Answer ALS: A Large-Scale Resource for Sporadic and Familial ALS Combining Clinical Data with Multi-Omics Data from Induced Pluripotent Cell Lines

Jeffrey Rothstein (jrothstein@jhmi.edu)
Johns Hopkins University https://orcid.org/0000-0003-2001-8470

James Berry
Massachusetts General Hospital

Clive Svendsen
Cedars-Sinai Board of Governors Regenerative Medicine Institute https://orcid.org/0000-0001-8696-3446

Leslie Thompson
Univ California Irvine

Steven Finkbeiner
Gladstone Institute https://orcid.org/0000-0002-3480-394X

Jennifer Van Eyk
Advanced Clinical Biosystems Research Institute, The Smidt Heart Institute, Cedars Sinai Medical Center https://orcid.org/0000-0001-9050-148X

Ernest Fraenkel
Massachusetts Institute of Technology https://orcid.org/0000-0001-9249-8181

Merit Cudkowicz
Massachusetts General Hospital

Nicholas Maragakis
Johns Hopkins University https://orcid.org/0000-0002-7311-9614

Dhruv Sareen
Cedars-Sinai Medical Center https://orcid.org/0000-0002-0898-9656

Raquel Norel
IBM Research https://orcid.org/0000-0001-7737-4172

Victoria Dardov
Cedars Sinai

Alyssa Coyne
Johns Hopkins University https://orcid.org/0000-0002-3658-5325

Aaron Frank
Cedars Sinai

Andrea Matlock
Cedars Sinai
Rakhi Pandey
Cedars Sinai
Vineet Vibhav
Cedars Sinai
Leandro Lima
Gladstone Institute
Jie Wu
Univ California Irvine
Divya Ramamoorthy
MIT https://orcid.org/0000-0001-9438-0419
Ryan Lim
Univ California Irvine
Julia Kaye
Gladstone Institute
Jonathan Li
Massachusetts Institute of Technology
Terri Thompson
One Point Scientific
Emily Baxi
Johns Hopkins University
Answer ALS
Johns Hopkins University

Biological Sciences - Article

Keywords: Longitudinal Study, Smartphone-based App, Speech Patterns, Integrated Clinical and Biological Signatures

Posted Date: October 28th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-96858/v1

Version of Record: A version of this preprint was published at Nature Neuroscience on February 3rd, 2022. See the published version at https://doi.org/10.1038/s41593-021-01006-0.

Abstract

Answer ALS is a comprehensive multi-omics approach to ALS to ascertain, at a population level, the various clinical-molecular- biochemical subtypes of sporadic ALS. This national program enrolled 1046 ALS and ALS/FTD patients along with a cohort of 100 matched control patients followed longitudinally over at least one year. A smartphone-based app was employed to collect deep clinical data including fine motor activity, speech, breathing and linguistics/cognition. Analytics of the speech patterns revealed a strong correlation between clinical progression indices and speech. In parallel, blood-derived iPS motor neurons were generated from each patient and the cells underwent multi-omics analytics including whole genome sequencing, RNA transcriptomics, ATAC-Seq and proteome along with quality assurance standards. HIPPA compliant cloud data bases were employed to store all data. There are more than 6 billion clinical and molecular data points per patient generated in the program. The program was designed, and patient consented, to be open access to all clinical, biological and molecular data as well as public release of all generated iPS cell lines. A web portal is available to academics as well as commercial researchers. The ultimate intent of this 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. Overall, this community based clinical and science program provides for the identification of distinct reliably identifiable subgroups among the sporadic and familial patients and the great utility in iPS based approaches to disease pathophysiology and therapy discovery. Although the data is ALS centric, given the large number of both ALS and control data sets, it would also be enormously useful to others studying frontotemporal dementia, Alzheimer's, Parkinson's disease and others.

Introduction

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, novel 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 and Huntington's disease, however, have not seen such advances. Clinical trials in human subjects, often based on findings from non-human model systems, have repeatedly proven disappointing. Although there are likely many reasons for such failures (e.g. poor pharmakokinetics, 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. Because 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. In the case of familial ALS (fALS), the study of genetic models of ALS has led to significant progress in our understanding of disease mechanisms. But for the majority of ALS patients, 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 induced pluripotent stem cell (iPSC) technology and the unprecedented data and specimen collection efforts of Answer ALS, we can now take an iPSC-based approach to unraveling mechanisms that may cause or contribute to the heterogeneous clinical spectra of sALS, such as pattern and speed of spread and certain non-motor manifestations. Notably, multiple genes 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 to 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.

With the Answer ALS program (AALS), we take a step away from the previous focus on genetic rodent models of ALS by generating iPSCs from a large number of people with sALS and applying well-established molecular, biochemical, and imaging techniques to understand the heterogeneity of sALS. After ensuring that results were reproducible, we assembled comprehensive biological datasets from iPSC work and linked them to the longitdudinal clinical data. In contrast to smaller previous iPSC experiments, studies of iPSCs from a large population, like AALS, provide the first opportunity to explore biologically relevant subgroups of sALS. This program was designed with the core goals of providing large clinical and biological datasets in an open source-like application that affords researchers the proper tools to identify biological subgroups and an extensive collection of IPSC lines with which to test ALS therapies and hypotheses about ALS pathogenesis.

Program Outline And Process

Overall Design (Figure 1). The overall AALS program, from clinical enrollment, to app data collection, iPSC line generation, biological data generation, and data storage is outlined in Figure 1 (ClinicalTrials.gov: NCT02574390). Below, we explain in detail the individual elements of the program.

1. **Enrollment, Clinical Characterization and Sample Collection**

The clinical portions of AALS were coordinated through Massachusetts General Hospital and Johns Hopkins University. 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 Neurology Group at Texas (Extended Table 1 and Extended Figure 1). The clinics were chosen for their geographic distribution, expertise in ALS clinical research and ability to recruit participants rapidly based on prior Northeast ALS Consortium (NEALS) clinical research studies (https://www.neals.org).

The study was approved by local institutional review boards, and all participants provided written informed consent prior to undergoing any study procedures. Consent was uniform across all sites and included agreement to share data broadly for medical research, in accordance with the mission of the overall program (see Data Access below for specifics). Subjects with sALS, fALS and related motor neuron diseases, including those with primary lateral sclerosis, progressive bulbar palsy, and progressive muscular atrophy, were enrolled in AALS. Age matched control participants without ALS or a family history of ALS were also enrolled.

Upon enrollment, participants were assigned a NeuroGUID (global unique identifier; https://nctu.partners.org/neurobank), used to link participant data within AALS and across studies. Clinical data were entered and stored in a centralized, custom web based electronic data capture system, (NeuroBank). All people over 18 years of age diagnosed with ALS or related motor neuron diseases were eligible to join the study irrespective of disease severity. Control participants were recruited at the same clinics - many were spouses, partners, or caregivers.

Participants were monitored every three months for a year. When possible, the ALS Functional Rating Scale-Revised (ALSFRS-R) was conducted by telephone every three months for another year thereafter. Visits included collection of baseline descriptors followed by measures of ALS progression. Baseline descriptors included: demographics and vital signs, genetic and family history of MND, general medical history, CNS -lability 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 (DTR), Ashworth Spasticity Scale, Hand Held Dynamometry (HHD), ALS Functional Rating Scale – Revised (ALSFRS-R), and pulmonary slow vital capacity (SVC). (See ExtendedTables 2,3 and Supplementary data forms).

At each in-clinic visit, approximately 50–100 ml of blood was collected from each participant and processed according to methods outlined in the Supplemental Methods. In addition, at the first visit, whole blood was collected, processed (see Supplemental methods) and shipped to Cedar-Sinai for generation of primary peripheral blood mononuclear cell (PBMC)-derived induced pluripotent cell lines.

Answer ALS Smartphone Application

To increase the density of the clinical data collection, we deployed a custom smartphone app to a self-selected subset of participants. AALS app study participants were active participants in the main study,

over 18, and used a personal smartphone device regularly. This AALS-accompanying study was approved by local institutional review boards, and all participants provided written consent prior to undergoing any study procedures. Data were collected and stored briefly on the phone before upload via cell carrier data plan or WiFi to a HIPAA-compliant Box account. Access was controlled using a secure password.

The app has seven modules designed to gather information about upper limb motor function, respiration, bulbar function and cognition (See Supplementary Data and Methods). The app, designed for both iOS and Android platforms, was made available in both the Apple App and Google Play stores. During an AALS main study visit, the app was downloaded to the participant's personal smartphone from the app store, activated by study staff using a unique code to initiate data collection, and used to collect select baseline demographic data (e.g. handedness). The participant was then able to carry out study app activities.

The app captured detailed information about actively performed tasks. It prompted participants to perform one task daily, though unperformed tasks remained accessible for the entire week. All seven tasks were repeated weekly. Six modules measured arm function: finger tapping, finger tracing, and phone tilt tracing, each performed using the right and left hand separately (Figure 2a). The speech module (Figure 2c), consisted of three tasks, rotated weekly to reduce learning effect: 1) Single-Breath Count, in which participants were instructed to draw in a deep breath and count at a measured pace (a surrogate for forced vital capacity)[1] 2) Read Aloud Passage, in which participants read aloud one of four standardized passages from their screen, and 3) Picture Description, in which participants described one of three line-art illustration over 30-120 seconds.

We analyzed compliance over time (total tasks completed per week) to evaluate for app engagement. We extracted relevant features for each task. For arm function tasks, error metrics such as Hausdorff and dynamic time warping distances were calculated, and the number of points acquired by the device during the tracing task was used to estimate movement speed. For speech task analysis, we used standard acoustic features to assess motor speech degradation such as pitch variations, prosody features, vowel space, vowel quality, noise measurements, mel frequency cepstral coefficients (MFCCs), tremor features and others.

Picture description recordings were manually transcribed because automated speech-to-text engines were unable to reliably transcribe dysarthric speech. From the transcripts, we extracted linguistic features to evaluate word diversity and complexity of thought such as semantic similarity, dispersion, and frequency. More details of the methodology have been reported[2]. To evaluate the potential of the tasks to assess different clinical variables used to monitor ALS (e.g. ALSFRS-R, vital capacity, cognitive behavioral screen), the extracted features were entered into three machine learning algorithms (linear, ridge, lasso regression) and validated using 10-fold cross validation.

Return of Answer ALS Results

Participants with ALS were offered the opportunity to receive the results of their whole genome sequencing for five ALS genes (C9orf72, SOD1, FUS, TARDBP, and TBK1), as well 59 genes designated as medically actionable by the American College of Medical Genetics[3], as part of a sub-study, Return of Answer ALS Results (ROAR). ROAR participants completed a separate online consent after enrollment in the parent study. A separate variant interpretation pipeline was applied for the purposes of return of results. Clinical confirmation of each identified variant interpreted as Pathogenic or Likely Pathogenic and genetic counseling by a licensed genetic counselor was offered to all participants in this study.

2. **iPSC generation from PBMCs and motor neuron differentiation**

iPSC line generation

Blood from participants with motor neuron disease and controls was sent to a central iPSC generation lab (Cedars Sinai) by overnight service where PBMCs were isolated, logged and frozen until iPSC generation (see **Extended Tables 4 and 5** for all participant demographics). iPSCs were then generated by reprogramming the cryopreserved and non-expanded PBMCs using a method based on a non-integrating episome. Clones were isolated, expanded and maintained according to standard feeder-free protocols and characterized extensively as described in Extended Table 6. iPSC lines were generated from ~25 patients per month and stored frozen until they were differentiated (Figure 3A). As of October 2020 ~700 iPS cell lines from participants have been generated. PBMCs were used instead of fibroblasts to limit the potential for genetic defects and facilitate sampling from the large number of patients enrolled in our study. Overall, blood draws are less invasive and carry lower risk for patients than skin biopsies, which improved the overall risk-to-benefit ratio for the study. In addition, it was widely felt that patients would be less likely to consent to a skin biopsy than a blood collection.

