Answer ALS, a large-scale resource for sporadic and familial ALS combining clinical and multi-omics data from induced pluripotent cell lines

Emily G. Baxi1,2, Terri Thompson3, Jonathan Li4, Julia A. Kaye5, Ryan G. Lim6, Jie Wu7, Divya Ramamoorthy4, Leandro Lima5, Vineet Vaibhav8, Andrea Matlock8, Aaron Frank9, Alyssa N. Coyne1,2, Barry Landin10, Loren Ornelas9, Elizabeth Mosmiller2, Sara Thrower11, S. Michelle Farr12, Lindsey Panther9, Emilda Gomez9, Erick Galvez9, Daniel Perez9, Imara Meepe9, Susan Lei9, Berhan Mandefro13, Hannah Trost13, Louis Pinedo9, Maria G. Banuelos13, Chunyan Liu9, Ruby Moran9, Veronica Garcia13, Michael Workman13, Richie Ho13, Stacia Wyman5, Jennifer Roggenbuck14, Matthew B. Harms15, Jennifer Stocksdale16, Ricardo Miramontes6, Keona Wang16, Vidya Venkatraman8, Ronald Holewinski8, Niveda Sundararaman8, Rakhi Pandey8, Danica-Mae Manalo8, Aneesh Donde4, Nhan Huynh9, Miriam Adam4, Brook T. Wassie4, Edward Vertudes5, Naufa Amirani5, Krishna Raja5, Reuben Thomas5, Lindsey Hayes2, Alex Lenail4, Aianna Cerezo2, Sarah Luppino11, Alanna Farrar11, Lindsay Pothier11, Carolyn Prina15, Todd Morgan17, Arish Jamii18, Sarah Heintzman15, Jennifer Jockel-Balsarotti19, Elizabeth Karanja19, Jesse Markway19, Molly McCallum19, Ben Joslin20, Deniz Alibazoglu20, Stephen Kolb15, Senda Ajroud-Driss20, Robert Baloh13, Daragh Heitzman17, Tim Miller19, Jonathan D. Glass18, Natasha Leanna Patel-Murray4, Hong Yu11, Ervin Sinani11, Prasha Vigneswaran11, Alexander V. Sherman11, Omar Ahmad2, Promit Roy2, Jay C. Beavers21, Steven Zeiler2, John W. Krakauer2, Carla Agurto10, Guillermo Cecchi10, Mary Bellard22, Yogindra Raghav4, Karen Sachs4, Tobias Ehrenberger4, Elizabeth Bruce22, Merit E. Cudkowicz11, Nicholas Maragakis12, Raquel Norel10, Jennifer E. Van Eyk8, Steven Finkbeiner5, James Berry11, Dhruv Sareen9,13, Leslie M. Thompson6,7,16,23, Ernest Fraenkel4, Clive N. Svendsen9,13 and Jeffrey D. Rothstein1,2

Answer ALS is a biological and clinical resource of patient-derived, induced pluripotent stem (iPS) cell lines, multi-omics data derived from iPS neurons and longitudinal clinical and smartphone data from over 1,000 patients with ALS. This resource provides population-level biological and clinical data that may be employed to identify clinical–molecular–biochemical subtypes of amyotrophic lateral sclerosis (ALS). A unique smartphone-based system was employed to collect deep clinical data, including fine motor activity, speech, breathing and linguistics/cognition. The iPS spinal neurons were blood derived from each patient and these cells underwent multi-omics analytics including whole-genome sequencing, RNA transcriptomics, ATAC-sequencing and proteomics. The intent of these data is for the generation of integrated clinical and biological signatures using bioinformatics, statistics and computational biology to establish patterns that may lead to a better understanding of the underlying mechanisms of disease, including subgroup identification. A web portal for open-source sharing of all data was developed for widespread community-based data analytics.

Over the last several decades, tremendous progress in the optimization of therapies for various medical conditions, such as cancer, has been realized. Many factors underlie this therapeutic success, including optimization of clinical trial design, new pathway-specific pharmaceuticals and the coordination of participant recruitment efforts across clinics. Perhaps one of the most powerful and fundamental reasons for the success of some cancer therapies is the ability to sample diseased tissues and thereby
distinguish the biological and molecular events responsible for individual diseases or disease subgroups within a disease cluster. Thus, skin, breast or prostate biopsies have been important starting points for the investigation of various types of melanomas and breast or prostate cancers. Neurodegenerative diseases such as ALS, Alzheimer’s disease and Huntington’s disease have, however, not been as extensively studied. Clinical trials in humans, often based on findings from nonhuman model systems, have repeatedly proven disappointing. Although there are probably many reasons for such failures (for example, poor pharmacokinetics, wrong biological pathway, lack of target engagement), a critical reason is the inability to identify disease pathways in patient tissues and to segment patients for clinical trials according to these pathways. As a result of the high risk of disability, brain and spinal cord biopsies for tissue analysis are not feasible in neurodegenerative diseases and therefore, unlike the biopsy of other organs and tissues, obtaining neural tissue during the disease course is a significant hurdle to effective therapeutic development.

An alternative is to use stem cell technology and infer disease pathways from cell lines derived from the patients’ own blood. Evidence for this approach is beginning to emerge. Early work employing iPS spinal neurons from patients with C9orf72 ALS/ frontotemporal dementia led the way to the development of the first antisense-based gene therapy for this common familial form of ALS (iALS), with an international clinical trial already under way (clinicaltrials.gov: NCT03626012). But for most patients with ALS, who have sporadic disease (sALS), these discoveries have yet to translate into meaningful therapies. A major barrier has been the lack of a predictive preclinical human model for sALS. However, with advances in iPS cell technology and the unprecedented data and specimen collection efforts of Answer ALS, we can now take an iPS cell-based approach to unraveling mechanisms that may cause or contribute to the heterogeneous clinical spectra of ALS, such as pattern and speed of spread and certain nonmotor manifestations. Notably, multiple gene mutations are already known to cause fALS and represent quite diverse pathways: RNA metabolism, nuclear transport, protein aggregation, axonal trafficking, glial dysfunction, etc. Curiously, the variability in clinical features is nearly as great when comparing patients with any single mutated gene as it is when comparing across genes or with sALS. Little is known about the derangements in specific biological pathway(s) driving sALS or whether there are ALS subgroups defined by specific biological derangements. Knowledge of these biological subgroups may be critically important and the success of disease-modifying therapies may depend on treating the right ‘subgroup’ with the proper pathway-targeting drug.

The Answer ALS (AALS) program was conceived as a program to generate iPS cell lines from a large number of patients with ALS and apply well-established molecular, biochemical and imaging techniques to understand the heterogeneity of sALS in these patient-derived spinal neurons, to serve as a ‘biopsy-like’ equivalent. After ensuring that results were reproducible, we assembled comprehensive biological datasets from individual subject iPS cell lines and combined them with the longitudinal clinical data. In contrast to smaller previous iPS cell experiments, studies of iPS cells from a large population, like AALS, provide the first opportunity to explore biologically relevant subgroups of sALS. This resource program was designed with the core goals of providing large clinical and biological datasets in an open source-like application that could and biological datasets in an open source-like application that offers researchers the proper tools to identify biological subgroups and an extensive collection of iPS cell lines with which to test ALS therapies and hypotheses about ALS pathogenesis.

Results
Clinical demographics and clinical data generation. Population demographics. The enrolled participant population for the AALS program (Fig. 1a, Extended Data Fig. 1, Supplementary Information and Supplementary Tables 1–5) had clinical characteristics comparable to past large sALS population demographics, with a slightly higher number of male than female participants, site of disease onset predominantly a limb rather than bulbar and a mean age of disease onset of approximately 57 years. The mean delay in clinical diagnosis for ALS patients included in the study was 14.8 months. A higher percentage of patients with rapid progression had bulbar-onset disease. There was a wide range of disease progression rates over the time period of observation (Fig. 1b,c), with an average follow-up duration of 12.5 months and an average rate of decline of 0.77 points per month (Fig. 1b,c). The smaller population of patients with fALS in the resource had typical representations of the common gene mutations including C9orf72 and SOD1 (Table 1), with a small subset of patients with C9orf72 and non-C9orf72 ALSs developing cognitive decline during the study (https://data-portal.AnswerALS.org). A small number of individuals were ALS mutation carriers (asymptomatic ALS) without overt neurological disease (Table 1). Non-ALS motor neuron disease (MND) included patients with predominantly upper MND, not formally categorized as ALS (for example, primary lateral sclerosis), and their demographic information is included in Supplementary Table 4. The healthy control subject population consisted of age-matched participants without ALS or a family history of ALS.

