**Supplemental Methods**

**Biofluid Collection and Processing**

At each in-clinic visit, approximately 50–100 ml of blood was collected from each participant. Plasma and serum were processed at the site using standard protocols, divided into 0.5mL aliquots and flash-frozen at -80°C within 30 minutes of collection. In addition, at the first clinic visit ~ 24 ml of blood was collected and processed for PBMC isolation as indicated in the PBMC processing section below. Whole blood was also collected in EDTA tubes and sent to New York Genome Center (NYGC) for DNA extraction and whole genome sequencing. Whole blood collected in EDTA tubes was aliquoted unprocessed and flash-frozen. Cerebrospinal fluid (CSF) was optionally collected, centrifuged at a standard speed to eliminate cellular components, aliquoted and flash-frozen at -80°C. At follow-up visits, plasma and serum were collected, and whole blood, urine and CSF were collected optionally, processed and stored as above. If blood was not collected at the initial visit or quality control problems prevented appropriate analysis or iPSC creation, then blood was collected at a subsequent visit. Serum, plasma, and CSF samples were shipped on dry ice to a centralized biofluid repository at Massacchusetts General Hospital (NeuroBank) to be stored at -80°C (See Supplemental Table 3).

**Answer ALS Smartphone**

Limb function tasks included finger tapping, finger tracing, and gyroscope-based tracing that called for the participant to tilt the phone to trace a complex line/figure. Each task was performed using the right and left hand separately, for a total of six individual tasks (Figure 2B,C).

For the speech module (Figure 2C), participants were asked to perform one of 3 tasks: 1) Single-Breath Count, a surrogate for forced vital capacity 1 in which participants were instructed to draw in a deep breath and count at a measured pace; 2) Read Aloud Passage, in which text such as the Bamboo Passage, a standard text used to assess motor speech characteristics, was displayed on the screen and participants were asked to read it aloud, and 3) Free Speech (picture description)(Figure 2D), in which participants were presented with a cartoon showing various activities, and asked to record their description of the particular scene over a period spanning 30 seconds to two minutes. These tasks alternated each week, so each one was performed once every 3 weeks. In addition, for the reading and free speech tasks, we alternated between 4 passages and 3 different pictures to reduce learning effects. We analyzed compliance over time, calculating the average number of tasks (total tasks, limb tasks, and bulbar tasks) completed per week of use to evaluate continued engagement with the app.

 Depending on the type of task, different features were extracted. To characterize the data obtained from the arm function tasks, errors metrics such as Hausdorff and dynamic time warping distances were calculated. In addition, the number of points acquired by the device during the tracing task was also obtained as a measurement of speed. On the other hand, for speech tasks, we employed standard acoustic features that are known to assess speech degradation such as pitch variations, prosody features, vowel space, vowel quality, noise measurements, mel frequency cepstral coefficients (MFCCs), tremor features, among others. Since one of the speech tasks also evaluates cognition (free speech), recordings were manually transcribed because automated speech-to-text engines were unable to reliably detect 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 reported2. 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 different machine learning algorithms (linear, ridge, lasso regression) and validated using a 10-fold cross-validation approach.

**IPSC Line Supplemental Methods:**

**PBMC Processing**: Fresh blood was collected into 3 8-ml Sodium Citrate BD Vacutainer CPT Tubes (BD, Cat 362761) according to the manufacturer’s instructions. Samples were centrifuged at 18-25oC in a horizontal rotor centrifuge for 20 minutes at 1800 RCF within 2 hours of collection. After centrifugation, tubes were inverted to mix the separated buffy coat and plasma layer together. The tubes were then packaged and shipped to Cedars-Sinai via overnight delivery. Once received at Cedars-Sinai, the plasma/buffy coat mixture was collected and centrifuged for 15 mins at 300 RCF. Isolated PBMCs were counted and cryopreserved at 5 million cells per vial in a 1:1 mixture of plasma and CyrostorCS10. Vials were placed in an alcohol freezing container (Mr. Frosty, Nalgene) overnight at -80oC before being transferred to liquid nitrogen for long-term storage.