Generation of motor neurons

The iPSCs were differentiated into motor neurons according to the direct iPSC-derived motor neuron (diMNs) protocol (Figure 3B, Extended Table 6, Supplemental Methods), which comprises three main stages. In stage 1, neural induction and hindbrain specification of iPSCs is achieved by dual inhibition of the SMAD and GSK3β pathways. During stage 2, specification of spinal motor neuron precursors is achieved by addition of Shh agonists and retinoic acid. Maturation of these precursors into neurons with more complex processes and neurites occurs during stage 3 with the addition of neurotrophins and Notch pathway antagonists. This protocol generates a mixed population of neurons consisting of ~75% (+/-9%) β_{III}-tubulin (TuJ1) and ~70% (+/-11%) NF-H positive cells, ~19% (+/-7%) Islet-1 and ~35% (+/-10%) Nkx6.1 positive spinal motor neurons, and ~18% (+/13%) S100B and ~38% (+/-17%) nestin positive progenitors 32 days after the onset of differentiation. As of October 2020, successful motor neuron differentiations from ~ 350 iPSC lines have been completed by the AALS program.

Program Quality Controls: Cell generation batch controls.

Reproducibility of disease signatures from iPSC-based experiments can be confounded not only by genetic differences between donors (diseased and healthy controls), but also by experimental variability in iPSC differentiation experiments that can be impacted by variations in differentiation efficiency, cellular composition, transcript and protein abundance. To detect and compensate for such confounders all differentiations were conducted in a single facility and included two key control groups of biological samples: batch differentiation controls (BDC), were differentiated with each batch from the same original line to assess inter-batch variablity of iPSC diffrentiation to diMNs and Batch Technical controls (BTC), consisting of a single differentiation of the same line was frozen, aliquoted and distributed with each batch to assess technical variability of the omics assay batch runs were performed as detailed in Supplemental Methods.

3. **Multi-omics data generation for each iPSC-derived motor neuron line**

At the end of the 32-day differentiation protocol, the spinal neurons were harvested for RNA-Seq, proteomics, or epigenome profiling as detailed below and in Supplemental Methods. Whole-genome sequencing was performed on PBMCs.

Whole-genome sequencing and analysis.

PBMCs were sent by each clinic to the New York Genome Center (NYGC) for DNA extraction and subsequent whole-genome sequencing (WGS) on the Illumina X10. Sequence data were processed on a NYGC automated pipeline. Paired-end 150-bp reads were aligned to the GRCh38 human reference using the Burrows-Wheeler Aligner (BWA-MEMv0.7.8) and processed using the GATK best-practices workflow, which includes the marking of duplicate reads with Picard tools (v1.83, http://picard.sourceforge.net), local realignment around indels, and base quality score recalibration (BQSR) via Genome Analysis Toolkit (GATK v3.4.0). [4] [5]

The variant calls from NYGC were assessed by examining the reads for alignment issues and spot-checking the BAM files for specific variants in IGV and were determined to be of good quality. The VCFs were converted into GVCFs, and we performed custom annotation and intersected a subset of the omics data (RNA-Seq, ATAC Cluster) with the WGS data. The annotation pipeline was customized to incorporate elements from ANNOVAR[6] and KGGseq[7], from which a report was generated that included the genotypes of all samples. The annotate genes and exonic variants that have clinical significance, we incorporated the Clinical Genomic Database (CGD)[8], the Online Mendelian Inheritance in Man (OMIM)[9], ClinVar[10] and genes listed in the American College of Medical Genetics and Genomics (ACMG)[11] database as well. Intervar, which is based upon the ACMG and AMP standards and guidelines for interpretation of variants, was also incorporated. This tool uses 18 criteria to ascribe clinical significance and classifies genes based on a five-tier system[12]. To flag ALS genes, we incorporated ALS gene lists and variants from ALSoD[13] (http://alsod.iop.kcl.ac.uk/), a highly curated list from Dr. John Landers, Dr. Matt Harms and ALS Association from the DisGeNet database[14]. For each variant, we also incorporated functional *in silico* predictions from nine programs, including databases such as SIFT[15], PolyPhen2[16], and Mutation

Taster[17] and those described in Li et al., 2013[18]. Additional databases were included that assess the variant tolerance of each gene using the RVIS[19] and the Gene Damage Index (GDI)[20] and are adding LoFTool[21]. For variants in genes that are highly expressed in the brain, we incorporated data from the Human Protein Atlas[22] (http://www.proteinatlas.org) and expression data from GTEx portal [23 24], (https://gtexportal.org/home/) for the cortex and spinal cord. Frequency information from three databases on all known variants from ExAC[25], the NHLBI Exome Sequencing Project (ESP)[26], and the 1000 Genomes Project[27].

A separate annotation pipeline was developed for variants that map to intergenic and regulatory regions. We report the variant as found next to the closest gene, and as either intronic, upstream or downstream (up to 4 KBs from the 5' and 3' UTRs). The annotation came from RegulomeDB, which annotates variants with known or predicted regulatory elements such as transcription factor binding sites (TFBS), eQTLs, validated functional SNPs and DNase sensitivity[28]. The source data comes from ENCODE [29 30] and GEO[31]. We also included other regulatory databases such as Target Scan, an algorithm that uses 14 features to predict and identify microRNA target sites within mRNAs[32] and miRBase[33-35].

RNA-Seq

Total RNA was isolated from the iPSC-derived neuronal lines from each AALS subject (control or ALS) using QIAshredders and RNeasy mini RNA prep kits (Qiagen). Samples with RIN >8 were used for subsequent library preparation and were entered into an electronic tracking system and processed at the University of California, Irvine, Genomics High-Throughput Facility. rRNAs were removed and libraries generated using TruSeq Stranded Total RNA library prep kit with Ribo-Zero (Qiagen). 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 100 Illumina (Novaseq 6000) paired end (PE) sequencing cycles to obtain over 50 million PE reads per sample. After sequencing, raw reads were subject to QC measures and reads with quality scores over 20 collected and analyzed. Reads were mapped to the GRCh38 reference genome, QCed, and gene expression and differential expression were quantified using Hisat2, featureCounts[36] and DESeq2[37]. Normalized and transformed count data were then used for exploratory analysis and differentially expressed (DE) genes (FDR <0.1) were analyzed with commercial and open-source pathway and network analysis tools, including Ingenuity Pathway Analysis (IPA), GSEA, GOrilla, Cytoscape, and other tools to identify transcriptional regulators, predict epigenomic changes, and determine potential effects on downstream pathways and cellular functions.

ATAC seq

We used the Assay for Transposase-Accessible Chromatin using Sequencing (ATAC-Seq) to assess chromatin accessibility and identify functional regulatory sites involved in driving transcriptional changes associated with ALS. ATAC-Seq detects open chromatin sites genome-wide and maps transcription factor binding events in global regulatory elements without needing prior information about which proteins are present. ATAC-seq sample prep, sequencing, and peak generation was carried out by Diagenode Inc as

previously described (Supplemental Methods). [38] After sequencing, chromatin accessibility signatures were generated for each sample individually using the peak-calling software MACS2 and then determined differentially open sites using DiffBind with DESeq2 (FDR<0.1). The sequencing quality 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-calling software MACS2 [39] and determined differentially open sites using DESeq2 (FDR<0.1). Peaks were assigned to unique genes using the default HOMER parameters, and gene ontology analysis was performed using GOrilla. [40]

Proteomics

Frozen diMNs were processed following Expedeon FASP protocol. Processed samples were subjected to acquisition on the SCIEX 6600. Samples were acquired in data-dependent acquisition (DDA) mode for library building and in data-independent acquisition (DIA) mode over 100 variable windows similar to previously described acquisition protocols [41,42]. DDA files were run through Trans Proteome Pipeline (TPP) using a human canonical FASTA file (Uniprot). A consensus peptide library with decoys was generated. Previously described DDA library build principles [43] were utilized to generate a cell-specific library, which allowed for more accuracy in matching DIA data to the DDA library during OpenSWATH, as indicated by higher d-scores in PyProphet. These data were also analyzed using commercial and open-source pathway and network analysis tools, including Ingenuity pathway analysis and GOrilla to identify upstream regulators and determine the cellular pathways affected.

Longitudinal single-cell imaging and analysis

Differentiated diMNs from a subset of the AALS iPSC lines were plated on 96-well plates for longitudinal single-cell imaging with robotic microscopy. Days after plating, cells were transfected with expression marker plasmids to visualize cell morphology and viability. After transfection, cells were automatically imaged with robotic microscopy once per day for 10–14 days. A fiducial mark on each 96-well plate was used to bring each plate back to its initial position at each imaging run, which allowed the system to collect images of the same microscope fields over the course of the experiment and to identify and track individual diMNs. Images of different microscope fields from the same well were stitched together into montages and montages assembled at different time points were organized into composite files in temporal order. A computational pipeline constructed within the open source program Galaxy was used to perform cell survival analysis and other morphological measurements. Independently, images from diMNs from ALS patients and healthy volunteers were analyzed with machine learning and deep learning methods in a relatively unbiased fashion to discover if they could be stratified or substratified to predict which diMNs were derived from ALS patients.

4. Data Storage and Data Integration/Analytics

Answer ALS was designed to be an “open source” program. All of the clinical data sets, the various omics results, including whole genome, proteome, transcriptome and epigenome along with the data integration

have been posted to a portal for data sharing and crowd sourcing (http://data.answerals.org; Extended Table 3). Data are available for download to all academic and commercial researchers. A required data use agreement provides assurance that users will not attempt to violate the GUID privacy, as well as share or sell the raw data without Answer ALS permissions. There are no intellectual property restrictions on the use of the data.

Web-based analytics. We have included online analytics for the many ALS researchers who will neither need nor want to download the full dataset. The current set of tools available at http://data.answerals.org/analyze allow users to select genes/pathways of interest and visualize them using braid maps, heat maps, volcano plots, bar charts or networks (Figure 5).

Results

AALS Clinical Demographics and Clinical Data Generation

Enrollment. All participants were enrolled at the eight geographically distributed clinics from Jan 2016 through June 2019 (Figure 4; Extended Figure1; ExtendedTable 1). Enrollment proceeded as planned at a rapid and regular pace (Figure 6). To ensure rapid enrollment, clinics were assigned a designated clinic coordinator to work full, half or quarter time, depending on historical rates of monthly enrollment from previous ALS clinical trials. Full clinics were assumed to enroll 5–10 participants/mo, half clinics 2.5–5 participants/mo and quarter clinics 1–3 participant/mo.

Population Demographics (Table 1 and Extended Table 4, 5). The enrolled participant population had clinical characteristics comparable to past large sporadic ALS population demographics, with a slightly higher number of male than female participants, a clinical site of onset predominantly limb rather than bulbar, and a mean age of disease onset of approximately 57. The mean delay in clinical diagnosis for ALS participants included in the study was 14.8 months. In agreement with prior population studies, a higher percentage of patients with rapid progression had bulbar onset disease.

ALS Progression: ALSFRS-R. Of the 1046 ALS participants enrolled, over 570 were seen in at least 3 follow-up visits, allowing us to generate disease progression curves based on the ALSFRS-R total assigned values. As shown in Figure 4B,C, there was a wide range of disease progression rates over the time period of observation, which ranged from 3 to 40 months. A subset of participants declined rapidly, as defined by a dip in ALSFRS-R slope by 1.8 or more points/month.

App-Based Voice Recordings – Motor and Motor Speech Analyses.

Compliance. Compliance for using the smartphone app was analyzed over 28 months from the beginning of the app rollout to a subset of study subjects. Data from 80 participants was analyzed. Only a modest decrease in compliance was observed with increased duration of use (see Figure 5A).

Smartphone App data. Features derived from the voice tasks (Single-Breath Count, Read Aloud Passage, and Free Speech; Extended Figure 2) each correlated highly with the bulbar subdomain of the ALSFRS-R

(Pearson R = 0.8; slope = 1.14, Pearson R = 0.89; slope = 0.98; Pearson R = 0.71; slope = 1.12 respectively). Features from the finger tracing showed modest individual correlations with the ALSFRS-R total score (Figure 5B, Extended Figure 2). Importantly, the combination of features from all of these tasks correlated very highly with the ALSFRS-R total score (Pearson R = 0.89; slope = 1.16; Figure 5. App 5C).

Features obtained from the single breath counting task correlated well with vital capacity (R=0.63). This task also predicted well the ALSFRS-R speech sub score (see Figure 5B); 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[2]. Semantic analysis of the picture description task was highly correlated with the ALS-CBS (R=0.72) and less correlated with the CNS lability scale (R=0.45).