App-based voice recordings—motor and speech analyses. A core tool to gather more comprehensive longitudinal clinical data, ultimately to integrate with the biological datasets, was the development of a new smartphone app, designed to inform elements of motor activity, speech, breathing, voice and cognition (Supplementary Information) while patients were at home. Given the nature of this progressively disabling disorder, the reliability of utilization is an important variable. Compliance for using the smartphone app was analyzed over 18 months from the beginning of the app rollout to a subset of 80 study subjects. Surprisingly, only a modest decrease in compliance was observed with increased duration of use (Fig. 2a).

App data accurately predicted clinical progression. From speech recordings, we extracted linguistic features to evaluate word diversity and complexity of thought such as semantic similarity, dispersion and frequency, as recently detailed. Features derived from the voice tasks (single-breath count, read-aloud passage and free speech; Extended Data Fig. 2) each correlated highly with the bulbar subdomain of the ALS Functional Rating Scale-Revised (ALSFRS-R; Pearson's R = 0.8, slope = 1.14; Pearson's R = 0.89, slope = 0.98; and Pearson's R = 0.71, slope = 1.12, respectively). Features from the finger tracing showed modest individual correlations with the ALSFRS-R total score (Fig. 2b and Extended Data Fig. 2). Importantly, the combination of features from all of these tasks correlated very highly with the ALSFRS-R total score (Pearson's R = 0.89, slope = 1.16; Fig. 2c).

Features obtained from the single-breath counting task correlated well with vital capacity (R = 0.63) and strongly suggest that voice analysis could be a proxy for vital capacity measurements in a clinic. Similar results by others employing sustained phonation are in agreement with our new observations.

Importantly, semantic analysis of the picture description task was highly correlated with the ALS-Cognitive Behavioral Screen (CBS) (R = 0.72) and less correlated with the central nervous system (CNS) lability scale (R = 0.45). These studies then also suggest that at-home app analytics can be useful for longitudinal cognition analytics.

This task also predicted well the ALSFRS-R speech subscore (Fig. 2b); however, models using features from the reading task
outperformed the counting and picture description tasks. A more detailed account of these results is reported elsewhere.

These results demonstrate that the modules implemented to assess hand function and speech may be useful to quantify ALS function when patients are not in clinic and can substantially aid in the acquisition of progressively declining clinical indices. Furthermore, the picture description task may be useful to evaluate cognitive function in ALS. The potential to record voice and store it encrypted in the cloud could provide a powerful clinical tool to assess change over time that could be used clinically and in ALS trials.

**Production of the iPS cell line.** A core design and strength of the program are the set of iPS cell lines from a large population of >1,000 patients with ALS and control subjects, all deeply pheno-typed, provided to the research community. To date, more than 850 of the iPS cell lines have been generated and are available through the web portal. Out of the ~850 unique samples, only 18 lines (~2%) failed reprogramming. As there are multiple different protocols to generate iPS cells and differentiate them into motor neurons, it was essential that the uniformity of the generated cultures be evaluated, thereby establishing the reliability of this new and renewable biological resource. To address this central issue, we evaluated the iPS
Table 1 | Answer ALS basic clinical demographics

| Variable                        | Level                        | Overall: no. (%) | ALS: no. (%) | Asymptomatic ALS: no. (%) | Healthy control: no. (%) | Non-ALS MND: no. (%) | ALS versus healthy control |
|---------------------------------|------------------------------|------------------|--------------|----------------------------|--------------------------|----------------------|---------------------------|
| Participants                    | n                            | 100.0 (1,047)    | 82.2 (861)   | 1.1 (12)                   | 10.3 (108)               | 6.3 (66)             |                           |
| Sex                             | Female                       | 40.6 (423)       | 37.4 (320)   | 58.3 (7)                   | 66.4 (71)                | 37.9 (25)            | <0.001                    |
|                                 | Male                         | 59.4% (618)      | 62.6 (536)   | 41.7 (5)                   | 33.6 (36)                | 62.1 (41)            | <0.001                    |
|                                 | [missing]                    | (6)              | (5)          | (0)                        | (1)                      | (0)                  | N/A                       |
| Race                            | Native American              | 0.2 (2)          | 0.1 (1)      | 0.0 (0)                    | 1.0 (1)                  | 0.0 (0)              | 0.078                     |
|                                 | Asian                        | 2.0 (21)         | 1.5 (13)     | 0.0 (0)                    | 5.7 (6)                  | 3.0 (2)              | 0.004                     |
|                                 | Black                        | 4.8 (49)         | 5.0 (42)     | 0.0 (0)                    | 4.8 (5)                  | 3.0 (2)              | 0.928                     |
|                                 | Pacific Islander             | 0.1 (1)          | 0.1 (1)      | 0.0 (0)                    | 0.0 (0)                  | 0.0 (0)              | 0.724                     |
|                                 | [missing]                    | (18)             | (15)         | (0)                        | (3)                      | (0)                  | N/A                       |
| Ethnicity                       | Hispanic or Latino           | 4.8 (50)         | 5.3 (45)     | 0.0 (0)                    | 2.8 (3)                  | 3.1 (2)              | 0.271                     |
|                                 | Not Hispanic or Latino       | 95.2 (989)       | 94.7 (810)   | 100.0 (12)                 | 97.2 (104)               | 96.9 (63)            | 0.271                     |
|                                 | [missing]                    | (8)              | (6)          | (0)                        | (1)                      | (1)                  | N/A                       |
| Age at baseline (years)         | Mean (s.d.)                  | 58.9 ± 11.6      | 59.3 ± 11.1  | 48.3 ± 10.3                | 55.0 ± 14.1              | 61.9 ± 12.0          | <0.001                    |
|                                 | Time between symptom onset and diagnosis (months) | 15.9 ± 20.4 (−5.7, 286) | 14.8 ± 16.8 (−5.7, 185) | N/A | N/A | 40.8 ± 5.62 (0.1, 286) | N/A |
|                                 | Time between symptom onset and study enrollment (months) | 32.0 ± 39.4 (0.6, 458) | 29.8 ± 35.6 (0.6, 458) | N/A | N/A | 78.4 ± 75.3 (11.1, 353) | N/A |
| BMI at screening visit          | Mean (s.d.)                  | 26.8 ± 6.39      | 26.5 ± 4.83  | 29.2 ± 3.38                | 29.2 ± 14.9              | 27.3 ± 5.61          | <0.001                    |
|                                 | ALSFRS-R at first ALSFRS-R visit | 33.8 ± 8.65 (0.0, 47.0) | 33.8 ± 8.67 (0.0, 47.0) | N/A | N/A | 33.5 ± 8.44 (7.0, 46.0) | N/A |
|                                 | ALSFRS-R slope               | −0.73 ± 0.87 (−5.1, 1.4) | −0.77 ± 0.88 (−5.1, 1.4) | N/A | N/A | −0.11 ± 0.40 (−1.6, 1.0) | N/A |
|                                 | FVC (percentage predicted) at first ALSFRS-R visit | 69.9 ± 24.0 (4.0, 126) | 69.6 ± 23.9 (4.0, 125) | N/A | N/A | 73.7 ± 25.3 (17.0, 126) | N/A |
|                                 | FVC slope                    | −1.5 ± 2.53 (−16.1, 1.0) | −1.6 ± 2.59 (−16.1, 1.1) | N/A | N/A | −0.12 ± 0.86 (−1.9, 2.1) | N/A |
|                                 | Follow-up duration            | Months (mean (s.d.)) | 13.3 ± 17.3 (0.0, 340) | 12.5 ± 12.6 (0.0, 94.1) | N/A | 24.0 ± 47.2 (0.0, 340) | N/A |
|                                 | Time from onset to death      | Months (mean (s.d.)) | N/A | 34.7 ± 27.6 (8.3, 187) | N/A | N/A | N/A |

BMI, body mass index; N/A, not available.