**iPSC methods.** Briefly, iPSC colonies were maintained on Matrigel-coated 6-well plates (Falcon 353934) at a concentration of 1 mg Matrigel / plate. The Batch Technical Control lines (BTC) through Batch 14 were cultured in mTeSR1 (Stemcell Technologies). All subsequent batches were cultured in mTeSR+. For conciseness, “mTeSR media” will hereafter refer to both mTeSR1 and mTeSR+. Each cell line was thawed and cultured for two to three weeks before passaging for differentiation. Cell lines were differentiated in batches of up to eleven lines.

The direct induced motor neuron (diMN) protocol comprises three stages (Fig 4). At the outset of Stage 1, plates from each iPSC cell line were washed with 1 mL DPBS (Corning 21-031-CV)/well and then incubated in 1 mL Accutase (EMD Millipore SCR005)/well for 5 minutes at 37°C. After incubation, 1 mL DPBS/well was added, cells were quickly collected into multiple 15-mL conical tubes (Falcon 352097). Typically, one 6-well plate was collected per tube. Tubes were then centrifuged at 161 g for 2 minutes. The supernatant was aspirated and discarded, and each pellet was re-suspended in 1 mL mTeSR media by gentle trituration using a P-1000 pipette. Once resuspended, all pellets were combined in a final volume of up to 10 mL mTeSR media. Viability and concentration were determined by automated cell counting (Nexcelom Auto 2000). Based on the cell concentration, up to four Matrigel-coated 6-well plates were seeded at a density of 5.0x10^5 cells/well in 2 mL mTeSR media/well. Twenty-four hours following platedown, mTeSR media was exchanged for Stage 1 media (refer to Supplemental Methods Table 1 for composition). Stage 1 media was exchanged daily until Day 6.

Day 6 began Stage 2 of the differentiation process. For each cell line, all wells were washed with 1 mL DPBS/well and incubated in 1 mL Accutase/well for 5 minutes at 37°C. After incubation, 1 mL DPBS/well was added, cells were quickly collected into multiple 15-mL conical tubes and centrifuged at 161 g for 2 minutes. The supernatant was aspirated and discarded, and each pellet was re-suspended in 1 mL Stage 2 Platedown Media (ST2PD, Supplemental Methods Table 2) by gentle trituration using a P-1000 pipette. Once resuspended, all pellets were combined in a final volume of up to 10 mL St2PD. Viability and concentration were then determined by automated cell counting. Based on the cell concentration, up to nine Matrigel-coated 6-well plates were seeded at a density of 7.5x10^5 cells/well in 2 mL St2PD/well. 24 hours following platedown, St2PD was exchanged for Stage 2 media (Supplemental Methods Table 3). Stage 2 media was exchanged every other day until day 12.

Day 12 began Stage 3 of differentiation. For each cell line, Stage 2 media was completely aspirated from all wells and replaced with 2 mL Stage 3 media/well. Stage 3 media (Supplemental Methods Table 4) was exchanged every other day until Day 32 of differentiation. During feedings, approximately 75% of old media was aspirated and 2 mL Stage 3/well was added dropwise in a circular manner in order to minimize disruption of the cell monolayer.

On Day 32 of differentiation, cell lines were collected and pelleted as illustrated in Fig. 4. Prior to collection and pelleting, one 6-well plate was selected from each line for brightfield imaging (Molecular Devices ImageExpress Micro). Six regions of interest were captured per well at a magnification of 10X. After imaging, the plates were collected with their respective lines.

For each cell line, the number of wells in which the cell monolayer became detached was recorded (Figure 4). The mean detachment rate was ~19% (SD +/-0.09). Any of these “lifted” monolayers were not included in the pellet. In addition, four wells/cell line were set aside for short tandem repeat (STR) analysis. For all remaining adherent wells, Stage 3 media was aspirated and replaced with 1 mL DPBS/well. Adherent cell monolayers were manually scraped with a cell scraper (Falcon #353085) and collected using a serological pipette into 15-mL conical tubes. Typically, two 6-well plates were collected per 15-mL conical, and up to eight 6-well plates were collected per line. The 15-mL conicals were centrifuged for 2 minutes at 161 g. The supernatant was then aspirated and discarded, and the pellets were re-suspended in 1 mL DPBS by gentle trituration using a P-1000 pipette. Once resuspended, all pellets were combined in a final volume of approximately 10 mL DPBS and centrifuged for 2 minutes at 161 g. Again, the supernatant was aspirated and discarded. The pellet was then resuspended in 6mL DPBS using a 5-mL serological pipette and aliquoted to six 1.7-mL Eppendorf tubes (1 mL/Eppendorf tube). The Eppendorf tubes were centrifuged for 4 minutes at 161 g, and the supernatants were aspirated and discarded. Four of the Eppendorf tubes were snap-frozen in an ethanol/dry ice slurry and stored at -80˚C until shipment to omics centers for analysis. The remaining two pellets were re-suspended in 1 mL each of CryoStor CS10 (Biolife Solutions #210102) using a P-1000 pipette (typically, 2-4 triturations were sufficient to resuspend the pellets) and each pellet was transferred to an individual cryovial (Thermo Scientific #5000-1020). CryoStor vials were then stored in a Mr. Frosty (Nalgene #5100-0001) at -80˚C for 24 hours, at which time they were transferred to sample boxes and stored at -80˚C until shipment to omics center for processing.