These results demonstrate that the modules implemented to assess hand function and speech may be useful to quantify ALS function. 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.

Induced Pluripotent Cell Line Production

PBMC Processing. A total of 1,030 whole-blood samples were collected and sent to Cedars-Sinai for PBMC isolation and cryopreservation. Of the 1,030 samples, 32 were unusable due to issues with sample collection or shipment and 34 samples were redrawn. The average cell count was ~25 million PBMCs per sample with an average cell viability of 91%. In total, the iPSC Core at Cedars-Sinai has frozen 2579 vials of PBMCs from 964 unique participants, comprising of 860 ALS participants and 104 healthy controls.

iPSC Reprogramming. The iPSC Core at Cedars-Sinai Medical Center reprograms PBMCs using a non-integrating episomal plasmid method. As of October, 2020, the iPSC Core has initiated the reprogramming of 637 unique PBMC samples. Thus far, ~700 iPSC lines have been successfully reprogrammed and one clone line banked and characterized per donor. Out of the ~700 unique samples, only 18 lines (~3%) failed reprogramming. When reprogramming fails, a new attempt is made from a blood sample collected at a follow-up clinic visit (Figure 3A). An average of two additional iPSC clones per donor were banked at an early passage and reserved as backup. Each iPSC line was banked in an average of 50 vials from multiple passages, including 25 vials at the distribution bank around passage 20. In all, the AALS program has created ~30,000 iPSC vials from all the individual PBMCs reprogrammed thus far.

Generation of Multi-Omics data

Genomics

We analyzed 830 whole-genome sequences from AALS participants. Of these, 706 were ALS cases, 92 were controls without neurological disease, 16 were individuals diagnosed with a motor neuron disease that is not ALS, 5 had another neurological disorder, 5 were pre-familial ALS (pre-fALS), and 6 had undiagnosed clinical syndromes (**Extended Figure 3; Extended Table 4**). We evaluated this AALS cohort using NYGC's ancestry pipeline[44]. The majority of participants were Caucasian and had European descent (91.45%), the remainder had ancestry consistent with the Americas (1.69%), Africa (4.94%), East Asia (1.33%) and South Asia (0.6%) (**Extended Figure 3**). On average, each sample harbored a total of ~ 4.1 million variants and ~ 9,800 protein-altering variants, including SNPs, frameshift and non-frameshift deletions and insertions, and protein–truncating variants (**Table 2 and Extended Figure 3A**). This is similar to what has been previously reported[51]. The samples with African descent had a higher number of variants than other ethnic populations, as expected[45] (**Extended Figure 3B**).

We used principal component analysis (PCA)[46,47] to visualize the ancestry background of the AALS cohort and a set of 2504 samples from the 1000 genomes project with well-defined ancestry. We used a set of 10,000 randomly chosen autosomal SNPs (singletons and multiallelic SNPs were removed) that were present in both datasets, and removed correlated SNPs by LD-pruning. We implemented randomized PCA[48] using the Python library scikit-allel package[49] (**Extended Figure 4**). PC1 showed that African samples (green) clustered apart from the other populations, and PC2 that Asian samples (red/brown) were distinct from European samples (purple), with admixed American were located in between. The majority of the AALS samples were clustered with the European samples, although some were closer to the African group and a few clustered with the Asian group (**Extended Figure 4),** corroborating the NYGC ancestry results (**Extended Figure 3B**).

We first evaluated pathogenic or likely pathogenic variants reported in ClinVar (C-PLP) for all genes. We observed between 22 and 48 C-PLP variants per individual, with an average of ~ 34 variants per ALS case and control, similar to what has been reported for Caucasian individuals[45]. The number of rare (<1%) C-PLP variants was approximately 5.2 per ALS case and 5 per control (**Table 2 and Extended Table 1**). We also examined pathogenic variants called by Intervar Li, (PMC3326332) (I-PLP), and observed that a typical sample (from an ALS case or control) harbored approximately 3 I-PLP variants (**Table 2 and Extended Table 2**). The majority of I-P/LP variants called by Intervar were rare.

Lastly, we investigated the number of predicted damaging variants as called by *in silico* prediction tools, where 6 or more out of 9 algorithms predicted a variant to have a functional impact (IS-D). We observed between 82 and 195 IS-D variants per individual. ALS cases had an average of 112 IS-D variants per sample, and controls had 113 IS-D variants per sample (**Table 2 and Extended Table 3**).

Variants in ALS genes

There are 33 genes in which mutations have been associated with ALS[50,51], specifically: *ALS2*[52,53], *ANG*[54-56], *ANXA11*, *ATXN2*, *C21orf2*, *C9orf72*, *CAMTA1*, *CCNF*, *CHCHD10*, *DAO*, *DCTN1*, *FIG4*, *FUS*, *HNRNPA1*, *HNRNPA2B*, *KIF5*, *MATR3*, *MOBP*, *NEK1*, *OPTN*, *PFN1*, *SCFD1*, *SETX*, *SOD1*, *SQSTM1*, *TAF15*, *TARDBP*[57-59], *TBK1*, *TUBA4A*, *UBQLN2*, *UNC13A*, *VAPB* and *VCP*[60]**.** We refer to these as the "33-ALS" genes. Within the 830 samples, we observed 440 exonic variants in the 33-ALS genes that were less than 1% frequent (**Table 2 and Extended Table 4**). Both controls and ALS cases averaged 1.5 rare ALS variant per individual within the 33-ALS genes. 79% of these were single nucleotide polymorphisms (SNPs), 13% uncharacterized, ~1% splicing, ~1% non-frameshift deletion, 1% frameshift deletion, 1% frameshift insertion, 2% frameshift insertion, 2% non-frameshift insertion, 1% stop-gain (**Extended Table 4**).

We first evaluated how many pathogenic or likely pathogenic existed as reported in ClinVar (C-PLP) in the 33-ALS genes. We found that 12% of ALS cases harbored a C-PLP variant within one of the 33-ALS gene (**Extended Table 4 and 5).** All of these C-PLP variants were rare (<1% frequency within the population) except 2 found within the OPTN gene. For example, we observed 5 *SOD1* C-PLP variants (within 8 ALS patients), 2 *TDP43* C-PLP variants (within 2 ALS patients), and 1 C-PLP *FUS* variant in an ALS patient (**Extended Table 4 and 5**). C-PLP variants were also detected in individuals that did not show signs of ALS at the time of the clinic visit, and there were 11 C-PLP variants within control samples (within *ALS2*, *SETX*, *OPTN* and *PFN1*), 4 C-PLP variants in the Pre-fALS cohort (within *FIG4*, *OPTN* and *CHCHD10*), 3 C-PLP variants within individuals with other motor neuron disease (OMND) (within *SQSTM1*, *OPTN*, and *PFN1*) and 3 C-PLP variants in uncharacterized individuals (within *SQSTM1* and *SETX*) (**Extended Table 5**). In summary, rare C-PLP variants were observed in 3.11% (22 total) in ALS cases and 1% in controls (1 out of 92 samples).

The majority of samples that harbored a C-PLP variant harbored only one. However, there were a few examples where individuals sustained more than one pathogenic variant. We observed a C-PLP variant in *SQSTM1* and one in *OPTN* in one ALS case, 2 C-PLP variants in the *SETX* genes in an uncharacterized individual and a control, C-PLP variants in both *OPTN* and *CHCHD10* in a Pre-fALS individual, and 2 C-PLP variants, in *FIG4* and *CHCHD10*, in another Pre-fALS individual (**Extended Table 5**).

Intervar called significantly fewer P/LP variants (3.54%, 25 variants total) than Clinvar. However, Intervar called more *SOD1* P/LP variants than flagged by ClinVar (12 variants within 16 individuals with ALS; **Extended Table 4 and 6**). We observed no I-PLP variants in controls, preF-ALS, OMND, or unknown samples. Using a 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 11 genes—*ALS2*, *CCNF*, *CHCHD10*, *FUS*,

OPTN, *PFN1*, *SOD1*, *TARDBP*, *TBK1*, *UBQLN2*, *VAPB*, and *VCP*—has been curated and designated the H-PLP variants. We observed 24 H-PLP variants in our cohort of ALS cases (**Extended Table 4 and 7**). We observed one H-PLP variant in *PFN1* in a control, and one in an (Other Motor Neuron Disease) OMND individual.

We also investigated IS-D variants in the 33-ALS genes. 98 individuals harbored at least 1 IS-D variant in the 33-ALS genes, and 9 individuals harbored 2 IS-D variants in the 33 genes associated with ALS as described above [50,51] (**Extended Table 4 and 8**). In general, Intervar called a variant as pathogenic or likely pathogenic much less than Clinvar and hence appears to be more stringent in its pathogenic determination. However, its unclear if Intervar provides a more rigorous calling of *bona fide* ALS variants.

Expansions in C9orf72 and ATXN2

Using Expansion Hunter[61] to identify repeat expansions within WGS data, we found 601 expanded regions in the 830 samples. In total, 41 ALS patients and 4 Pre-fALS subjects harbored hexanucleotide expansions in *C9orf72* that were greater than 26 repeats (**Extended Figure 5 and Extended Table 9.** We also observed 35 ALS patients, 4 controls and one uncharacterized individual harboring CAG triplet repeat expansions in *ATXN2* greater than 26 repeats (**Extended Figure 5 and Extended Table 10**). For carriers of expansions in both *ATXN2* and *C9orf72* simultaneously, we found no correlation between age of ALS onset and expansion size (**Extended Figure 6 and Extended Tables 9 and 19**). Interestingly, 4 ALS patients harbored both *C9orf72* and *ATXN2* expansions greater than 27 repeats (**Extended Table 10**). We also observed that one Pre-fALS patient that had a *C9orf72* expansion also harbored two C-PLP variants, one in *FIG4* and one in *CHCD10*. In addition, one individual who harbored an *ATXN2* expansion of 28 repeats, also harbored a C-PLP variant in *TDP-43*. Another individual with ALS that harbored an ATXN2 expansion of 26 had a I-PLP and H-PLP variant in *SOD1* (V69A).

ACMG Genes

Pathogenic or likely pathogenic variants in 59 genes are currently considered medically actionable by the ACMG, due to the potential for medical intervention to modify morbidity and mortality in carriers of such variants.[3]. Within the 830 samples, we identified 73 C-PLP variants within 32 ACMG genes (**Extended Table 11**). 50.4% of individuals 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. 66 of these variants found within 110 individuals were rare (<1%) (**Extended Table 11**). We also found 42 I-PLP variants within ACMG genes within 51 individuals, all of which were rare (**Extended Table 12**). Participants were offered to receive the results of these medically actionable genes through the RoAR substudy.

Transcriptomics

For each of the omics assays, vials from an identical pool of differentiated motor neurons were processed to ensure comparability (Extended Figure 7), including BDCs and BTCs from the control 2AE8 line. For the RNA-Seq data, the initial set of 102 samples were processed and passed all quality controls (QC) metrics including RNA integrity (**Figure 6A**), library, and sequencing QC metrics. After read mapping and expression quantification, we evaluated data composition and quality. To assess data quality and technical batch effects, pairwise gene level SERE scores (Simple Error Rate Estimate, 0 = identical samples) were generated for the batch differentiation controls (BDCs), batch technical controls (BTCs), and all other samples (**Figure 6D**). These data show low SERE scores in the BTC and BDC controls, relative to all other samples, indicating minimal to no technical confounders and low batch effects between differentiations. The highest SERE values were found between different individuals. A heatmap of SERE scores between all samples with hierarchical clustering (**Extended Figure 7**) shows tht while BTCs form their own cluster, the rest of the samples fall info multiple small clusters with no clear relation to their disease status.