---

cell-derived spinal neurons from a large cohort of 217 control and ALS iPS cell lines. Specifically, we examined expression of five different cell-identifying markers for neurons and glia, including cell markers NKX6.1, SMI32, ISL1, TUJ1 and S100beta. This differentiation protocol (Extended Data Fig. 3) generates a mixed population of neurons consisting of ~75% (±5%) βIII-tubulin- (TUJ1-) and ~70% (±10%) NF-H-positive cells, ~19% (±6%) Islet-1- and ~34% (±9%) Nlx6.1-positive spinal motor neurons, and ~18% (+/13%) S100B-positive progenitors 32 d after the onset of differentiation (Fig. 3 and Supplementary Table 6). As shown in Fig. 3, there was great uniformity in the cellular composition of the cultures for this large selection of human lines. This was important, because past work or methods can lead to variable cultures, making the interpretation of downstream analysis complicated. Notably the cellular composition was not substantially different between the ALS and control iPS cell-derived neurons. As expected, these cultures presented a mixture of motor neurons, neurons and, to a lesser extent, glia. This was important, because ALS is not simply a motor neuron disease, but is a disorder of multiple different nervous system cell types, as reflected in these uniformly generated cultures.

**Generation of multi-omics data.** Genomics. As an appreciation of the overall diversity of the program’s ALS and control population, especially valuable for future global analytics, we evaluated the AALS cohort using New York Genome Center’s (NYGC’s) ancestry pipeline. Most participants were white and of European descent (91.45%); the remainder had ancestry consistent with the Americas (1.69%), Africa (4.94%) and east (1.33%) and south Asia (0.6%)
On average, each sample harbored a total of ~4.1 million variants and ~9,800 protein-altering variants, including SNPs, frameshift and nonframeshift deletions and insertions, and protein-truncating variants (Table 2 and Fig. 4a–d), similar to previous reports\(^6\). Notably, the samples with African descent had a higher number of variants than other ethnic populations, as expected (Fig. 4b)\(^6\).

We used PCA\(^12,13\) to visualize the ancestry background of the AALS cohort and a set of 2,504 samples from the 1000 Genomes Project with well-defined ancestry. We find that most of the samples clustered with the NYGC’s European samples, although some were closer to the African group and a few clustered with the Asian group (Fig. 4e), corroborating the NYGC ancestry results and probably consistent with the local recruiting clinics geographic locations (Extended Data Fig. 1).

**Variants in ALS genes.** As most of the ALS lines were derived from patients with sALS, an analysis of the genomic variants is important, especially as future opportunities for researchers to correlate the observed variants along with the deep clinical and multi-omics data, as well as the future use of the living cell lines. Within the 830 samples, we observed 440 exonic variants in the 33-ALS genes (Supplementary Information) that were <1% frequent (Fig. 4c,d, Table 2 and Supplementary Table 7). Both controls and ALS cases averaged 1.5 rare ALS variants per individual within the 33-ALS genes. Of these, 79% were SNPs, 13% uncharacterized, ~1% splicing, ~1% nonframeshift deletion, 1% frameshift deletion, 1% frameshift insertion, 2% frameshift insertion, 2% nonframeshift insertion and 1% stop-gain (Supplementary Table 7).

As future biological pathways in ALS subgroups could reflect the expression of genetic variants of established ALS genes, we first evaluated how many pathogenic or probably pathogenic variants existed as reported in ClinVar (CP) in the 33-ALS genes. We found that 12% of ALS cases harbored a CP variant within one of the 33-ALS genes (Supplementary Tables 7 and 8). All of these CP variants were rare (<1% frequency within the population) except two found within the OPTN gene. For example, we observed five SOD1 CP variants (within eight patients with ALS), two TDP43 CP variants (within two patients with ALS) and one CP FUS variant in a patient with ALS (Supplementary Tables 7 and 8). CP variants were also detected in individuals who did not show signs of ALS at the time of the clinic visit, and there were eleven CP variants within control samples (within ALS2, SETX, OPTN and PFN1), four CP variants in the pre-fALS cohort (within FIG4, OPTN and CHCHD10), three CP variants within individuals with other MNDs (within SQSTM1, OPTN and PFN1) and three CP variants in uncharacterized individuals (within SQSTM1 and SETX; Supplementary Table 8). In summary, rare CP variants were observed in 3.11% (22 total) of ALS cases and 1% of controls (1 out of 92 samples). We also investigated the number of P/LP variants called by Intervar (IP), in silico prediction (ISD variants) and a new combination of ACMG gene criteria as well as the in silico prediction and family-based segregation data, a list of high-confidence causal variants in 12 genes—ALS2, CCNF, CHCHD10, FUS, OPTN, PFN1, SOD1, TARDBP, TBK1, UBQLN2, VAPB and VCP—which have been curated and designated as the HP (Harms P/LP, Supplementary Table 7) variants. These are reported in Supplementary Tables 7–11.

---

**Fig. 2 | Smartphone use and analytics (n = 80 biologically independent samples).** a. Smartphone app compliance mean and 95% confidence interval (CI). Compliance was calculated using the average number of tasks done per day and per subject. b. Results of inferring ALSFRS-R total. Pearson’s values are shown in black contoured bars (left, y axis) and mean absolute errors of the prediction are shown in color bars with 95% CI (right, y axis). Performance values were obtained using each individual task as well as the combination of all the tasks. The highest performance was obtained using all tasks (R = 0.89, P < 1x10\(^{-5}\)) LH, left hand; RH, right hand. c. Results of inferring ALSFRS-R scores using only speech-related tasks. Pearson’s values are shown in black contoured bars (left, y axis) and the mean absolute errors of the prediction are shown in color bars with 95% CI (right, y axis). Performance values were calculated independently for each of the three speech tasks to infer FVC and ALSFRS-R speech and bulbar subscores. Highest performance was obtained using information from the reading task for both ALSFRS-R subscores, obtaining up to R = 0.89 (P < 1x10\(^{-5}\)) or ALSFRS-R bulbar subscore. On the other hand, counting task information produced the best result when inferring the FVC score (R = 0.65, P = 2x10\(^{-2}\)).
We investigated CP, IP and ISD variants found across all genes in 830 samples and these are listed in Supplementary Tables 12, 13 and 14. Expansions in C9orf72 and ATXN2. Genomic expansions of both C9orf72 and ataxin 2 are associated with both fALS and sALS. The availability of large numbers of iPS cell lines and the matched multi-omics data from this phenotypically variable genetic sub-group provide a unique future opportunity to investigate these genes that alternatively lead to ALS and/or FTD. Using Expansion Hunter to identify repeat expansions within whole-genome sequencing (WGS) data, we found 601 expanded regions in the 830 samples14. In total, 41 patients with ALS and 4 pre-fALS subjects in the AALS study population harbored hexanucleotide expansions in C9orf72 that were >26 repeats (Fig. 4f and Supplementary Table 15). We also observed 35 patients with ALS, 4 controls and 1 uncharacterized individual harboring CAG triplet repeat expansions in ATXN2 >26 repeats (Fig. 4g and Supplementary Table 16). All patients with ALS with >26 ATXN2 repeats had clinical phenotype characteristics of MNDs and no other reported neurological abnormalities. Notably, in this population of patients and cell lines, for carriers of expansions in both ATXN2 and C9orf72 simultaneously, we found no correlation between age of ALS onset and expansion size (Fig. 4h,i and Supplementary Tables 15). However, future multi-omic studies of the patient iPS spinal neurons may reveal different biological pathways/properties when both mutations are co-expressed in humans.