**Quality Control of diMNs**

As referenced in Figure 4, on Day 32 we reserved one 6-well plate from each cell line for immunostaining. We stained each line with the following markers of neuronal differentiation: SMI32(NF-H), TUBB3(TUJ), ISL1, NKX6.1, S100β, and Nestin, as well as DAPI to obtain a total nuclear stain.

To begin the process, each plate was fixed as follows: old media was aspirated and each well washed with 1 mL DPBS +Ca/+Mg (Corning #21-030-CV)/well. Cells were then incubated in 1 mL 4% paraformaldehyde (PFA) solution/well for 10 minutes at room temperature. After incubation, PFA was aspirated and each well carefully washed with 1 mL DPBS (Ca+/Mg+). Finally, 3 mL DPBS (Ca+/Mg+)/well was added and the plates stored at 4˚C until immunostaining.

For immunostaining, each well was blocked for 1 hour at room temperature (5% normal donkey serum (EMD Millipore #S30), 0.2% Triton X-100 (Sigma-Aldrich #T9284) in DPBS (Ca+/Mg+)). Following blocking each well was incubated with primary antibody (refer to Table# for antibody reference and dilution) for 1 hour at room temperature. Following primary incubation, each well was washed in 1 mL washing solution (0.1% Triton X-100 in DPBS (Ca+/Mg+))/well for 2-3 minutes. Following the wash, secondary antibody (Supplemental Methods Table 5) was added to each well and allowed to incubate for 1 hour at room temperature in the dark. Following secondary incubation, each well was washed with 1 mL washing solution/well for 2-3 minutes. Following the wash, each well was incubated with DAPI solution (Supplemental Methods Table 4) for 3 minutes at room temperature. Wells were then washed again with 1 mL DPBS (Ca+/Mg+)/well. Finally, 1 mL DPBS (Ca+/Mg+) was added to each well, and the plates were covered with aluminum adhesive film and stored at 4˚C until image acquisition using the ImageExpress Micro system (Molecular Devices). During image acquisition, 64 regions of interest were captured per stained well.

Supplemental Methods Table 1

|  |  |  |  |
| --- | --- | --- | --- |
| **Stage 1****(Day 0-Day 6)** | **Manufacturer** | **Catalog #** | **1X Concentration** |
| IMDM | Life Technologies | 12440061 | 47.5% |
| F12 | Life Technologies | 11765062 | 47.5% |
| NEAA | Life Technologies | 11140-50 | 1% |
| B27 | Life Technologies | 17504044 | 2% |
| N2 | Life Technologies | 17502048 | 1% |
| Anti/Anti | Life Technologies | 15240062 | 1% |
| LDN193189 | Cayman Chemical | 19396 | 0.2 µM |
| CHIR99021 | Xcess bioscience | M60002 | 3 µM |
| SB431542 | Cayman Chemical | 13031 | 10 µM |

Supplemental Methods Table 2

|  |  |  |  |
| --- | --- | --- | --- |
| **Stage 2 Platedown****(Day 6)** | **Manufacturer** | **Catalog #** | **1X Concentration** |
| IMDM | Life Technologies | 12440061 | 47.45% |
| F12 | Life Technologies | 11765062 | 47.45% |
| NEAA | Life Technologies | 11140-50 | 1% |
| B27 | Life Technologies | 17504044 | 2% |
| N2 | Life Technologies | 17502048 | 1% |
| Anti/Anti | Life Technologies | 15240062 | 1% |
| All-trans RA | Stemgent | 04-0021 | 0.1 µM |
| SAG | Cayman Chemical | 11914 | 1 µM |
| LDN193189 | Cayman Chemical | 19396 | 0.2 µM |
| CHIR99021 | Xcess bioscience | M60002 | 3 µM |
| SB431542 | Cayman Chemical | 13031 | 10 µM |
| Rock Inhibitor (Y-27632) | Stemcell Technologies | 72308 | 10 µM |