Annotation of quantified reads revealed various RNA species that were captured during the sequencing with protein coding RNAs accounting for the majority of all RNAs as well as lncRNAs **(Figure 7A**). A low proportion of reads mapped to small RNAs and a very minimal portion to rRNAs, which were depleted during library preparation and act as a technical quality assessment. The use of total RNAseq and deeper sequencing allows for differential alternative splicing and exon usage analyses as well as circular RNA and cryptic exon analyses. We chose to assess the ability of our cell model and RNAseq methods to capture known alt-splicing that has been previously reported in ALS samples. **Figure 7B** shows a sashimi plot of *POLDIP3,* a gene that has been previously shown to have alterations in splicing associated with loss of TDP-43 in ALS tissues [62]. Similar to previously reported data, splice variant 2 (with exon 3 skipped) is enriched in our ALS samples. These data indicate that both gene expression differences and RNA splicing differences could be captured by our iPSC model and through our RNAseq methods. These data can be explored for additional novel alterations in ALS and potential associations with ALS subtype and clinical data, and with other omics data that are being captured from these samples.

Proteomics

Proteomics data was generated for an initial 66 samples which were processed as a single batch and run sequentially on the MS instrument in blocks of 14. Each block of samples was comprised of case, control, BDC (differential batch control) samples and HEK293 cell control samples (the latter processed on the 96-well digestion plate for use as a sample plate digestion control). The numbers of proteins and peptides quantified for all 66 samples were very consistent (**Figure 6C**), a QC measure which indicates accurate processing consistency and the stability of the intra-batch data acquisitions on the instrument across all samples. The % coefficient of variation (CV) for the proteins quantified were calculated for the

BTC and BDC samples (**Figure 6F)**. 80% of the proteins identified in the technical replicates of BTC and BDC samples across all MS batches have % CV less than or equal to 25%, indicating proteomics data acquisitions between batches were highly reproducible. Individual samples are normalized to the total MS2 spectra intensity across the chromatographic profile of eluting peptides to smooth any inconsistencies in sample loading onto the MS instrument thereby eliminating systemic variation in signal intensities (**Figure 7E**). Finally, in a correlation plot of the protein level data for all 66 samples, we find BTCs and BDCs (both originating from 2AE8 CTR cell line) cluster tightly (**Extended Fig 7C**) indicating minimal drift between the MS batches. In total, greater than 25,000 peptides corresponding to more than 3,600 proteins per sample were quantified. Although cases and control iMNs clusters are interspersed, indicating their overall similarity, these iMN models have significant individual protein level differences and we selected four representative proteins ANXA2, PCKGM, ECH1, SYPL1) that show significant ($p \leq 0.05$) differences, based on what is seen in the differential analysis-based evidence (**Figure 6F**).

Epigenomics

ATAC-seq data quality was determined according to ENCODE [63]. The distribution of fragment sizes across all samples revealed a clear nucleosome-free region and regular peaks corresponding to n-nucleosomal fractions (**Figure 6B**). Mitochondrial DNA contamination was low (mtDNA fraction: 0.07 ± 0.01), and the fraction of reads in called peak regions (FRiP) indicated a good signal-to-noise ratio for our library (**Figure 7C**, mean ± SD = 0.160 ± 0.048), with no difference in quality score between ALS and control samples ($p = 0.32$).As expected, replicates from our batch control line were highly correlated with each other, with batch technical controls (BTC) having an even smaller variation in correlation values compared to batch differentiation controls (BDC) (**Figure 6E**).

Next, we generated a consensus set of peaks present in >10% of samples using DiffBind (**Extended Figure 8**) and characterized transcription factor motif enrichment within these peaks using HOMER[40]. Consistent with our expected cell composition, we observed an overrepresentation of transcription factors implicated in neuronal differentiation, such as Pdx1, Cux2, and the Lhx family (**Figure 7D**). 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 Figure 8**). Subjects did not cluster by disease status, presence of C9 mutation, sex, or by processing batch.

Data Dissemination: Data Portal

The Answer ALS Data Portal (http://data.answerals.org/; Extended Table 3) was designed to provide information about the various types of biological data generated by the AALS partners and to allow easy visualization/access to the metadata, data and biosamples released. To fulfill the first goal, the portal provides an overview of the data release notes, assays, data level descriptions and links to sites for viewing cell lines/biosamples associated with the program. To fulfill the second goal, the website provides users a way to browse all available metadata (using filter and text search functions), download

all data and metadata or a filtered subset and find iPSC lines that can be ordered from Cedars-Sinai Biomanufacturing Center. Additional details regarding the portal can be found in Supplemental Methods

Users interested in downloading datasets are required to submit an online form, acknowledge data use parameters and return a signed Data Use Agreement (DUA). These measures serve to protect our enrolled participants' privacy in compliance with HIPAA. In addition, results generated using AALS have the possiblity of being shared for collaborative and open science purposes.

Discussion

The pathogenesis of sporadic ALS remains a mystery and few comprehensive data collections, on a population scale, exist to truly inform researchers as to 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 regarding disease management[64,65].

The core goal of Answer ALS was to provide a comprehensive set of tools including deeply phenotyped longitudinal clinical data and biological tools such as iPSC lines and a multi-omics platform consisting whole genome, cell-enriched proteome, transcriptome and epigenome, in order to uncover underlying biological subgroups.

Our reagent collection includes individual iPSC lines from approximately 700 sproadic ALS and control participants (soon to reach >1100), the iPSC-derived spinal neurons form each participant, their longitudinal clinical data (collected over one year), sequentially amassed fluid biospecimens (blood and CSF) and the early multi-omics data generated from each participant's blood (whole genome) as well as from their "spinal cord biopsy equivalent"- diMN 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 and coupled with the iPSC lines from these participants.

This population and its dataset were never envisioned to enable the identification of new ALS genes: a cohort of 1000 ALS participants does not amount to a large enough database for new gene identifications, although sharing the whole genome sequences from this data set has aided in the identification of a new ALS gene, Kif5A[66]. In fact, the estimated 6+ billion data points generated from

each participant, combining the longitudinal clinical demographic and observational data, the longitudinal smartphone app data (motor activity, speech, breathing, cognition), and the aggregate multi-omics data (whole genome, epigenome, proteome, transcriptome), represent an exceptionally large set of data per participant. Furthermore, the core omics data reflect the human cells affected in individual ALS paricipants—spinal neurons—and act as an organ- or tissue-specific biopsy. When this combined longitudinal and multidimensional clinical and biological data is analyzed by integrative methods, such as artifical intelligence, clinical and biological subgroups might emerge, potentially assigning a unique risk or modifier gene or a unique molecular pathway to a specific patient subgroup, which might one day enable patient-specific interventions, or serve as drug target engagement marker or subgroup biomarker.

The other strong research advantage to such a dataset and living tools is the immediate ability to test for potentially ALS-relevant pathogenic pathways using the participant's own iPSCs/iPSC-derived spinal neurons to test drugs for candidate pathogenic pathways and, importantly, to develop CNS biomarkers from the iPSCs and validate drug target engagement. Libraries of iPSC lines derived from participants with neurological diseases, including Alzheimer's disease and frontotemporal dementia have been growing in the last several years and represent valuable tool to truly examine specific disease pathways[67,68]. Most of these iPSC libraries are relatively small, including our orginal library of 22 familial ALS iPS cell lines[69], with a few selected lines for each disease mutation and when appropriate, isogenic controls. None are representative of the far more common sporadic forms of the disease. Furthermore, most do not provide deep longitudinal clinical and extensive multi-omics data.

The overall clinical demographics and population genomics in the Answer ALS program reflect accurately the ALS subject population described in prior studies. This observation validates the Answer ALS iPSC 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 iPSC lines were generated to be freely accessible to all researchers, academic and commercial, free of restrictions other than standard HIPAA compliance rules. A web portal for downloading filtered datasets, e.g. proteome, whole genome, etc. has been set up with minimal but appropriate requirements for data access (Extended Table 3). The iPSC 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 longitudinally collected from patients (e.g. plasma) are also available (Extended Table 3).

Longitudinal Smartphone App utilization:

The preliminary results from our AALS app demonstrate that the modules implemented to assess limb-function and speech may be useful to identify early bulbar 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 ALS functional rating scale, now commonly used as 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, often a prominent secondary outcome measure in clinical trials. This test typically requires in-clinic testing, which limits enrollment or followup data collection in clinical trials. The application of this app test alone could greatly enhance patient participation in nationwide clinical trials—especially in those areas where travel to testing center is challenging. Overall, we observe that quantitative motor speech analysis holds tremendous promise in both identifying changes not only limited 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 data sets available for muti-domain biological and clinical correlations for future users.

Declarations

Acknowledgements: Program support provided by: 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 Inc., 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, Mike Armstrong, Bruce Edwards, Pape Adams, Muscular Dystrophy Association, Les Turner ALS Center, PGA Tour, Mike Armstrong, Lipp Foundation. The authors thank the following for overall AALS program guidance: Lucie Bruijn, Jay Fishman, Ed Rapp, Peter Warlick, Clare Durrett, Peter Foss, Leandro P. Rizzuto, Denis Rizzuto, Steve Gleason, Paul Varisco, Randy Fishman, Beverly Goulet and Margaret Sutherland.

Data Availablity. All data supporting the findings of this study are available within the paper and its supplementary information files and web portals listed in Extended Table 3.

Answer ALS Consortium:

Project Leadership: Emily G. Baxi[1, 2], Terri Thompson[3], Steven Finkbeiner[4], Ernest Fraenkel[5], Dhruv Sareen[6,7], James Berry[8], Nicholas Maragakis[2], Jennifer E. Van Eyk[9], Leslie M. Thompson[10,11,12,13], Merit E. Cudkowicz[8], Clive N. Svendsen[6,7], Jeffrey D. Rothstein[1, 2]*

Project Management: Emily G. Baxi[1, 2], Terri G. Thompson[3], Barry Landin[14], Loren Ornelas[6], Elizabeth Mosmiller[2], Sara Thrower[7], S. Michelle Farr[15]

iPSC Production: Loren Ornelas[6], Lindsey Panther[6], Emilda Gomez[6], Erick Galvez[6], Daniel Perez[6], Imara Meepe[6], Dhruv Sareen[6, 7]

iPSC Differentiation and Distribution: Aaron Frank[6], Susan Lei[6], Berhan Mandefro[7], Hannah Trost[7], Louis Pinedo[6], Maria G. Banuelos[7], Dhruv Sareen[6,7] and Clive N. Svendsen[6,7]

iPSC Differentiation Analysis: Aaron Frank[6], Susan Lei[6], Chunyan Liu[6], Ruby Moran[6], Veronica Garcia[7], Michael Workman[7], Richie Ho[7], Dhruv Sareen[6,7], Clive N. Svendsen[6,7]

Whole Genome Analysis and Genetics: Julia A. Kaye[4], Leandro Lima[4], Stacia Wyman[4], Jennifer Roggenbuck[16], Matthew Harms[17], Steven Finkbeiner[4]

Transcriptomics: Ryan G. Lim[10], Jie Wu[11], Jennifer Stocksdale[13], Ricardo Miramontes[10], Keona Wang[13], Leslie M. Thompson[10,11,12,13]

Proteomics: Andrea Matlock[9], Victoria Dardov[9], Vidya Venkatraman[9], Ronald Holewenski[9], Niveda Sundararaman[9], Rakhi Pandey[9], Danica-Mae Manalo[9], Vineet Vaibhav[9], Jennifer E. Van Eyk[9]

Epigenomics: Aneesh Donde[5], Nhan Huynh[5], Miriam Adam[5], Brook T Wassie[5], Ernest Fraenkel[5]

Cell Imaging and Phenotyping: Ashkan Javaherian[4], Jeanette Osterloh[4], Jaslin Karla[4], Krishna Raja[4], Julia A. Kaye[4], Steven Finkbeiner[4]

Cell-based Studies: Alyssa N. Coyne[1, 2], Lindsey Hayes[2], Jeffrey D. Rothstein[1, 2]

Integrative Analysis and Computational Modeling: Jonathan Li[5], Divya Ramamoorthy[5], Ryan Lim[10], Jenny Wu[11], Julia Kaye[4], Karen Sachs[5], Alex Lenail[5], Natasha Leanna Patel-Murray[5], Steven Finkbeiner[4], Leslie M Thompson[10,11,12,13], Ernest Fraenkel[5]