ACMG genes. Pathogenic or probable pathogenic variants in 59 genes are currently considered to be medically actionable by the American College of Medical Genetics and Genomics (ACMG), due to the potential for medical intervention to modify morbidity and mortality in carriers of such variants15. Within the 830 samples, we identified 73 C-PLP variants within 32 ACMG genes (Supplementary Table 17). Of the individuals, 50.4% did not harbor a C-PLP variant in an ACMG gene, 41.2% harbored 1, 7.6% harbored 2 and 0.84% harbored 3 C-PLP variants. Of these variants found within 110 individuals, 66 were rare (<1%; Supplementary Table 17). We also found 42 I-PLP variants within ACMG genes within 51 individuals, all of which were rare (Supplementary Table 18). Participants were offered to receive the results of these medically actionable genes through the return of genetic results substudy (Extended methods).

Transcriptomics. For each of the omics assays, vials from an identical pool of differentiated motor neurons were processed to ensure comparability, including batch differentiation controls (BDCs) and batch technical controls (BTCs) from the control 2AE8 line, as
detailed in Extended methods. Overall the analytics revealed minimal to no technical confounders and low batch effects between differentiation and no clear batch-related abnormalities with regard to disease status (Extended Data Figs. 4a,d and 5a).

Annotation of transcripts detected in the samples revealed various RNA species that were captured in the deep sequencing, with protein-coding RNAs accounting for most (~82%) of all RNAs, followed by long intergenic noncoding (linc)RNA (~13%) (Fig. 5a). A low proportion of reads mapped to small RNAs and a very minimal proportion to ribosomal RNAs, which were depleted during library preparation and act as a technical quality assessment. The use of total RNA-sequencing (RNA-seq) and deeper sequencing allows new alterations in ALS and potential associations with ALS subtype iPS cell model. Notably, these data can be explored for additional alternative splicing analyses, as well as circular RNA and cryptic exon analyses (Fig. 5c,f). As an example of RNA-seq analyses, we assessed the ability of our cell model and RNA-seq methods to capture common, alternative splicing types and found significant enrichment in skipped exon (SE, 52%) and retention of introns (RIs, 35%) when comparing male C9 samples with male controls (Fig. 5e). RNA-binding protein (RBP) motif enrichment analysis of the significant RI events (cryptic exons) predicts that the binding of HNRNPA2B1 (Fig. 5f) is upregulated in ALS samples. These findings are consistent with previous reports in human post-mortem brain tissue[6].

To assess pathway activities, we used gene set variation analysis (GSVA) to score samples against canonical Kyoto Encyclopedia of Genes and Genomes (KEGG) and Biocarta pathways from the MsigDB database, and identified pathways that are differentially regulated between subjects with bulbar and limb onset (Fig. 5g). Using these pathway activity scores, we also identified pathways that are positively or negatively correlated with the patient ALSFRS progression slope (Fig. 5h).

These data indicate that both gene expression differences and RNA-splicing differences could be captured by our differentiated iPS cell model. Notably, these data can be explored for additional new alterations in ALS and potential associations with ALS subtype and clinical data, and with other omics data that are being captured from these samples.

Epigenomics. Overall the quality of transposase-accessible chromatin using sequencing (ATAC-seq) data was high, with very good
reproducibility of BDCs and BTCs, as assessed by the simple error rate estimate (SERE) (Fig. 5b, Extended Data Figs. 4b,e and 5b, and Supplementary Information). Hypersensitive sites were distributed across the genome in the expected regions (Extended Data Fig. 6a,b), especially in previously annotated regulatory regions, with very few reads in ENCODE blacklist regions. Although, overall, samples did not cluster by genotype or disease status, many loci did show strong differences between patients and controls (Extended Data Fig. 6c). As an example of a potential application of the epigenomic data, we identified potential transcriptional regulators through analysis of sequence motifs in the open chromatin (Extended Data Fig. 6d). Consistent with the expected cell composition, we observed an overrepresentation of transcription factors implicated in neuronal differentiation, such as Pdx1, Cux2 and the Lhx family (Extended Data Fig. 6d).

**Proteomics.** In total, >25,000 peptides corresponding to >3,600 proteins per sample were quantified. As detailed in the Supplementary Information, for proteomic analyses, there was minimal drift between the batches (Fig. 5c and Extended Data Figs. 4c,f and 6c). Although patient and control iPS neuron clusters are interspersed, indicating their overall similarity, these iPS neuron models have significant individual protein-level differences and we selected representative proteins ECH1 and PCKGM (Fig. 5d) that show significant (P ≤ 0.05) differences, based on what is seen in the differential analysis-based evidence (Fig. 5d).

**Longitudinal single-cell imaging and analysis.** Validation of the identification of pathological phenotypes was achieved with longitudinal single-cell robotic imaging of mutant SOD1 patient-derived iPS spinal neurons as described previously (Fig. 6a)17. As shown in Fig. 6b, mutant SOD1 neurons exhibited an enhanced cell death profile, similar to that reported previously with spinal motor neurons18. Future data will be available on similar analytics of cohorts of the sporadic iPS cell-derived neurons from the AALS dataset.

**Data dissemination: data portal.** The AALS data portal (http://data.answeraals.org; Supplementary Table 3) was designed to provide information about the various types of biological and clinical data generated by the AALS partners and to allow easy visualization/access to the metadata and data, along with links to obtain biofluids and iPS cell lines. Additional details regarding the portal can be found in Extended methods. In the future, the portal will also host online data analytics and visualization tools.

---

**Table 2 | Summary table of variants in the AALS cohort**

| Variant type | Total variants in all genes in ALS cases | Total variants in all genes in CTRLs | ALS gene variants in ALS | ALS gene variants in controls | Number of variants per 33-ALS gene |
|--------------|----------------------------------------|-------------------------------------|--------------------------|-----------------------------|----------------------------------|
| All variants | Sum = 2,941,489,030 Average = 4,166,415 variants per ALS case | Sum = 379,092,863 Average = 4,120,575 variants per control | Sum = 1,092 Average = 1.5 variants per ALS case | Sum = 141 Average = 1.5 variants per control | ALS2 (20), ANG (5), ANXA11 (15), ATXN2 (29), C2orf2 (19), C9orf72 (5), CAMTA1 (24), CCNF (28), CHCHD10 (2), DAO (7), DCTN1 (24), FIG4 (14), FUS (6), HNRNPA1 (2), HNRNPA2B1 (1), KIF5A (9), MATR3 (10), MOBP (4), NEK1 (19), OPTN (10), PFIN (7), SCFD1 (13), SETX (57), SODI (14), SQSTM1 (14), TAF1S (16), TARDBP (11), TBK1 (18), TUBA4A (3), UBQLN2 (7), UNCT3A (19), VAPB (4), VCP (4). |
| ClinVar P/ LP (C-PLP) variants | Sum = 23,924 Average = 33.9 variants per ALS case Rare = 3,659 (5.2 variants per ALS case) | Sum = 3,097 Average = 33.7 variants per control Rare = 61 (5 variants per control) | Sum = 85 (2% of cases harbor) Rare only = 21 (3% of cases per control) | Sum = 11 (12% of controls harbor) Rare only = 3 (3.3% or control harbor) | ALS2 (1) ANG (2), CHCHD10 (1), FIG4 (2), FUS (1), OPTN (2), PFIN (2), SETX (4), SODI (5), SQSTM1 (3), TARDBP (2), UBQLN2 (2), VCP (1) |
| Harms P/ LP (H-PLP) variants | N/A | N/A | Sum = 4 (3.4% of cases harbor) | Sum = 1 (1% of controls harbor) | FUS (1), PFIN (2), SODI (11), TARDBP (3) UBQLN2 (1), VCP (1) |
| Intervar P/ LP (I-PLP) variants | Sum = 2,346 Average = 3.3 variants per sample Rare = 2272 Average = 3.21 variants per case | Sum = 288 Average = 3.1 variants per sample Rare = 276 Average = 3.2 per control | Sum = 25 (3.5% of cases harbor) | 0 (0%) | NEK1 (2), OPTN (1), SODI (12), SETX (1), TBK1 (2), VCP (2), |
| In silico prediction: 6/9 predicted to be damaging | Sum = 79,010 Average = 112 variants per sample Rare = 40,910 Average = 58 variants per sample | Sum = 5,464 Average = 113 variants per sample Rare = 5,464 Average = 59.4 variants per sample | Sum= 97 (13.7% of cases harbor) | Sum=11 (12% of controls harbor) | ALS2(2), ANXA11 (4), ATXN2 (3), C2orf2 (1), CAMTA1 (1), DAO (3), DCTN1 (5), FIG4 (3), FUS (1), HNRNPA2B1 (1), KIF5A (2) MOBP(1), NEK2(2), OPTN (1), PFIN (3), SCFD1 (2), SETX (14), SODI (11), SQSTM1 (2), TARDBP (4), TUBA4A (2), UBQLN2 (1), UNCT3A (3), VCP (2) |