Supplemental Methods Table 3

|  |  |  |  |
| --- | --- | --- | --- |
| **Stage 2****(Day 7-11)** | **Manufacturer** | **Catalog #** | **1X Concentration** |
| IMDM | Life Technologies | 12440061 | 47.5% |
| F12 | Life Technologies | 11765062 | 47.5% |
| NEAA | Life Technologies | 11140-50 | 1% |
| B27 | Life Technologies | 17504044 | 2% |
| N2 | Life Technologies | 17502048 | 1% |
| Anti/Anti | Life Technologies | 15240062 | 1% |
| All-trans RA | Stemgent | 04-0021 | 0.1 µM |
| SAG | Cayman Chemical | 11914 | 1 µM |
| LDN193189 | Cayman Chemical | 19396 | 0.2 µM |
| CHIR99021 | Xcess bioscience | M60002 | 3 µM |
| SB431542 | Cayman Chemical | 13031 | 10 µM |

Supplemental Methods Table 4

|  |  |  |  |
| --- | --- | --- | --- |
| **Stage 3****(Day 12-Day32)** | **Manufacturer** | **Catalog #** | **1X Concentration** |
| IMDM | Life Technologies | 12440061 | 47.5% |
| F12 | Life Technologies | 11765062 | 47.5% |
| NEAA | Life Technologies | 11140-50 | 1% |
| B27 | Life Technologies | 17504044 | 2% |
| N2 | Life Technologies | 17502048 | 1% |
| Anti/Anti | Life Technologies | 15240062 | 1% |
| \*SAG | Cayman Chemical | 11914 | 0.1 µM |
| db-cAMP | Millipore | 28745 | 0.1 µM |
| All-trans RA | Stemgent | 04-0021 | 0.5 µM |
| Compound E | Calbiochem | 565790 | 0.1 µM |
| DAPT | Cayman Chemical | 13197 | 2.5 µM |
| Ascorbic Acid | Sigma-Aldrich | A4403 | 200 ng/mL |
| BDNF (-80) | Peprotech | 450-02 | 10 ng/mL |
| GDNF (-80) | Peprotech | 450-10 | 10 ng/mL |

Supplemental Methods Table 5

|  |  |  |  |
| --- | --- | --- | --- |
| **Antibody or Stain** | **Manufacturer** | **Catalog #** | **Dilution** |
| SMI 32 | Biolegend | 801701 | 1:1000 |
| Human Islet-1 Antibody | R&D | AF1837 | 1:250 |
| Nkx-6.1 | DSHB | F55A10-s | 1:1000 |
| Tuj (TUBB3 polyclonal antibody) | Abnova | PAB7874 | 1:1000 |
| Anti-Nestin, Human Antibody | Sigma Aldrich | ABD69 | 1:1000 |
| S100beta | Sigma Aldrich | S2532 | 1:250 |
| Donkey anti-mouse secondary | Life Technologies | A-10037 | 1:1000 |
| Donkey anti-rabbit secondary | Life Technologies | A-31573 | 1:1000 |
| Donkey anti-goat secondary | Life Technologies | A-21447 | 1:1000 |
| DAPI | Life Technologies | D3571 | 0.1µg/mL |

**Batch technical control (BTC):** **The BTC controls for technical variability of a particular ‘Omics assay between different batch runs.** Briefly, one iPSC line from a healthy donor (CS2AE8iCTR-n6 line) was differentiated in a single large batch at the beginning of the project at the cell generation center (Cedars-Sinai iPSC Core). Multiple biological samples, including snap frozen cell pellets and cryopreserved cell pellets, were prepared to last over a significant period of the data generation component of the project. With each shipment batch, end users at each ‘Omics center receive the appropriate BTC biological sample. Each shipment batch comprises three to four batches of iPSC-derived motor neurons of ALS and healthy control (CTR) subjects, as well as the BTC biological sample, while each differentiation batch comprises 10-15 iPSC lines from different experimental subjects. Since BTC pellets were produced at same time with the same diMNs differentiation standard operating procedure (SOP), a given assay should technically return similar results for any BTC sample across multiple ‘Omics batch runs. The BTC thus controls for ‘Omics assay-specific variability.