Clinical Study Team: Elizabeth Mosmiller[2], Sara Thrower[8], Aianna Cerezo[2], Sarah Luppino[8], Alanna Farrar[8], Lindsay Pothier[8], Carolyn Prina[17], Todd Morgan[19], Arish Jamil[20], Sarah Heintzman[17], Jennifer Jockel-Balsarotti[21], Elizabeth Karanja[21], Jesse Markway[21], Molly McCallum[21], Ben Joslin[22], Deniz Alibazoglu[22], Stephen Kolb[17], Senda Ajroud-Driss[22], Robert Baloh[7], Daragh Heitzman[19], Tim Miller[21], Jonathan D. Glass[20], Jeffrey D. Rothstein[1,2], James, Berry[8], Nicholas Maragakis[2]

Clinical Data Management: Divya Ramamoorthy[5], Hong Yu[8], Ervin Sinani[8], Prasha Vigneswaran[8], Alex Sherman[8]

Smartphone App Development:, Omar Ohmad[2], Promit Roy[2], Jay Beavers[23], Emily G. Baxi[1, 2], Jeffrey D. Rothstein[1,2], John Krakauer[2]

Smartphone App Data Analytics: Raquel Norel[14], Carla Agurto[14], Guillermo Cecchi[14]

Web Portal Development: Terri Thompson[3], Barry Landin[15], Alex Lenail[5], Mary Bellard[23], Yogindra Raghav[5], Karen Sachs[5], Emily G. Baxi[1, 2], Tobias Ehrenberger[5,] Elizabeth Bruce[23], Ernest Fraenkel[5]

Author Affiliations:

[1]Brain Science Institute, [2]Department of Neurology, Johns Hopkins University School of Medicine, Baltimore MD 21205

[3]On Point Scientific Inc. San Diego, CA 92130

[4] Center for Systems and Therapeutics and the Taube/Koret Center for Neurodegenerative Disease, Gladstone Institutes and the Departments of Neurology and Physiology, University of California, San Francisco, San Francisco, CA 94158

[5]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139

[6]Cedars-Sinai Biomanufacturing Center, Cedars-Sinai Medical Center, Los Angeles, CA 90048

[7]The Board of Governors Regenerative Medicine Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90048

[87]Department of Neurology, Healey Center, Massachusetts General Hospital, Harvard Medical School, Boston MA 02114

[9]Advanced Clinical Biosystems Research Institute, The Barbra Streisand Heart Center, The Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90048

[10]UCI MIND; [11]Department of Biological Chemistry [12]Department of Neurobiology and Behavior, [13]Department of Psychiatry and Human Behavior and Sue and Bill Gross Stem Cell Center; University of California, Irvine, CA 92697

[14]Computational Biology Center, IBM T.J. Watson Research Center, Yorktown Heights, NY 10598

[15]Technome LLC. Herndon, VA 20171

[16]Zofia Consulting, Reston, VA 20190

[17]Department of Neurology and Genetics, Ohio State University Wexner Medical Center, Columbus, OH 43210

[18]Departmenr of Neurology, Columbia University, New York, NY 10032, USA

[19]Texas Neurology, Dallas, TX 75214

[20]Department of Neurology, Emory University, Atlanta, GA 30329

[21]Department of Neurology, Washington University, St. Louis, MO 63130

[22]Department of Neurology, Northwestern University, Chicago IL 60611

[23]Microsoft University Relations, Microsoft Corporation, One Microsoft Way, Redmond, WA 98052

Contributions: Individual contribution to each of the consortia operations are detailed above. JDR and CNS conceived the program. JDR, LMT, EF, SF, MC, JB, NM, JEVE, CNS, DS designed the overall program and oversaw all resource development. EGB and TGT oversaw overall program operations. JDR, EGB, LMT, EF, SF, MC, JB NM, JEVE, CNV, DS., JAK, JR, RN, JB, NM SK, and DR wrote the manuscript with input and edits from all the authors.

Corresponding Author:

Correspondance to: Jeffrey D. Rothstein MD, PhD, jrothstein@jhmi.edu

References

1 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, doi:10.1002/mus.24929 (2016) PMC4715713

2 Agurto, C. *et al.* Analyzing progression of motor and speech impairment in ALS. *Annu Int Conf IEEE Eng Med Biol Soc* **2019**, 6097-6102, doi:10.1109/EMBC.2019.8857300 (2019)

3 Kalia, 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. *Genetics in medicine : official journal of the American College of Medical Genetics* **19**, 249-255, doi:10.1038/gim.2016.190 (2017)

4 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010) PMC2928508

5 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498, doi:10.1038/ng.806 (2011) PMC3083463

6 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164, doi:10.1093/nar/gkq603 (2010) 2938201

7 Li, M. X., Gui, H. S., Kwan, J. S., Bao, S. Y. & Sham, P. C. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic acids research* **40**, e53, doi:10.1093/nar/gkr1257 (2012) 3326332

8 Solomon, B. D., Nguyen, A. D., Bear, K. A. & Wolfsberg, T. G. Clinical genomic database. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 9851-9855, doi:10.1073/pnas.1302575110 (2013) 3683745

9 Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic acids research* **43**, D789-798, doi:10.1093/nar/gku1205 (2015) 4383985

10 Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic acids research* **44**, D862-868, doi:10.1093/nar/gkv1222 (2016) 4702865

11 Green, R. C. *et al.* ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in medicine : official journal of the American College of Medical Genetics* **15**, 565-574, doi:10.1038/gim.2013.73 (2013) 3727274

12 Farrer, L. A. *et al.* Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *Jama* **278**, 1349-1356 (1997)

13 Abel, O. *et al.* Development of a Smartphone App for a Genetics Website: The Amyotrophic Lateral Sclerosis Online Genetics Database (ALSoD). *JMIR mHealth and uHealth* **1**, e18, doi:10.2196/mhealth.2706 (2013) 4114449

14 Pinero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research* **45**, D833-D839, doi:10.1093/nar/gkw943 (2017) 5210640

15 Sim, N. L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic acids research* **40**, W452-457, doi:10.1093/nar/gks539 (2012) 3394338

16 Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome research* **19**, 1553-1561, doi:10.1101/gr.092619.109 (2009) 2752137

17 Schwarz, J. M., Rodelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nature methods* **7**, 575-576, doi:10.1038/nmeth0810-575 (2010)

18 Li, M. X. *et al.* Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS genetics* **9**, e1003143, doi:10.1371/journal.pgen.1003143 (2013) 3547823

19 Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS genetics* **9**, e1003709, doi:10.1371/journal.pgen.1003709 (2013) 3749936

20 Itan, Y. *et al.* The human gene damage index as a gene-level approach to prioritizing exome variants. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 13615-13620, doi:10.1073/pnas.1518646112 (2015) 4640721

21 Fadista, J., Oskolkov, N., Hansson, O. & Groop, L. LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals. *Bioinformatics* **33**, 471-474, doi:10.1093/bioinformatics/btv602 (2017)

22 Uhlen, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419, doi:10.1126/science.1260419 (2015)

23 Consortium, G. T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-660, doi:10.1126/science.1262110 (2015) PMC4547484

24 Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585, doi:10.1038/ng.2653 (2013) PMC4010069

25 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016) 5018207

26 Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-69, doi:10.1126/science.1219240 (2012) 3708544

27 Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015) 4750478

28 Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome research* **22**, 1790-1797, doi:10.1101/gr.137323.112 (2012) 3431494

29 Consortium, E. P. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636-640, doi:10.1126/science.1105136 (2004)

30 Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100, doi:10.1038/nature11245 (2012) PMC4154057

31 Barrett, T. *et al.* NCBI GEO: archive for high-throughput functional genomic data. *Nucleic acids research* **37**, D885-890, doi:10.1093/nar/gkn764 (2009) 2686538

32 Agarwal, V., Bell, G. W., Nam, J. W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**, doi:10.7554/eLife.05005 (2015) 4532895

33 Griffiths-Jones, S. The microRNA Registry. *Nucleic acids research* **32**, D109-111, doi:10.1093/nar/gkh023 (2004) 308757

34 Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A. & Enright, A. J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic acids research* **34**, D140-144, doi:10.1093/nar/gkj112 (2006) 1347474

35 Griffiths-Jones, S., Saini, H. K., van Dongen, S. & Enright, A. J. miRBase: tools for microRNA genomics. *Nucleic acids research* **36**, D154-158, doi:10.1093/nar/gkm952 (2008) 2238936

36 Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169, doi:10.1093/bioinformatics/btu638 (2015) 4287950

37 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014) 4302049

38 Milani, P. *et al.* Cell freezing protocol suitable for ATAC-Seq on motor neurons derived from human induced pluripotent stem cells. *Sci Rep* **6**, 25474, doi:10.1038/srep25474 (2016) PMC4857123

39 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008) PMC2592715

40 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010) PMC2898526

41 Holewinski, R. J., Parker, S. J., Matlock, A. D., Venkatraman, V. & Van Eyk, J. E. Methods for SWATH: Data Independent Acquisition on TripleTOF Mass Spectrometers. *Methods Mol Biol* **1410**, 265-279, doi:10.1007/978-1-4939-3524-6_16 (2016)

42 Kirk, J. A. *et al.* Pacemaker-induced transient asynchrony suppresses heart failure progression. *Science translational medicine* **7**, 319ra207, doi:10.1126/scitranslmed.aad2899 (2015) 4858435

43 Parker, S. J., Venkatraman, V. & Van Eyk, J. E. Effect of peptide assay library size and composition in targeted data-independent acquisition-MS analyses. *Proteomics* **16**, 2221-2237, doi:10.1002/pmic.201600007 (2016)

44 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research* **19**, 1655-1664, doi:10.1101/gr.094052.109 (2009) PMC2752134

45 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015) PMC4750478

46 Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS genetics* **2**, e190, doi:10.1371/journal.pgen.0020190 (2006) PMC1713260

47 McVean, G. A genealogical interpretation of principal components analysis. *PLoS genetics* **5**, e1000686, doi:10.1371/journal.pgen.1000686 (2009) PMC2757795

48 Galinsky, K. J. *et al.* Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *Am J Hum Genet* **98**, 456-472, doi:10.1016/j.ajhg.2015.12.022 (2016) PMC4827102

49 Fabian Pedregosa, G. V., Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* doi: arXiv:1201.0490 (2012)

50 Alsultan, A. A., Waller, R., Heath, P. R. & Kirby, J. The genetics of amyotrophic lateral sclerosis: Current insights. *Degener. Neurol. Neuromuscul. Dis.* **6**, 49–64, doi:10.2147/DNND.S84956 (2016) PMC6053097

51 Mejzini, R. *et al.* ALS Genetics, Mechanisms, and Therapeutics: Where Are We Now? *Front Neurosci* **13**, 1310, doi:10.3389/fnins.2019.01310 (2019) PMC6909825

52 Sato, K. *et al.* Altered oligomeric states in pathogenic ALS2 variants associated with juvenile motor neuron diseases cause loss of ALS2-mediated endosomal function. *J Biol Chem* **293**, 17135-17153, doi:10.1074/jbc.RA118.003849 (2018) PMC6222102

53 Hadano, S. *et al.* A gene encoding a putative GTPase regulator is mutated in familial amyotrophic lateral sclerosis 2. *Nat Genet* **29**, 166-173, doi:10.1038/ng1001-166 (2001)

54 Greenway, M. J. *et al.* ANG mutations segregate with familial and 'sporadic' amyotrophic lateral sclerosis. *Nat Genet* **38**, 411-413, doi:10.1038/ng1742 (2006)

55 Wu, D. *et al.* Angiogenin loss-of-function mutations in amyotrophic lateral sclerosis. *Ann Neurol* **62**, 609-617, doi:10.1002/ana.21221 (2007) PMC2776820

56 Bradshaw, W. J. *et al.* Structural insights into human angiogenin variants implicated in Parkinson's disease and Amyotrophic Lateral Sclerosis. *Sci Rep* **7**, 41996, doi:10.1038/srep41996 (2017) PMC5296752

57 Rutherford, N. J. *et al.* Novel mutations in TARDBP (TDP-43) in patients with familial amyotrophic lateral sclerosis. *PLoS genetics* **4**, e1000193, doi:10.1371/journal.pgen.1000193 (2008) PMC2527686

58 Sreedharan, J. *et al.* TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis. *Science* **319**, 1668-1672, doi:10.1126/science.1154584 (2008)

