; RRID:AB_2534013
Donkey anti-Rabbit IgG Alexa Fluor™ 647 (1:1000)	ThermoFisher Scientific	Cat#A-31573; RRID:AB_2536183
Donkey anti-Goat IgG Alexa Fluor™ 647 (1:1000)	ThermoFisher Scientific	Cat#A-21447; RRID:AB_2535864
DAPI (4’,6-Diamidino-2-Phenylindole, Dilactate) (0.1 ug/mL)	ThermoFisher Scientific	Cat#D3571, RRID:AB_2307445
Chemicals, peptides, and recombinant proteins		
X-VIVO 10 Serum-free Hematopoietic Cell Medium	Lonza	Cat#04–380Q
MEMα	ThermoFisher Scientific	Cat#12561056
Primate ES Cell Medium	Reprocell	Cat#RCHEMD001
mTeSR1 media	StemCell Technologies	Cat#85850
IMDM	ThermoFisher Scientific	Cat#12440061
Ham’s F-12 Nutrient Mix	ThermoFisher Scientific	Cat#11765062
MEM Non-Essential Amino Acids Solution (100X)	ThermoFisher Scientific	Cat#11140050
B-27™ Supplement (50X), serum free	ThermoFisher Scientific	Cat#17504044
N-2 Supplement (100X)	ThermoFisher Scientific	Cat#17502048
Pen-Strep-Antimycotic (PSA)	ThermoFisher Scientific	Cat#15240062
LDN193189	Cayman Chemicals	Cat#17502048
SB431542	Cayman Chemicals	Cat#13031
CHIR99021	Xcess Biosciences	Cat#M60002
All-Trans Retinoic Acid (ATRA)	Stemgent	CAT#04–0021
Smoothened agonist (SAG)	Cayman Chemicals	Cat#11914
Y-27632 (ROCKi)	StemCell Technologies	Cat#72308
Compound E	Sigma-Aldrich	Cat#565790
DAPT	Cayman Chemicals	Cat#13197
Dibutyryl-cAMP (dbcAMP)	Sigma-Aldrich	Cat#28745
L-Ascorbic acid	Sigma-Aldrich	Cat#A4403
BDNF	PeproTech	Cat#450–02
GDNF	PeproTech	Cat#450–10

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Human IL-2 Recombinant Protein	ThermoFisher Scientific	Cat#PHC0026
Human Recombinant IL-3	StemCell Technologies	Cat#78040.1
Human Recombinant IL-6	StemCell Technologies	Cat#78050.1
Human Recombinant G-CSF	StemCell Technologies	Cat#78012.1
Human Recombinant GM-CSF	StemCell Technologies	Cat#78015.1
Dynabeads Human T-Activator CD3/CD28	ThermoFisher Scientific	Cat#11161D
Accutase	Sigma-Aldrich	Cat#SCR005
Fetal Bovine Serum (FBS), qualified	ThermoFisher Scientific	Cat#10437028
Matrigel® Growth Factor Reduced (GFR)	Corning	Cat#354230
Critical commercial assays		
TaqMan™ hPSC Scorecard™ Panel, 384-well	Thermo Fisher Scientific	Cat#A15870
IdentiClone® TCRB + TCRG T Cell Clonality Assay - Gel Detection	Invivoscribe	Cat#92000010
Amaxa Human T cell Nucleofector® Kit	Lonza	Cat#VVPA-1002
MycoAlert® Mycoplasma Detection Kit	Lonza	Cat#LT07–118
Deposited data		
RNA sequencing data, WGS data, and patient clinical information	Answer ALS Consortium	dataportal.answerals.org
Oligonucleotides		
Epstein-Barr virus nuclear antigen (EBNA), *forward primer*	Integrated DNA Technologies	*GGTCCCGAGAATCCCCATCC*
Epstein-Barr virus nuclear antigen (EBNA), *reverse primer*	Integrated DNA Technologies	*TTCATGGTCGCTGTCAGACAG*
GAPDH, forward primer	Integrated DNA Technologies	*GTGGACCTGACCTGCCGTCT*
GAPDH, *reverse primer*	Integrated DNA Technologies	*GGAGGAGTGGGTGTCGCTGT*
Recombinant DNA		
pEP4 E02S ET2K	Yu et al.[46]	RRID:Addgene_20927
pCXLE-hOCT3/4-shp53-F	Okita et al.[47]	RRID:Addgene_27077
pCXLE-hUL	Okita et al.[47]	RRID:Addgene_27080
pCXLE-hSK	Okita et al.[47]	RRID:Addgene_27078
pCXWB-EBNA1	Okita et al.[48]	RRID:Addgene_37624
Software and algorithms		
R Project for Statistical Computing	CRAN	RRID:SCR_001905
tidyverse	CRAN	RRID:SCR_019186
ROCR: Classifier Visualization in R	CRAN	RRID:SCR_008551
DESeq2	Bioconductor	RRID:SCR_015687
edgeR	Bioconductor	RRID:SCR_012802
biomaRt	Bioconductor	RRID:SCR_019214
variancePartition	Bioconductor	RRID:SCR_019204

REAGENT or RESOURCE	SOURCE	IDENTIFIER
karyoploteR	Bioconductor	RRID:SCR_021824
PCAtools	Bioconductor	N/A
Hisat2	Github	RRID:SCR_015530
STAR aligner	Github	RRID:SCR_004463
LeafCutter	Github	RRID:SCR_017639
Samtools	htslib.org	RRID:SCR_002105
DAVID Bioinformatics Resource	david.ncifcrf.gov	RRID:SCR_001881