**Batch differentiation control (BDC):** **The BDC controls for inter-batch variability in iPSC differentiation to diMNs.** Briefly, a differentiation batch comprises 10-15 iPSC lines from different ALS and CTR subjects. The same iPSC line used to produce the BTC (CS2AE8iCTR-n6 line) is differentiated in every batch with the other experimental iPSC lines at the cell generation center and is referred to as the Batch Differentiation Control (BDC). This line is thawed, expanded, differentiated, and pelleted in addition to the ALS or healthy control (CTR) lines in each batch. The repeated differentiation of this single line, therefore, serves as a differentiation control, reflecting the intrinsic variability in the iPSC to diMNs differentiation process of the same line across multiple differentiation batches. ‘Omics centers receive a BDC sample along with ALS and CTR samples for each differentiation batch. Thus, in addition to the BTC sample, a shipment to the ‘Omics center contains multiple BDC samples (one for each differentiation batch included in the shipment).

**RNA Methods**

Total RNA was isolated from each sample using the Qiagen RNeasy mini kit. RNA samples for each AALS subject (control or ALS) were entered into an electronic tracking system and processed at the University of California, Irvine GHTF. RNA QC was conducted using an Agilent Bioanalyzer and Nanodrop. Our primary QC metric for RNA quality is based on RIN values (RNA Integrity Number) ranging from 0-10, 10 being the highest quality RNA. Additionally, we collected QC data on total RNA concentration and 260/280 and 260/230 ratios to evaluate any potential contamination. Only samples with RIN >8 were used for library prep and sequencing.  Library prep processing was initiated with a total of 1 μg RNA using a Ribo-Zero Gold rRNA depletion and Truseq Stranded total RNA kit. Additionally, ERCC exFold spiked-in controls were used for further QC and downstream data analysis. Briefly, RNA was chemically fragmented and subjected to reverse transcription, end repair, phosphorylation, A-tailing, ligation of barcoded sequencing adapters, and enrichment of adapter-ligated cDNAs.  RNA-Seq libraries were titrated by qPCR (Kapa), normalized according to size (Agilent Bioanalyzer 2100 High Sensitivity chip). Each cDNA library was then subjected to Illumina (HiSeq 2500) paired end (PE), 100 cycle sequencing to obtain approximately 50-65 M PE reads. After sequencing fastq were subject to QC measures and reads with quality scores (>Q20) collected and analyzed using the pipeline described at http://neurolincs.org/tools/. Briefly, reads were mapped to the GRCh73 reference genome, QCed, and gene expression and differential expression were quantified using the pipeline outlined here: <http://galaxy.neurolincs.org/u/terri/p/neurolincs-data-analysis-workflows>, using tools HTseq1 and DESeq22. Normalized and transformed count data were then used for exploratory analysis and DE genes (FDR <0.1) were used for pathway, network, and gene ontology analysis. These primary data were subject to additional statistical and network-based data analyses using commercial and open-source pathway and network analysis tools, including Ingenuity Pathway Analysis (IPA), Gorilla, Cytoscape, and other tools to identify transcriptional regulators, predict epigenomic changes, and determine potential downstream pathway and cellular functional effects. Significant DEGs (FDR<0.1) were then analyzed against genes that were found to contain exonic enriched genetic variants from the WGS. The gene expression (voom normalized and transformed values) and genotype variant pairs were analyzed by fitting a linear regression model. Adjusted R2 and Benjamini-Hochberg adjusted p-values were calculated, significant genes were reported at (FDR<0.1).