Sum = the total number of variants found per group, ALS versus control. Variants <1% of OPTN variants listed here are high frequency, >1%.
Discussion

The pathogenesis of sALS remains a mystery and few comprehensive data collections, on a population scale, exist to truly inform researchers about the biological underpinnings of the disease or the possibility of disparate biological subgroups. To date, clinical studies alone have not yielded reliable data to suggest a common pathway or, more importantly, a means to target relevant biological subgroups. The identification of biological subgroups has been impactful in various cancers, where the ability to actually sample disease tissues from skin, liver, prostate or pancreas biopsies, coupled with clinical characteristics of tumor type, has led to marked improvements in therapeutic approaches, drug treatments and decisions about disease management.23–25.

The core goal of AALS is to provide a comprehensive set of tools including deeply phenotyped longitudinal clinical data and biological tools such as iPS cell lines, and a multi-omics platform consisting of whole-genome, iPS-derived, spinal neuron-enriched proteomes, transcriptomes and epigenomes, to uncover underlying biological subgroups. Previous studies have demonstrated the ability to generate small populations of fALS or sALS iPS cell-derived motor neurons and glia, as well as relatively limited multi-omics data. However, none approximates true population-based tools, with reproducible quality assurance protocols, necessary to accurately assess disease pathways or identify population subgroups combining longitudinal clinical, genomic and living multi-omics data.19,20.

The AALS reagent collection includes individual iPS cell lines from approximately 850 sALS and control participants (soon to reach 1,200), the iPS cell-derived spinal neurons from each participant, their longitudinal clinical data (collected over 1 year), sequentially assayed fluid biospecimens (blood and cerebrospinal fluid (CSF)) and the early multi-omics data generated from each participant’s blood (whole genome) as well as from their ‘spinal cord biopsy’-equivalent, iPS-derived neuronal cell lines. The collection also includes autopsy samples and pathology data from a subset of participants. The autopsy pathology data and CNS specimens will eventually be available through the AALS web portal and coupled with the iPS cell lines from these participants.

A reasonable question is the utility of patient-derived iPS cells to predict the disease-causing pathways in an adult-onset disease. Can reprogrammed human spinal neurons reflect adult-onset disease pathogenic cascades? Already multiple studies have documented that human iPS cell lines, in either two-dimensional cultures or three-dimensional organoids, can reproduce the pathology seen in human brain.26–28. One advantage of the iPS platform is the ability to dynamically detect early pathogenic events and even serially occurring events. In fact, early use of the AALS iPS cell lines has already provided evidence that the iPS collection can provide insights into new pathways (nuclear pore complex and nuclear transport defects)

Fig. 5 | Omics exploratory analysis of results. a, Violin plot showing counts of RNA species identified in the current AALS samples. As expected, protein-coding and lincRNAs represent the largest proportions whereas rRNAs, which have been depleted, are the lowest. Minimal variability has been observed among samples. Types represented are: protein coding, lincRNA, miRNA, small nuclear RNA, small nucleolar RNA and rRNA in green, red, gold, purple, blue and teal, respectively (n = 102 biologically independent samples). b, Peak functional annotations. Analysis of read distribution across all ATAC-seq samples shows an enrichment in known open chromatin regions, such as DNase 1-hypersensitive sites and previously annotated enhancers and promoters (n = 100 biologically independent samples). c, The log2(protein intensity distribution) unnormalized (top) and normalized (bottom). d, The log2(protein intensity) comparison of selected proteins (PCKGM, ECH1) showing differential expression between ALS and controls. Box plots in c and d indicate median, quartiles and range (n = 66 biologically independent samples). e, Pie chart of proportions of rMATs analysis of differentially alternative splicing identified events comparing male CA9orf72 ALS samples versus male controls. An FDR cutoff of 0.05 was used to define statistical significance. SE has the highest number of events (n = 617, 52%), followed by RIs (n = 409, 35%), f, The mAPTS2-based motif enrichment analysis of alternatively RIs (409 R1 events) shows that the RBP-binding motif HNRNPA2B1 is significantly enriched in the male control samples versus male CA9orf72 ALS samples near the RI sites. Wilcoxon’s rank-sum test (one sided) was used to get the P values for comparing up- and downregulated exons (RI) versus control/background exons. Motif scores are plotted in solid lines and P values are in dotted lines. Red designates control samples and blue the ALS. g, Heatmap of pathway activity scores defined by GSEA against MsigDB’s C2 canonical pathways from KEGG and Biocarta. The top 30 pathways are shown from comparing samples with bulbar versus limb ALS disease onset (FDR < 0.05). h, The top 14 pathways that have high Pearson’s correlation between GSEA enrichment scores and ALSFRS clinical progression slope.
original library of 22 fALS iPS cell lines21, with a few selected lines for each disease mutation and, when appropriate, isogenic controls. None represents the far more common sporadic forms of the disease. Furthermore, none provides long-term longitudinal clinical and extensive multi-omics data.

Aside from the biological data generated from the program, the results from the AALS smartphone app demonstrate that the modules implemented to assess limb function, speech and cognition may be useful to identify early bulbar and cognitive symptoms in ALS and track disease progression over time. Specifically, limb-function tests reveal that it can be useful to infer ALSFRS-R scores. Importantly, we observed that, by combining the features from multiple domains, motor tests and all the voice tests highly correlated with the ALSFRS-R, now commonly used as a primary or secondary outcome measure in ALS clinical trials, thereby providing a reliable tool for at-home longitudinal monitoring of patient progression. Furthermore, the single-breath testing also correlated well with in-clinic forced vital capacity (FVC), often a prominent secondary outcome measure in clinical trials. This test typically requires in-clinic testing, which limits enrollment or follow-up data collection in clinical trials. The application of this test alone could greatly enhance patient participation in nationwide clinical trials—especially in those areas where travel to a testing center is challenging. Overall, we observe that quantitative motor speech analysis holds tremendous promise in both identifying changes limited not only to ALS rating scales but also to others such as cognitive assessment. The potential to record voice, and store it encrypted in the cloud, could provide a powerful clinical tool to assess change over time for use clinically and in ALS trials. Overall, the app data, coupled with in-clinic data, provide deep and longitudinal clinical datasets available for multi-domain biological and clinical correlations for future users.

The overall clinical demographics and population genomics in the AALS program accurately reflect the ALS subject population described in previous studies. This observation validates the AALS iPS cell lines and multi-omics platform as a database that others can employ to generate and test biological hypotheses. Importantly, all the clinical data, multi-omic data and iPS cell lines were generated to be freely accessible to all researchers, academic and commercial, free of restrictions other than standard Health Insurance Portability and Accountability Act (HIPAA) compliance rules. A web portal for downloading filtered datasets, for example, proteome, whole genome, etc., has been set up with minimal but appropriate requirements for data access (Supplementary Table 3). The ALS and control iPS cell lines, matched to datasets, are also fully available for research studies, for a minimal fee (to cover the replacement of the depleted stock of cells). Biospecimens (for example, CSF and plasma) longitudinally collected from patients are also available (Supplementary Table 3). Future web-based links will include access to autopsied CNS tissues from patients matched to the iPS cell lines and iPS cell-based multi-omics.