59 Mackenzie, I. R. & Rademakers, R. The role of transactive response DNA-binding protein-43 in amyotrophic lateral sclerosis and frontotemporal dementia. *Curr Opin Neurol* **21**, 693-700, doi:10.1097/WCO.0b013e3283168d1d (2008) PMC2869081

60 Koppers, M. *et al.* VCP mutations in familial and sporadic amyotrophic lateral sclerosis. *Neurobiol Aging* **33**, 837 e837-813, doi:10.1016/j.neurobiolaging.2011.10.006 (2012)

61 Dolzhenko, E. *et al.* Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome research* **27**, 1895-1903, doi:10.1101/gr.225672.117 (2017) PMC5668946

62 Shiga, A. *et al.* Alteration of POLDIP3 splicing associated with loss of function of TDP-43 in tissues affected with ALS. *PLoS One* **7**, e43120, doi:10.1371/journal.pone.0043120 (2012) PMC3416794

63 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012) PMC3439153

64 Zhang, H. *et al.* Subgroup analysis reveals molecular heterogeneity and provides potential precise treatment for pancreatic cancers. *Onco Targets Ther* **11**, 5811-5819, doi:10.2147/OTT.S163139 (2018) PMC6140745

65 Ashley, E. A. Towards precision medicine. *Nat Rev Genet* **17**, 507-522, doi:10.1038/nrg.2016.86 (2016)

66 Nicolas, A. *et al.* Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. *Neuron* **97**, 1268-1283 e1266, doi:10.1016/j.neuron.2018.02.027 (2018) PMC5867896

67 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, doi:10.1016/j.neuron.2019.07.010 (2019)

68 Karch, C. M. *et al.* A Comprehensive Resource for Induced Pluripotent Stem Cells from Patients with Primary Tauopathies. *Stem Cell Reports* **13**, 939-955, doi:10.1016/j.stemcr.2019.09.006 (2019) PMC6895712

69 Li, Y. *et al.* A comprehensive library of familial human amyotrophic lateral sclerosis induced pluripotent stem cells. *PLoS One* **10**, e0118266 (2015)

Tables

Table 1. Answer ALS basic clinical demographics

			Cohort				
Variable	Level	Overall	ALS	Asymptomatic_ALS	Healthy_Control	Non_ALS_MND	Overall
Participants	N	100.0% (1047)	82.2% (861)	1.1% (12)	10.3% (108)	6.3% (66)	100.0% (200)
Sex	Female	40.6% (423)	37.4% (320)	58.3% (7)	66.4% (71)	37.9% (25)	38.0% (76)
	Male	59.4% (618)	62.6% (536)	41.7% (5)	33.6% (36)	62.1% (41)	62.0% (124)
	[missing]	.% (6)	.% (5)	.% (0)	.% (1)	.% (0)	N/A
Race	American Indian	0.2% (2)	0.1% (1)	0.0% (0)	1.0% (1)	0.0% (0)	N/A
	Asian	2.0% (21)	1.5% (13)	0.0% (0)	5.7% (6)	3.0% (2)	0.5% (1)
	Black	4.8% (49)	5.0% (42)	0.0% (0)	4.8% (5)	3.0% (2)	4.5% (9)
	Pacific Islander	0.1% (1)	0.1% (1)	0.0% (0)	0.0% (0)	0.0% (0)	0.5% (1)
	White	92.9% (956)	93.3% (789)	100.0% (12)	88.6% (93)	93.9% (62)	94.4% (187)
	[missing]	.% (18)	.% (15)	.% (0)	.% (3)	.% (0)	.% (2)
Ethnicity	Hispanic or Latino	4.8% (50)	5.3% (45)	0.0% (0)	2.8% (3)	3.1% (2)	4.0% (8)
	Not Hispanic or Latino	95.2% (989)	94.7% (810)	100.0% (12)	97.2% (104)	96.9% (63)	96.0% (191)
	[missing]	.% (8)	.% (6)	.% (0)	.% (1)	.% (1)	.% (1)
Age at baseline (yrs)	Years (mean(SD))	58.9±11.6 (20.0,91.0)	59.3±11.1 (24.0,91.0)	48.3±10.3 (33.0,62.0)	55.0±14.1 (20.0,82.0)	61.9±12.0 (26.0,85.0)	60.6±10.4 (29.0,85.0)
Time between symptom onset and diagnosis	Months (mean(SD))	15.9±20.4 (-5.7,286)	14.8±16.8 (-5.7,185)	N/A	N/A	40.8±52.6 (0.1,286)	16.5±21.0 (-5.7,157)
Time between symptom onset and study enrollment	Months (mean(SD))	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)	33.9±48.9 (0.6,458)
ALSFRS-R at first ALSFRS-R visit	mean(SD)	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)	32.9±8.20 (10.0,47.0)
ALSFRS-R slope		-.73±0.87 (-5.1,1.4)	-.77±0.88 (-5.1,1.4)	N/A	N/A	-.11±0.40 (-1.6,1.0)	-1.1±1.44 (-5.1,1.4)
FVC (%-pred) at first ALSFRS-R visit	mean(SD)	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)	66.8±21.8 (12.0,100)
FVC slope		-1.5±2.53 (-16,14.1)	-1.6±2.59 (-16,14.1)	N/A	N/A	-.12±0.86 (-1.9,2.1)	-1.9±2.66 (-16,3.0)

Table 2. **Table WGA1. 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 Genes* Variants in ALS	ALS Genes* 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=1092 Average= 1.5 variant per ALS case	Sum=141 Average= 1.5 variant per control	*ALS2* (20), *ANG* (5), *ANXA11* (15), *ATXN2* (29), *C21orf2* (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), *PFN1* (7), *SCFD1* (13), *SETX* (57), *SOD1* (14), *SQSTM1* (14), *TAF15* (16), *TARDBP* (11), *TBK1* (18), *TUBA4A* (3), *UBQLN2* (7), *UNC13A* (19), *VAPB* (4), *VCP* (4). Details of variants are found in Sup tables Table
ClinVar P/LP (C-PLP) Variants	Sum= 23,924 Average= 33.9 variant per ALS case Rare =3,659 5.2 variants per ALS case	Sum= 3,097 Average= 33.7 variant per control Rare= 461 5 variants per control	Sum= 85 12% of cases harbor Rare only = 21 (3% of cases harbor)	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)**, *PFN1* (2), *SETX* (4), *SOD1* (5), *SQSTM1* (3),*TARDBP* (2), *UBQLN2* (2), *VCP* (1)
Harms P/LP (H-PLP) Variants	N/A	N/A	Sum= 24 3.4% of cases harbor)	Sum=1 1% of controls harbor	*FUS* (1), *PFN1* (2), *SOD1* (11), *TARDBP* (3) *UBQLN2* (1), *VCP* (1)
Intervar P/LP (I-PLP) variants	Sum= 2,346 Average 3.3 variant per sample Rare= 2272 Average 3.21 per case	Sum= 288 Average 3.1 Variant per sample Rare= 276 Average 3.2 per control	Sum=25 3.5% of cases harbor	0 (0%)	*NEK1* (2), *OPTN* (1), *SOD1* (12), *SETX* (1), *TBK1* (2), *VCP* (2),
In silico Prediction 6/9 Predicted to be damaging	Sum= 79,010 Average 112 variant per sample Rare= 40,910 Average 58 variants per sample	Sum=5,464 Average 113 variant per sample Rare=5,464 Average 59.4 variant per sample	Sum= 97 (13.7% of cases harbor)	Sum=11 (12% of controls harbor)	*ALS2*(2), *ANXA11* (4), *ATXN2* (3), *C21orf2* (1), *CAMTA1* (1), *DAO* (3), *DCTN1* (5), *FIG4* (3), *FUS* (1), *HNRNPA2B1* (1), *KIF5A* (2)MOBP(1), NEK1(2), *OPTN* (1), *PFN1* (3), *SCFD1* (2), *SETX* (14), *SOD1* (11), *SQSTM1* (2), *TARDBP* (4), *TUBA4A* (2), *UBQLN2* (1), *UNC13A* (3), *VCP* (2)
Sum= the total number of variants found per group, ALS vs. Control. ** OPTN variants are high frequency < 1%.					

Figures

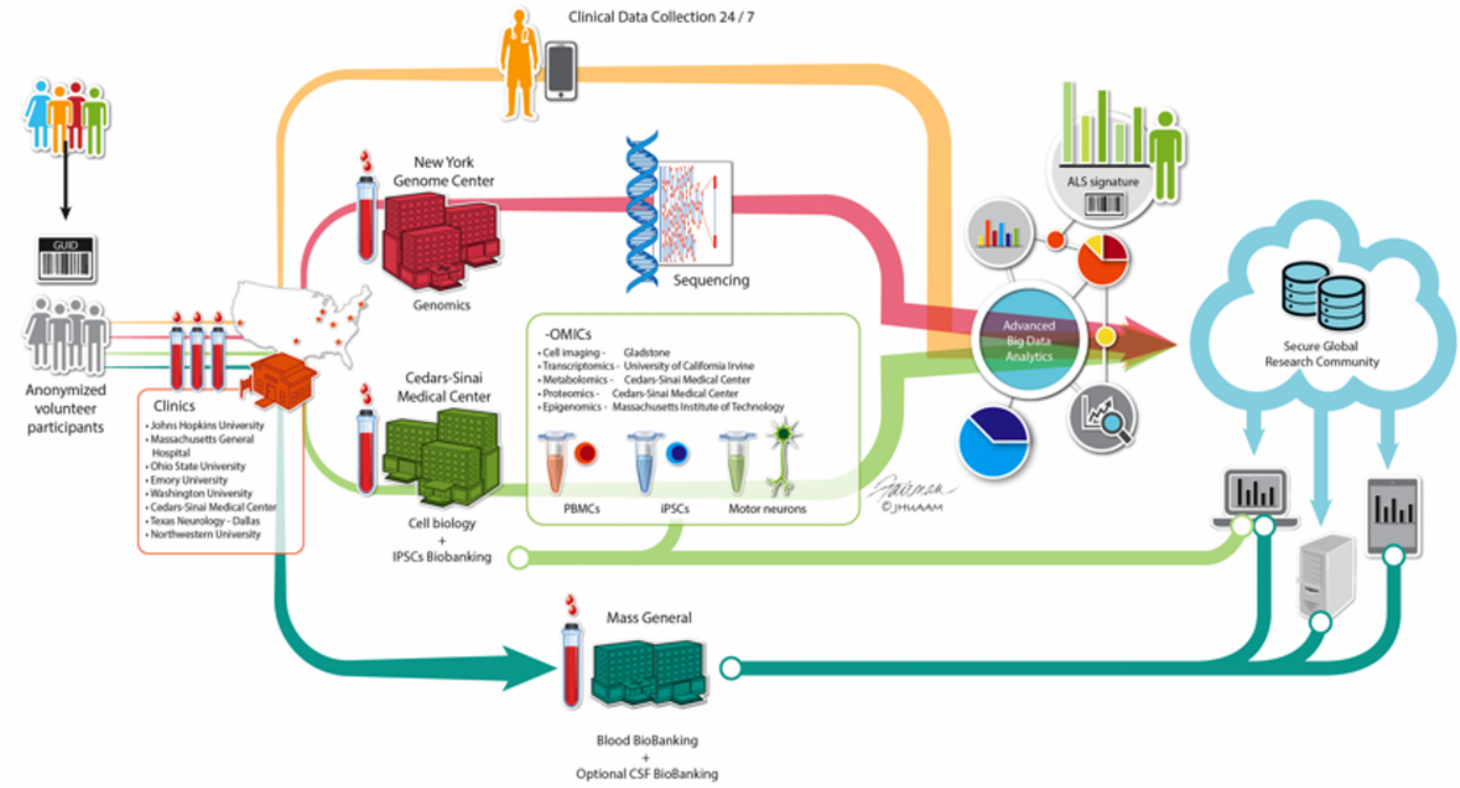

Figure 1

Answer ALS Research Program. Graphic illustration of overall program flow.