**ATAC seq Methods**

ATAC-seq was carried out as further described3*.* Briefly, cells were lysed in ATAC-seq resuspension buffer (RSB; 10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM MgCl2, protease inhibitors) with a mixture of detergents (0.1% Tween-20, 0.1% NP-40, 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. 50K nuclei were collected and centrifuged at 450 rcf for 5 min at 4 °C. The pellet was re-suspended in 50 µl of transposition mixture (25 µl 2X Illumina Tagment DNA Buffer, 2.5 µl Illumina Tagment DNA Enzyme, 16.5 µl PBS, 0.5 µl 1% digitonin, 0.5 µl 10% Tween-20, 5 µl water). The transposition reaction was incubated at 37°C for 30 min followed by DNA purification. An initial PCR amplification was performed on the tagmented DNA using Nextera indexing primers (Illumina). Real-time 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 (paired-end, 75nt kit). All samples passed quality control checks that included morphological evaluation of nuclei, fluorescence-based electrophoresis of libraries to assess size distribution, and real-time 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-calling software MACS24 and determined differentially open sites using DESeq2 (FDR<0.1). Peaks were assigned to unique genes using the default HOMER5 parameters, and gene ontology analysis was performed using GOrilla6.

**Proteome Methods**

Snap frozen cell pellets were stored at -80oC and transferred to the CSMC proteomics lab on dry ice where it was stored at -80oC until use. Samples were lyophilized and aliquoted into 600 ul polystyrene microcentrifuge tubes containing lysis buffer (6M Urea, 1 mM DTT in 1.5 M NH4HCO3). Sample was sonicated (QSonica Q800R1) by alternating 10 seconds on and 10 seconds off at 70% amplitude while rotating in a 4°C water bath until the solution was homogenized (~20 mins). Samples were centrifuges and the protein concentration determined on the supernatant according to manufactures’ instructions (PierceTM BCA Protein Assay Kit). 200 ug of each sample was transferred to a 96 well plate in aliquots and processed on the Biomek i7 Automated workstation (Beckman Coulter) as outlined previously. Briefly, samples underwent the following: reduction of disulfide bonds in 3 mM TCEP, alkylated in 5 mM IAA. Addition of Beta-galactosidase at 2 ug and protein digestion in solution using equal molar Trypsin and LysC enzyme mixture (Promega, product #: V5111) at 1:40 enzyme to protein ratio under optimized digestion conditions (4 hours at 37°C). Digested proteins were desalted on a 5 mg Oasis HLB 96 well plate (Waters; product #: 186000309 ) and eluted in 50% acetonitrile. Samples were dried to completion using a speed-vac system and stored at -80oC until MS analysis. For MS analysis, digested peptides were resuspended in 0.1% 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. DDA data used for the generation of a sample specific peptide ion library. DIA data files were analyzed with OpenSWATH pipeline against the sample specific peptide ion library generated. Protein level quantitation is calculated by summing transition level intensities for all the proteotypic 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 iPSC lines were plated on 96-well plates for longitudinal single cell imaging with robotic microscopy. At day 25, cells were transfected with expression marker plasmids such as synapsin::mApple7 to visualize cell morphology and viability. After transfection, cells were imaged in an automated fashion with robotic microscopy once per day for 10–14 days. A fiducial mark from the plate was imaged during the first imaging run and then used each time thereafter to register the position of the plate and align it to its initial position. This enabled the system to collect images of the same microscope fields over the course of the experiment and to identify and track individual iMNs. Images of different microscope fields from the same well were stitched together into montages, and montages of the same well collected at different time points were organized into composite files in temporal order. Some image analysis was performed in a computational pipeline constructed within the open source program Galaxy, to identify and track individual cells and perform survival analysis and other morphological measurements. Independently, images from iMNs from ALS patients and healthy volunteers were analyzed with machine learning methods (described below) in a relatively unbiased fashion to discover if they could be stratified or sub-stratified to predict which iMNs were derived from ALS patients.