Online content
Any methods, additional references, Nature Research reporting summaries, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-021-01006-0.

Received: 20 May 2021; Accepted: 16 December 2021; Published online: 3 February 2022

References
1. Hooiveld, V. et al. Medulloblastomas revisited: biological and clinical insights from thousands of patients. Nat. Rev. Cancer 20, 42–56 (2020).
2. Katyal, N. & Govindarajan, R. Shortcomings in the current amyotrophic lateral sclerosis trials and potential solutions for improvement. Front. Neurol. 8, 521 (2017).
3. Phillips, T. & Rothstein, J. D. Rodent models of amyotrophic lateral sclerosis. Curr. Protoc. Pharm. 69, 56761–21 (2015).
4. Donnelly, C. J. et al. RNA toxicity from the ALS/FTD C9ORF72 expansion is mitigated by antisense intervention. Neuron 80, 415–428 (2013).
5. Sareen, D. et al. Targeting RNA foci in iPSC-derived motor neurons from ALS patients with a C9ORF72 repeat expansion. Sci. Transit. Med. 5, 208ra149 (2013).
6. Taylor, J. P., Brown, R. H. Jr. & Cleveland, D. W. Decoding ALS: from genes to mechanism. Nature 539, 197–206 (2016).
7. Agurto, C. et al. Analyzing progression of motor and speech impairment in ALS. Annu. Int. Conf. IEEE Eng. Med Biol. Soc. 2019, 6097–6102 (2019).

8. Stegmann, G. M. et al. Estimation of forced vital capacity using speech acoustics in patients with ALS. Amyotroph. Lateral Scler. Frontotemporal Degener. 22, 14–21 (2021).

9. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19, 1655–1664 (2009).

10. Auton, A. et al. A global reference for human genetic variation. Nature 526, 68–74 (2015).

11. Genomes Project, C. et al. A global reference for human genetic variation. Nature 526, 68–74 (2015).

12. Patterson, N., Price, A. L. & Reich, D. Population structure and ancestry in unrelated individuals. Genome Res. 27, 1895–1903 (2017).

13. Kalai, S. S. et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genet. Med. 19, 249–255 (2017).

14. Prudencio, M. et al. Distinct brain transcriptome profiles in C9orf72-associated and sporadic ALS. Nat. Neurosci. 18, 1175–1182 (2015).

15. Lindsay, J. W. et al. Automated four-dimensional long term imaging enables single cell tracking within organotypic brain slices to study neurodevelopment and degeneration. Commun. Biol. 2, 155 (2019).

16. 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).

17. Zhang, H. et al. Subgroup analysis reveals molecular heterogeneity and clinical outcomes in ALS. Neurology 84, 135–143 (2015).

18. Vass, R. et al. Risk genotypes at TMEM106B are associated with cognitive impairment in amyotrophic lateral sclerosis. Acta Neuropathol. 121, 373–380 (2011).

19. Elden, A. C. et al. Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. Nature 466, 1069–1075 (2010).

20. Kwart, D. et al. A large panel of isogenic APP and PSEN1 mutant human iPSC neurons reveals shared endosomal abnormalities mediated by APP beta-CTFs, not abeta. Neuron 104, 256–270.e255 (2019).

21. 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).

22. Marchetto, M. C. et al. A model for neural development and treatment of Rett syndrome using human induced pluripotent stem cells. Cell 143, 527–539 (2010).

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

© The Author(s) 2022
Methods

Program process, Overall design (Extended Data Fig. 1). The overall AALS program, from clinical enrollment to smartphone app data collection, iPS cell-line generation, biological data generation and data storage is outlined in Extended Data Fig. 1 (ClinicalTrials.gov: NCT02573490). Methods for each element of the program are provided below and in Supplementary Methods.

Enrollment, clinical characterization and sample collection. The clinical portions of AALS were coordinated through Johns Hopkins University and Massachusetts General Hospital. The eight enrolling neuromuscular clinics were distributed across the USA and included Johns Hopkins University, Massachusetts General Hospital, Ohio State, Emory University, Washington University, Northwestern University, Cedars-Sinai and Texas Neurology (Supplementary Table 1 and Extended Data Fig. 1). The study was approved by local institutional review boards, and all participants provided written informed consent. Consent was uniform across all sites and included agreement to share data broadly for medical research (also see Data access in Supplementary Information). Subjects with aALS, fALS and related MNDs (referred to as non-ALS MNDs), including those with primary lateral sclerosis, progressive bulbar palsy and progressive muscular atrophy, along with asymptomatic ALS gene mutation carriers, were enrolled in AALS. Age-matched control participants without ALS or a family history of ALS were also enrolled. Additional enrollment details are provided in Supplementary Information.

Participants were monitored every 3 months for a year and, when possible, the ALSF5-R was conducted by telephone every 3 months for another year thereafter. Baseline descriptors included the following: demographics and vital signs, genotyping of MND, general medical history, CNS exam, and a brief focused history of environmental exposures. Concomitant medications and past medical history were collected at enrollment and updated throughout study participation. Measures of ALS progression included: deep tendon reflexes, Ashworth Spasticity Scale, Hand Held Dynamometry, ALSF5-R and pulmonary slow vital capacity (Supplementary Tables 2 and 3 and Supplementary Information). To enhance depth of longitudinal clinical data collection, a secure and HIPAA-compliant smartphone app, with a specific focus on motor activity, voice and cognition, was created for home data collection (Fig. 2 and Extended Data Fig. 2). At each in-clinic visit, blood was collected and processed according to the methods outlined in Supplementary Information. At the first visit, whole blood was collected for generation of primary peripheral blood mononuclear cell (PBMC)-derived iPS cell lines.

Biofluid collection and processing. At each in-clinic visit along with follow-up visits, approximately 50–100 ml of blood was collected from each participant. Plasma and serum were processed for storage and PBMC isolation. Whole blood was sent to the NYGC for DNA extraction and WGS. CSF was optionally collected and flash frozen at −80°C (Supplementary Table 6). Additionally, PBMCs were isolated from whole blood for generation of primary-iPS cells. Additional details are provided in Supplementary Information.

Return of AALS results. To provide medical and ethically appropriate feedback, study participants with ALS were offered the opportunity to receive the results of WGS. Whole blood was sent to the NYGC for DNA extraction and WGS. CSF was optionally collected and flash frozen at −80°C. CSF from each cell line for immunostaining was reserved for QCs, which included the following: markers of neuronal differentiation: SMN2 (SMN), TUBB3 (TUB), ISL1, NKX6.1, S100B and Nestin. This protocol generates a mixed population of neurons consisting of ~75% (±8%) β3-tubulin (Tub) and ~70% (±10%) NF-H-positive cells, ~19% (±6%) Islet-1 and ~34% (±9%) NKx6.1-positive motor neurons, and ~18% (±13%) S100B-positive progenitors 32 d after the onset of differentiation (Fig. 3). Additional details are provided in Supplementary Information.

Generation of iPS cell-derived motor neuron line. At the end of the 32-d differentiation protocol, the spinal neurons were harvested for RNA-seq, proteomics or epigenome profiling as detailed in Supplementary Methods. WGS was performed on PBMCs. Day 32, chosen from independent experiments with selected C9orf72 ALS/FTD iPS cell-derived spinal neurons, demonstrated phenotypic and molecular changes in nuclear pore complex and biology, matching that seen in patient autopsies, by this time point.

Program QCs: cell generation batch controls. To detect and compensate for cell culture-associated confounders, all differentiations were conducted in a single facility and included two key control groups of biological samples: BDCs were differentiated with each batch from the same original line to assess interbatch variability of iPS cell differentiation to dMNAs and BTCs, consisting of a single line from the same cell bank and characterized per donor. To assess technical variability of the omics assay batch runs, were performed as detailed in Supplementary Information. Complete details for the design and implementation of these critical operational controls (Extended Data Figs. 4 and 5) can be found in Supplementary Information.