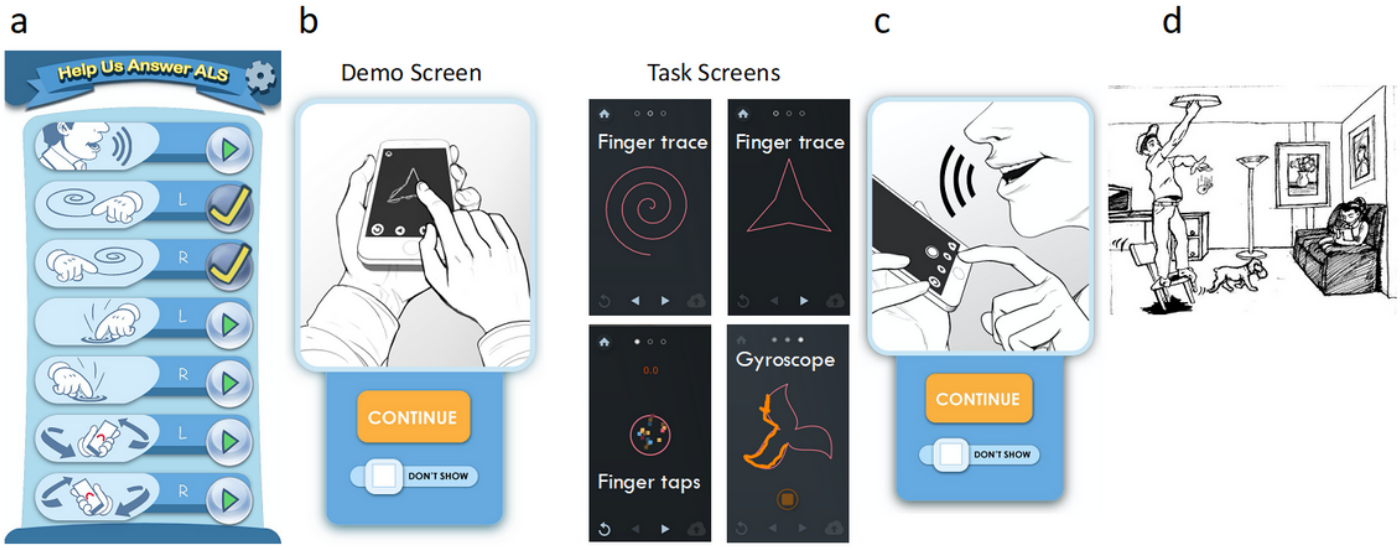

Figure 2

Smartphone App. 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.

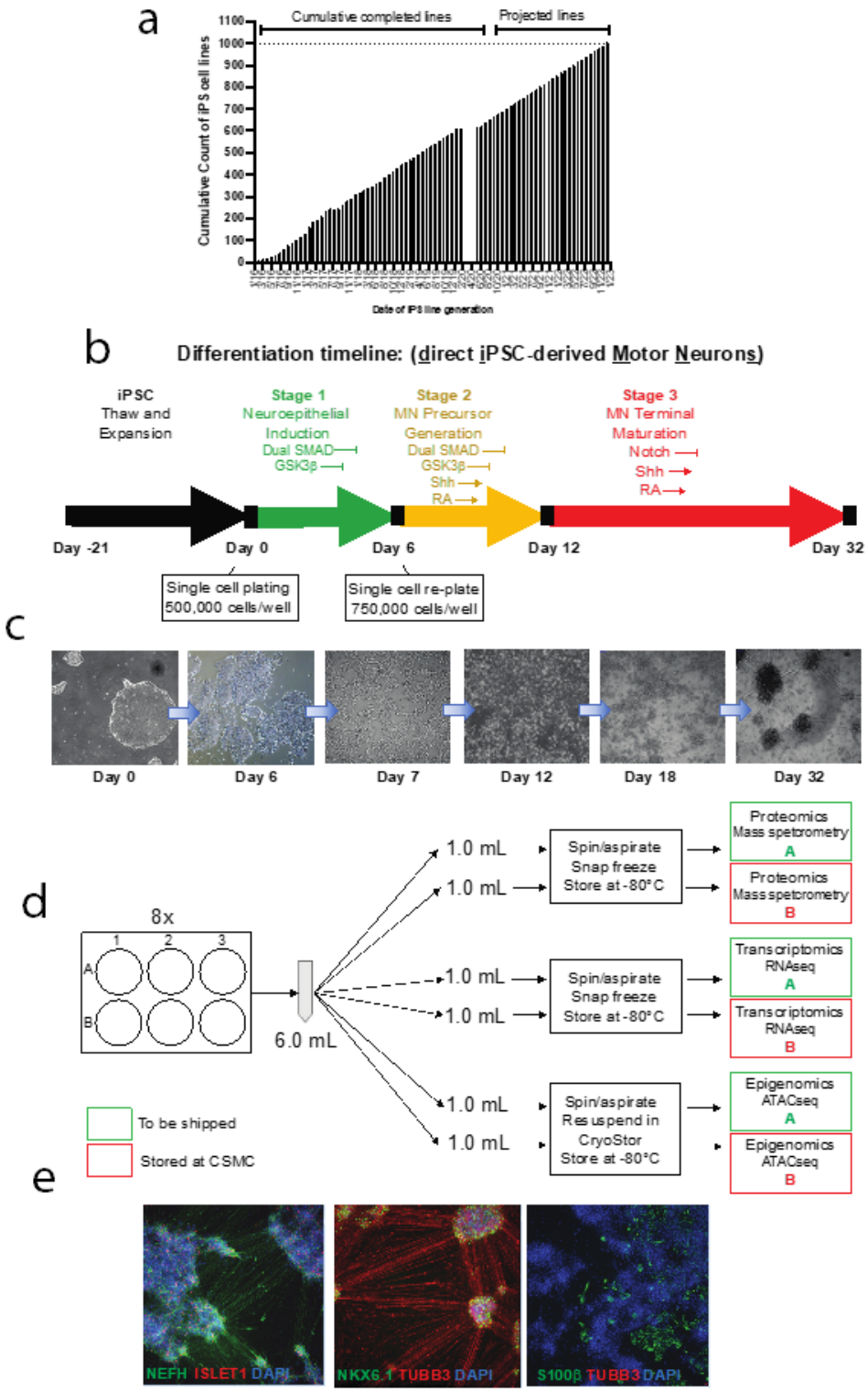

Figure 3

Production of ALS and control iPSC spinal motor neurons. A. Example of IPS Generation Schedule. B. Method of generating iPSC derived motor neuron cell lines using the diMNs protocol. C. Brightfield images show the morphology of the cells during differentiation from iPSC 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 iPSCs is performed by imaging of

representative wells for immunohistochemical staining with neuronal, motor neuron and glial markers after 32 days of differentiation¬¬

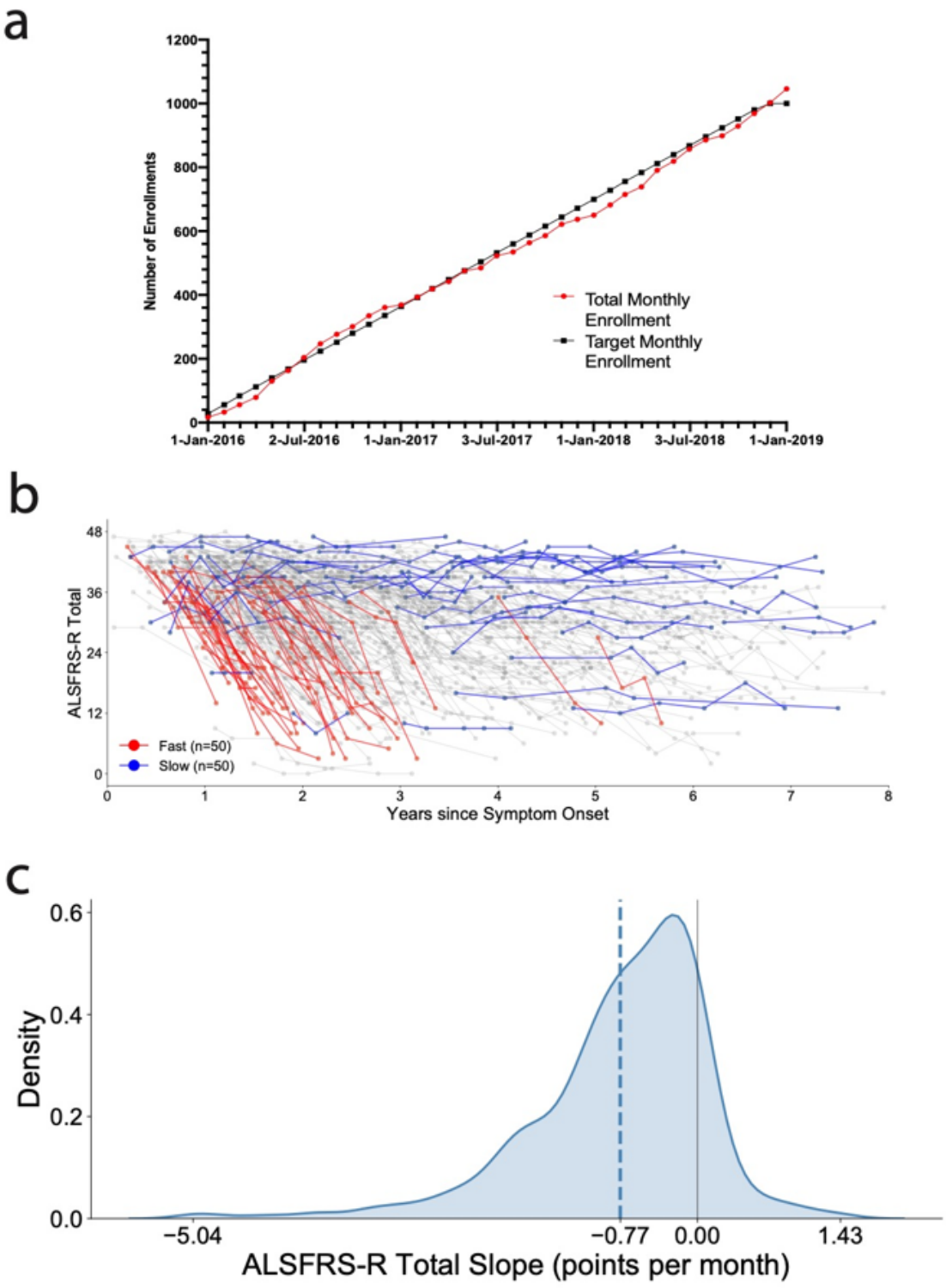

Figure 4

Clinical Enrollment and Characteristics: ALSFRS-R progression curves for all Answer ALS clinic enrolled subjects over a 40-month period. A. Answer ALS patient and control subject enrollment. B. ALSFRS-R Total Slope Distribution. Kernel density estimation with Gaussian kernels used to estimate probability

density function of ALSFRS-R slope. Dashed line indicates the mean ALSFRS-R slope. C. Longitudinal ALSFRS-R measurements with fast and slow progressors. Participants with 3 or more visits and a maximum visit date within 8 years of symptom onset included. N indicates the number of participants in fast and slow progressing groups, sorted by ALSFRS-R slope.