**Image-Based Machine Learning Methods**

Supervised Machine Learning (SML) and Deep Learning (DL) methods were used to determine if there are morphological characteristics that can differentiate ALS diMNs from controls. RM images from control and ALS lines on differentiation day ~30 were processed using our custom-built image processing pipeline using Galaxy software. From each image, crops were generated from detected objects to include only the cell body, the cell body/soma and neurite, or exclusively neurite. These crops were then pre-processed using the enhanced contrast feature with a pixel saturation of 0.5% and normalization using FIJI8. The neurites and soma were sorted into different groups using a Python (<https://www.python.org/>) ellipse-based soma selection script, and confirmed for morphology afterwards by eye. To obtain the final pool of object crops for feature detection, images were selected that had a high signal-to-noise ratio which could define the neurite or the soma from the background without image artifacts. Equal numbers of ALS and control images were then subjected to the feature detection algorithms so the input data were balanced. The feature detection algorithms were written in Python and contained within Jupyter notebooks (<https://jupyter.org/>). They used various image processing packages (OpenCV, Pillow, Skimage, etc.) and numerous algorithms as such bright spot analysis9, overall intensity10, distance mapping**11**, Minkowski-Bouligand fractal dimension12, contour analysis13, and Fast Fourier Transform14, and many others. These features were used to train several models including Logistic Regression15, Support Vector Machine16 (SVM), Artificial Neural Network17, MLP Classifier18, Random Forests19, XGBoost20, and Majority Voting21, with a train/test split of 80/20 and 5 fold K- cross validation. These scores are output into a CSV file for both ALS and control crops was then fed into the machine learning Jupyter notebook. CSV files were pre-processed by ensuring all non-string values were numeric, there were no blank/NaN entries that would make the CSV unbalanced, and there were no unexpected additional columns after reading the CSV into the notebook. Furthermore, the features underwent standard scaling to remove the mean and scales to unit variance as preparation for machine learning.

**Genomics.**

PBMCs were sent by each clinic to The New York Genome Center (NYGC) (<https://www.nygenome.org/>) for DNA extraction and sample QC. Whole-genome sequencing libraries were prepared, and sequencing was performed on an Illumina NovaSeq 6000 sequencer using 2X150 bp cycles. Sequence runs were assessed and only FASTQ data that were of high quality (exhibiting a 99.9% base call accuracy) were used for processing. Paired-end 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 marking of duplicate reads by the use of 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)22,23. The variant calls from NYGC were assessed by examining the actual reads for alignment issues and spot-checking the BAM files for specific variants in IGV determined to be of good quality. The VCFs were converted in to GVCFs, and joint genotyping calling was run using Sentieon v. 201911 (<https://www.sentieon.com/>), applied variant quality score recalibration (VQSR) was done using GATK v. 3.8 (truth sensitivity level = 99.0), and the files were annotated using Annovar v. 2018Apr1624 .

The NYGC developed an ancestry pipeline that estimates individual genome-wide average ancestries from a set of SNP genotypes using the ADMIXTURE tool, which is a maximum likelihood-based method. The pipeline takes a gVCF generated by Haplotype Caller as input, runs through a series of processing steps in PLINK, and passes the processed PLINK output to ADMIXTURE, which performs ancestry determination. The pipeline estimates ancestries for individual samples at the 1000 Genomes defined “super population” level, which are: AFR: African, AMR: Americas, EAS:  East Asian, EUR:  European, and SAS:  South Asian (<http://www.internationalgenome.org/category/population/>). Samples from the MXL (Mexican Ancestry from Los Angeles USA) and ASW (Americans of African Ancestry in SW USA) populations were excluded from the reference because they might be putatively admixed. The values range from 0–1 to represent the estimated fraction of each population to which the sample belongs.

The annotation pipeline incorporated elements from ANNOVAR25 and generated reports, including genotypes for all samples. These reports are available upon request. The following annotation was used: For genes and exonic variants that have clinical significance, the Clinical Genomic Database (CGD)26 , the Online Mendelian Inheritance in Man (OMIM)27, ClinVar28, and genes listed in the American College of Medical Genetics and Genomics (ACMG)29 database were incorporated. We also incorporated Intervar, which is based upon the ACMG and AMP standards and guidelines for interpretation of variants30-33. This tool uses 18 criteria to prescribe the clinical significance and classifies based on a five-tiered system34. To flag ALS genes, ALS gene lists and variants were incorporated from ALSoD35 (<http://alsod.iop.kcl.ac.uk/>), a list provided by Dr. Matthew Harms, a gene list from Dr. John Landers, and associations from DisGeNet36. Functional predictions were based on *in silico* prediction from nine databases: SIFT37, PolyPhen238-40 (HDIV and HVAR), LRT\_Prediction 40, Mutation Taster41, Mutation assessor 42, FATHMM prediction43-45, and dbNSFP (RadialSVM\_pred and LR\_pred)46-48. Databases that assess the variant tolerance of each gene using the RVIS49 and the Gene Damage Index (GDI)50 were also included, and LoFTool51 will be incorporated. To identify variants in genes that are highly expressed in the brain, data from the Human Protein Atlas52 (<http://www.proteinatlas.org>) and the GTEx portal53,54, (<https://gtexportal.org/home/>) for the cortex and spinal cord were used. Frequency information was derived from ExAC55, the NHLBI Exome Sequencing Project (ESP)56, and the 1000 Genomes Project57.