Data quality and batch effect assessments. RNA-seq. For the RNA-seq data samples were processed and passed all QC metrics including RNA integrity (Extended Data Fig. 4a), library and sequencing QC metrics. To assess data quality and technical batch effects, sample-to-sample SERE scores (0 identical samples) were generated using gene expression for three groups: the BDCs, BTCs and all other samples (Extended Data Figs. 4 and 5).

A heatmap of SERE scores between all samples with hierarchical clustering (Extended Data Fig. 5) shows that, although BTCs form their own cluster, the rest of the samples fall into multiple small clusters with no clear relationship to their disease status.

Proteomics. Each block of samples comprised case, control, BDC samples and HEK293 cell control samples. The numbers of proteins and peptides quantified for all 66 samples were very consistent (Extended Data Fig. 4e). The percentage coefficient of variation for the proteins quantified were calculated for the BTC and BDC samples (Extended Data Fig. 4f). Individual samples are normalized to the total MS2 spectra intensity across the chromatographic profile of eluting peptides to smooth any inconsistencies in sample loading on to the mass spectrometry (MS) instrument, thereby eliminating systemic variation in signal intensities (Extended Data Fig. 4c). We found that BTCs and BDCs (both originating from the 2AE8 CTR cell line) cluster tightly (Extended Data Fig. 6c), indicating minimal drift between the MS batches.
Epigenetics. ATAC-seq data quality was determined according to ENCODE46. The distribution of fragment sizes across all samples revealed a clear nucleosome-free region and regular peaks corresponding to nucleosomal fractions (Extended Data Fig. 3a). As expected, replicates from our biological control line were highly correlated with each other, with BTCs having an even smaller variation in correlation values compared with BDCs (Extended Data Fig. 4e). We also generated a consensus set of peaks present in >10% of samples using DiffBind (Extended Data Fig. 6) and characterized transcription factor motif enrichment within these peaks using HOMER. There was an overrepresentation of transcription factors implicated in neuronal differentiation, such as Pax6, Cux2 and the Lhx family (Extended Data Fig. 6d). We then obtained a counts matrix of reads mapped to each peak in the consensus peakset across all samples and performed hierarchical clustering using the same approach as the RNA-seq data (Extended Data Figs. 4, 5 and 6). Subjects did not cluster by disease status, presence of C9 mutation, sex or processing batch. Additional data on quality control can be found in Supplementary Methods.

Whole-genome methods: WGS and analysis. PBMCs were sent by each clinic to the NYGC (https://www.nygenome.org) for DNA extraction and sample QC and WGS; matched peripheral blood mononuclear cells (PBMCs) and patient-determined pathogenic variants reported in ClinVar (C-PLP) for all genes. We also examined pathogenic variants called by Intervar Li+ (I-PLP) and predicted damaging variants as called by silico prediction tools (IS-D), which are reported in Table 2 and Supplementary Table 8. The variant calls from NYGC were assessed by examining the actual reads mapped to the IUAM file and spot checking the specific variants in Integrative Genomic Viewer determined to be of good quality. The variant call formats (VCFs) were converted into genomic VCFs (GVCFs), and joint genotyping calling was run using Sonentie v.201911 (https://www.sonentie.com); applied variant quality score recalibration (VQSR) was done using GATK v.3.8 (true sensitivity level = 99.0), and the files were annotated using Annovar v.2018Apr16 (ref. 39). This also included expression and SNPs from nine programs, including databases such as SIFT40, PolyPhen2 (ref. 41) and Mutation Taster42, Mutation assessor68, FATHMM69, LRT_Prediction67, Mutation Taster42, and those described in Li et al.43. Additional databases were included that assess the variant tolerance of each gene using the Residual Variation Interpretation LoFTool46 will be incorporated. To identify variants in genes that are highly expressed in the brain, we incorporated data from the Human Protein Atlas (http://www.proteinatlas.org) and the GTEx portal75,76 (https://gtexportal.org/home) for expression data, and the Gene Expression Omnibus80. Additional regulatory databases such as miRNA target sites within messenger RNAs (miRNA) target sites within messenger RNAs and miRNA* target sites in messenger RNAs and miRNA* were also used. Extensive details on the methods for whole-genome analytics can be found in Supplementary Methods.

RNA methods. Total RNA was isolated from each sample using the QIAGEN RNasy mini-kit. RNA QC was conducted using an Agilent Bioanalyzer and Nanodrop. Our primary QC metric for RNA quality is based on RNA integrity number (RIN) values ranging from 0 to 10, 10 being the highest quality RNA. In addition, we collected QC data on total RNA concentration and 260:280 ratio to evaluate any potential contamination. Only samples with RIN > 8 were used for library prep and sequencing. The RNA were removed and libraries generated using TruSeq Stranded Total RNA library prep kit with Ribo-Zero (QIAGEN). RNA-seq libraries were trimmed by quantitating (qiPCR (Kapa), normalized according to size (Agilent Bioanalyzer 2100 High Sensitivity chip). Each complementary DNA library was then subjected to 100 Illumina (Novaseq 6000) sequencing cycles on a 50 million high throughput sequencing, raw reads were subject to QC measures and reads with quality scores >20 collected and analyzed. Reads were mapped to the GRCh38 reference genome using Hisat2, QCed and gene expression quantified with featureCounts34, and differential expression was quantified using DESeq2 (ref. 47). Normalized and transformed count data were also used for exploratory analysis and differentially expressed genes (false discovery rate [FDR] <0.1) were analyzed with commercial and open-source pathway and network analysis tools, including Ingenuity Pathway Analysis, gene set enrichment analysis (GSEA), GOriilla, Cytoscape and other tools to identify transcriptional regulators, predict epigenomic changes and determine potential effects on downstream pathways and cellular functions.

ATAC-seq methods. We used the assay for ATAC-seq to assess chromatin accessibility and identify functional regulatory sites involved in driving transcriptional changes associated with ALS. ATAC-seq sample prep, sequencing and peak generation were carried out by Digenode Inc. as further described48. 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 nucleotides 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 2x 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 augmented DNA using Nextera sequencing 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 NextSeq platform (PE, 75–nt kit). All samples passed QC checks that included morphological evaluation of nuclei, fluorescent-based electrophoresis of libraries to assess size distribution and RT-qPCR to assess the enrichment of open chromatin sites. The quality of the sequencing was assessed using FastQC and the reads were aligned to GRCh38 genome build using Bowtie2. We identified open chromatin regions separately for each sample using the peak caller MACS2 and determined open chromatin peaks using DESeq2 (FDR <0.1). Peaks were assigned to unique genes using the default HOMER parameters, and gene ontology analysis was performed using GOriilla48.

Proteome methods. Whole-proteome extracts from frozen diMNs were digested with trypsin and LysC and subjected to acquisition on the SCIEX 6600 as detailed below. Snap-frozen cell pellets were stored at −80°C and transferred to the Cedars-Science Medical Center proteomics lab on dry ice, where it was stored at −80°C until use. Samples were lyophilized and aliquoted into 600-μl polystyrene microcentrifuge tubes containing lysis buffer (6 mM urea and 1 mM diithothreitol in 1.5 M NH4HCO3). The sample was sonicated (Qsonica Q800R1) by shaking at 10 s on and 10 s off at 70% amplitude while rotating in a 4°C water bath until the solution was homogenized (~20 min). Samples were centrifuged and the protein concentration determined on the supernatant according to manufacturer’s instructions (Pierce BCA Protein Assay Kit). Then 200 μg of each sample was transferred to a 96-well plate in aliquots and processed on the Biomek i7 Automated workstatation (Beckman Coulter) as outlined previously49. Briefly, samples underwent the following: reduction of disulfide bonds in 3 mM tris(2-carboxyethyl)phosphine hydrochloride solution, alkylated in 5 mM iodoacetamide, and digested with trypsin. Digestion of proteins in 3 mM tris(2-carboxyethyl)phosphine hydrochloride solution, alkylated in 5 mM iodoacetamide, and digested with trypsin. Digested proteins were desalted on a 5-mg Oasis HLB 96-well plate (Waters, catalog no. 186003039) and eluted in 50 μl acetonitrile. Samples were dried to completion using a vacuum-acid and stored at −80°C until MS analysis. For MS analysis, digested peptides were resuspended in 0.1% formic acid (FA) and analyzed on a 6600 Triple TOF (Sciex) in data-independent acquisition (DIA) mode and on the 6600 Triple TOF (Sciex) for data-dependent acquisition (DDA).
mode. Specifically, samples were acquired in DDA mode for ion library generation and in DIA mode over 100 variable windows, similar to previously described acquisition protocols. 