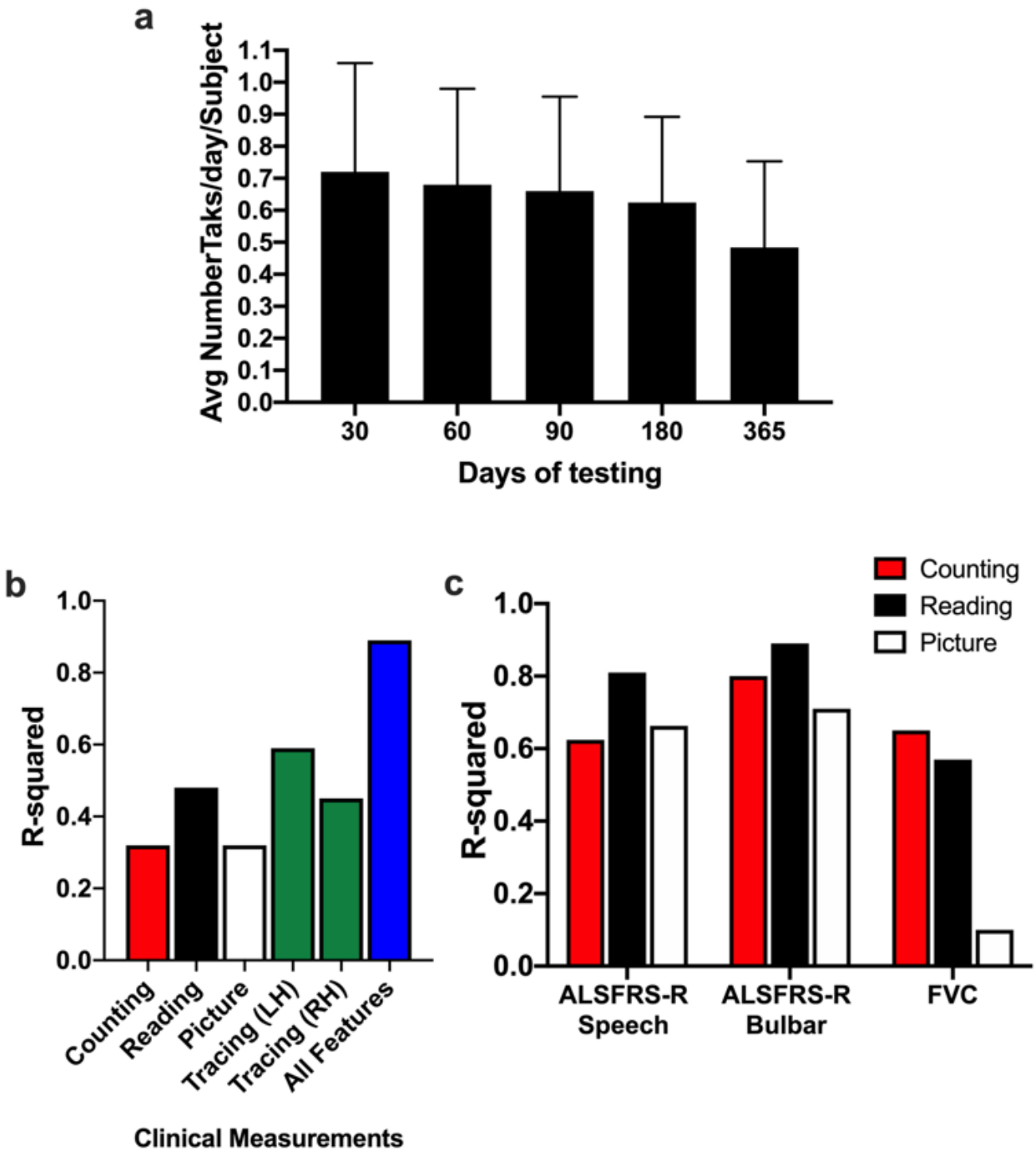

Figure 5

Smartphone use and analytics. A) Smartphone app compliance. Compliance was calculated using the average number of tasks done per day and per subject. B) Results of inferring ALFRS-R total. Pearson correlation 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<1E-5), C) Results of inferring ALSFRS-R scores using only speech related tasks. Pearson values were calculated independently for each of the 3 speech tasks to infer FVC and ALSFRS-R speech and bulbar sub-scores. Highest performance was obtained using information from the reading task for both ALSFRS-R sub-scores obtaining up to R=0.89 (p<1E-5) for ALSFRS-R bulbar sub-score. On the other hand, counting task information produced the best result when infer FVC score (R=0.65, p=2E-2).

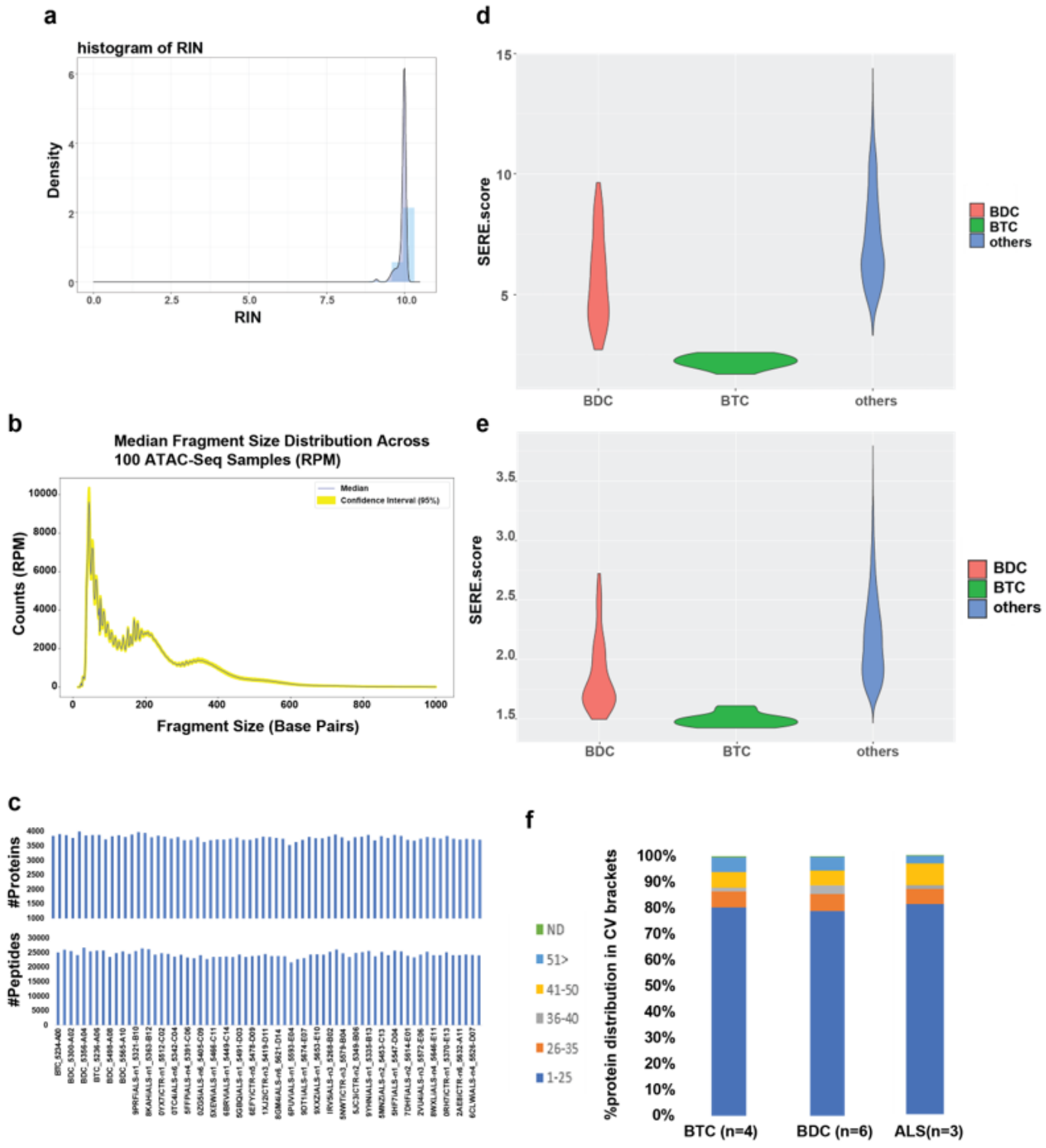

Figure 6

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 ~147bp, 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. Violing 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%.

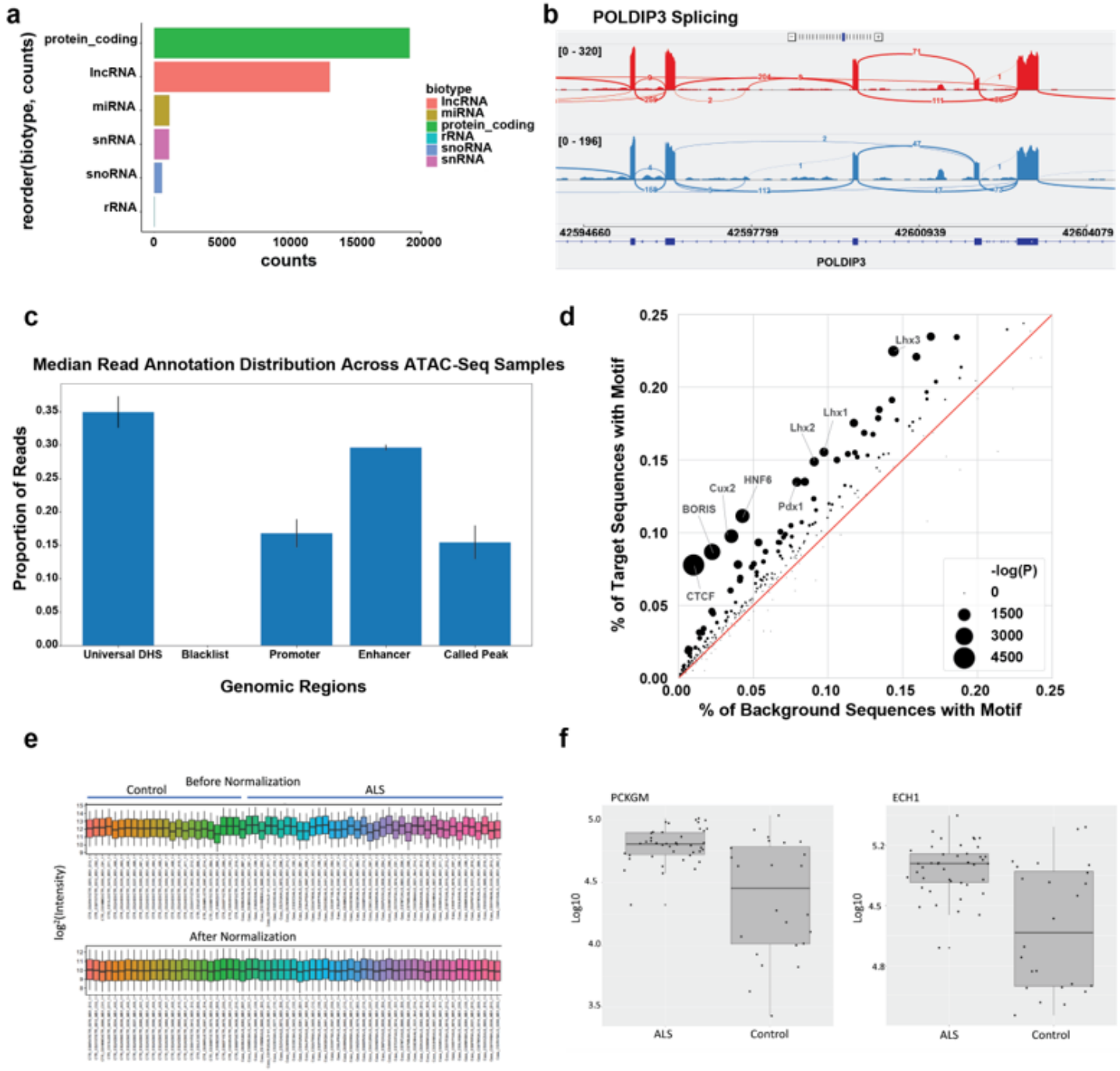

Figure 7

Omics exploratory analysis of results. A. Barplot showing counts of RNA species identified in the current AALS samples. Barplot showing the different number of RNA species/types across all current AALS samples. As expected, protein coding and lncRNAs represent the largest proportion while rRNA, which have been depleted, are the lowest. Types represented are: protein coding, lncRNA, miRNA, snRNA, snoRNA, and rRNA in green, red, gold, purple, blue, and lightblue, respectively. B. Sashimi Plot showing POLDIP3 Splicing in AALS samples. Sashimi plot shows alt-splicing results of POLDIP3 in the current AALS samples, CTR in red and ALS in blue. Showing increased exon 3 skipping in ALS samples, which has previously been described in ALS. C. 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 (DHS) and previously annotated enhancers and promoters. 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. E. log2 protein intensity distribution unnormalized (top) and normalized (bottom). F. Log 10 protein intensity comparison of selected proteins (PCKGM, ECH1) showing differential expression between ALS and Controls.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SuppTables112Oct20.xlsx
- SupplementaryForms1ALSCRFsV4.pdf
- ExtendedDataTables123456Oct20.docx
- ExtendedFiguresOct20FINAL.docx
- SupplementalMethodsOct14JDR.docx