DDA data were used for the generation of a sample-specific peptide ion library. DDA files were run through a trans-proteome pipeline using a human canonical FASTA file (Uniprot). A consensus peptide library with decoys was generated and used to quantify ions identified in DIA data files. Previously described DDA library build principles were utilized to generate a cell-specific library, which allowed for greater accuracy in matching DIA data to the DDA library during OpenSWATH, as indicated by higher d scores in PyProphet. The differential protein expression between ALS and control samples was calculated using mapDIA.

DIA data files were analyzed using OpenSWATH pipeline against the sample-specific peptide ion library generated. Protein-level quantification is calculated by summing transition level intensities for all the prototypic peptides identified. Differential protein expression between ALS and control samples analyzed was calculated using mapDIA.

Imaging methods. Longitudinal single-cell imaging and analysis. Differentiated iMNs from a subset of the AALS iPS cell lines were plated on 96-well plates for longitudinal single-cell imaging using robotic microscopy as previously described. At day 25, cells were transduced with expression marker plasmids such as synapsin::EGFP to visualize cell morphology and viability. After acquisition protocols were applied, samples were acquired in DDA mode for ion library generation.

Specifically, samples were acquired in DDA mode for ion library generation. This nomenclature is applied consistently to all metadata and data tokenization. This nomenclature is applied consistently to all metadata and data

The first underscore token of the sample name consists of the following components: whether the sample is from a diseased patient or healthy control patient, the de-identified patient GUID, the sample vial number, the date of the data collection, the organ source (muscle, fluid, organ, etc.), and the experiment identifier.

Prefixes are applied systematically to all metadata and data. The first underscore token of the sample name consists of human identifier (first name, last name), and a unique identifier for the sample. The prefix separates the first from the subsequent file information, allowing for easy tokenization. This nomenclature is applied consistently to all metadata and data files, making it easy to establish relationships with a single study participant.

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

Data availability

All data supporting the findings of the present study are available within the paper, its Supplementary information files and the AALS web portals listed in Supplementary Table 3 (or via data.answers.org).

References

43. Elsheikh, B. et al. Correlation of single-breath count test and neck flexor muscle strength with spirometry in myasthenia gravis. *Muscle Nerve* 53, 134–136 (2016).
48. Consortium, G. T. Human genomics. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45, 580–585 (2013).
51. Tennessen, J. A. et al. Evolution and functional impact of rare coding genetic variation in 121,514 humans. *Science* 347, 1260419 (2015).
53. Petrovskii, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PloS Genet.* 9, e1003709 (2013).
55. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. A catalog of human genes and genetic disorders. *Am. J. Hum. Genet.* 88, D862–D868 (2016).
56. Landrum, M. J. et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* 44, D1038–D1046 (2016).
57. Green, R. C. et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* 15, 565–574 (2013).
59. Richards, C. S. et al. ACMG recommendations for standards and guidelines for interpretation and reporting of sequence variations: revisions 2017. *Genet. Med.* 19, 294–300 (2017).
Competing interests
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. R.N., C.A. and G.A.C. disclose that their employer, IBM Research, is the research branch of IBM Corporation. R.N, C.A. and G.A.C. own stock in IBM Corporation.

Ethics statement
The AALS trial and smartphone app were approved by the Johns Hopkins institutional review board (nos. 00082277 and 00240008). The AALS program is registered at clinicaltrials.gov (NCT02574390).

Additional information
Extended data are available for this paper at https://doi.org/10.1038/s41593-021-01006-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-021-01006-0.

Correspondence and requests for materials should be addressed to Jeffrey D. Rothstein.

Peer review information Nature Neuroscience thanks John Landers and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | Answer ALS Operations. Top. Answer ALS Research Program. Graphic illustration of overall program flow. Bottom. Clinical Sites. Participating clinics were districted nationally at 8 academic or private neurology clinics specializing in ALS clinical care and research.
Extended Data Fig. 2 | Smartphone App. a. Smartphone App. Illustrations from app of various activities. a’. Main Menu, b’. Upper limb motor tests, c’. Bulbar activities, including single breath counting, speech and cognition, d’. Example of cartoon used for speech/cognition analytics. b. Examples of speech and fine motor tasks performed by the smartphone app study participants. Data are collected with an app called “Help us Answer ALS”. Each week, the app asks the participant to perform different tasks. The tasks involve motor control in the upper body, speech and cognition. Each task is performed once per week. The speech tasks include describing a picture (a,b,c), reading a passage (d,e,f), and counting until the subject runs out of breath (not represented). Describing a picture also serves as a cognition task. The motor task involves tracing 3 different contours in sequential order (h,i,j), alternating hand each day of the week.
Extended Data Fig. 3 | Production of ALS and control iPS cell spinal motor neurons.  

**a.** Example of IPS Generation Schedule.  

**b.** Method of generating iPS cell-derived motor neuron cell lines using the diMNs protocol.  

**c.** Brightfield images show the morphology of the cells during differentiation from iPS cell stage to the generation of motor neurons over a period of 32 days.  

**d.** Production flow and harvesting schematic of diMNs for multi-omics analyses.  

**e.** Quality control of the diMNs produced from iPS cells is performed by imaging of representative wells for immunohistochemical staining with neuronal, motor neuron and glial markers after 32 days of differentiation. Scale bar=400μm. Images representative of over 600 patient cell lines.
Extended Data Fig. 4 | See next page for caption.
Extended Data Fig. 4 | Omics Quality Control metrics. a. Histogram of RNA integrity numbers for current AALS samples. Density plot and histogram of RIN values for all current AALS samples with RNAseq data. Plot shows all processed samples have RIN > 8. b. fragment size distribution Size distribution of ATAC seq data, with peaks representing different n-nucleosomal fragments and clear nucleosome-free regions separated by ~147 bp, the size of a nucleosome. c. Number of Proteins and peptide identification consistency in the data generation batches of AALS samples. d. Violin plot of SERE values for RNAseq data for current AALS samples. Violin plot showing variance of SERE values in BTC (green) and BDC (red) control samples relative to all other (blue) current AALS samples. BTC shows lowest score with the least amount of variance indicating that samples are true technical replicates, while BDC and other samples show increase variance. e. Violin plot of SERE values for ATACseq data for current AALS samples. Similar to RNA data the BTC (green) show lowest variability indicating low technical confounds. f. Coefficient of Variation (CV) for Batch Technical Control (BTC) and Batch differentiation control (BDC) replicates showing 80% proteins to be under a CV of 25%.
Extended Data Fig. 5 | Heatmap and hierarchical clustering of current AALS samples. a&b. Heatmap and hierarchical clustering of SERE values using RNA/ATACseq data. Heatmap and clustering of current AALS samples using SERE values from the (a) RNAseq and (b) ATACseq data. Samples are annotated with gender, genotype, and C9orf72 mutation. No distinct clustering separates samples by these categories, but BTC sample cluster together. c. Spearman correlation matrix plot for the AALS proteomics data.
Extended Data Fig. 6 | ATACSeq data. a and b. CDFs. The number of all peaks (a) and promoter peaks (b) that are common to different numbers of samples. (c) PLEKHG4b locus. (Left) ATAC-seq read density upstream of the PLEKHG4b gene for ALS (middle) and CTR (bottom) samples. Average coverage for each group is shown at the top. (Right) Zoomed in region around the starred peak. d. Motifs. The most overrepresented genomic motifs corresponding to known transcription factors as determined by the HOMER discovery algorithm for ATAC-seq. Motifs for transcription factors implicated in neuronal identity, such as Pdx1, Cux2, and the Lhx family, are significantly enriched.