A separate annotation pipeline was developed for variants in intergenic and regulatory regions. Variants are reported relative the closest gene, whether intronic, upstream and downstream (up to 4 KBs from the start and stop of a gene) or in 5‘ and 3‘ UTRs. The annotation was based on RegulomeDB, which annotates variants with known or predicted regulatory elements such as transcription factor binding sites (TFBS), eQTLs, validated functional SNPs and DNase sensitivity58, with source data from ENCODE59,60 and GEO61. Additional regulatory databases such as Target Scan, an algorithm that uses 14 features to predict and identify microRNA target sites within mRNAs62, and miRBase63-65, were also used.

**Genomic Machine Learning.**

To calculate the probability that a person would develop ALS, and to find the variants and genes that are most predictive, we used single-nucleotide variants (SNPs) from the genomic data as input to machine learning models. Genotype data from the joint-genotyped, annotated VCF was transformed into numeric values (0, 1 and 2, depending on the number of alternate alleles that the sample has) and annotated using Annovar and our custom-based annotation pipeline as described above. A combination of variant filters was applied (using Python/Pandas) in an additive way to select specific variant types of interest, such as exonic, rare (minor allele frequency ≤ 0.01), low RVIS score, 6 or more in-silico tools damage prediction, etc. We also filtered for specific genes sets from WikiPathways66 or GO pathways67,68. Control data from 1000 genomes 57, which was recently sequenced on the NYGC platform with the same pipeline (tools and methods), were used since AALS lacks data from healthy controls. As the predominant ethnicity of the Answer ALS dataset is Caucasian, only the Caucasian samples from the 1000 genomes were used (CEU: Utah Residents with Northern and Western European Ancestry, TSI: Toscani in Italy, FIN: Finnish in Finland, GBR: British in England and Scotland, and IBS: Iberian Population in Spain). To assess accuracy of our models and avoid overfitting, we applied 10-fold cross-validation to all training and testing datasets. Samples were split into 10 random sets and, within each, 10% of the data were used for testing and the other 90% to train the models. Various machine learning algorithms were used to assess prediction power, including: Logistic Regression 69, Support Vector Machine/Classifier 70, Random Forests71,72, XGBoost73, Multinomial Naïve Bayes74,75, and Majority Voting76, all implemented by the Python library Scikit-Learn77.

**Data Portal**

The data portal provides users with information about the AALS program, the data, relevant terminology and data release notes. Users can download a metadata package associated with each versioned release. This versioned package contains comprehensive clinical, iPSC and inventory metadata. In addition, processes for enrolling patients, producing iPSC lines and performing Whole Genome Sequencing (WGS) are explained with links provided to the relevant facilities/institutions. Explanations for sample collection and analysis of Epigenomics, Proteomics and Transcriptomics data are available. Finally, precise definitions are provided for our data levels, which are ways to stratify all the various omics data coming from our analyses (Supplemental Methods Table 1).

**Supplemental Table 6: Data Level Definitions**

|  |  |
| --- | --- |
| **Data Levels** | **Description** |
| Level 1 | Immutable raw data |
| Level 2  | Raw data mapped to reference  |
| Level 3  | Most-processed sample-specific data |
| Level 4  | Analysis from joining Level 3 data for specific assay |

In addition, the data portal provides the user a means to access metadata, data and biosamples. The portal provides visual tools allowing researchers to find data by sample and participant features. Each sample is described by its omics assay, experiment type, sample name and subject ID. Samples can be removed from the visualization based on filter selection (e.g. filtering for only male patients). Once filters are selected, the user can download metadata or data associated with the filtered samples. For example, a researcher can retrieve metadata and data for patients who are older than 50 and have a known C9ORF72 mutation. Users are also able to find iPSC lines to order from Cedars-Sinai Biomanufacturing Center using the same filtering tools.

**Data Organization and Naming.**

The organization and naming of our data, regardless of data type, is an essential component of the program. We organize and name data products in a unified and systematic manner to allow a smooth end-user experience. A key component to data organization in our program is the usage of data levels. Data levels are a categorization schema to group similar types of omics data products together. Supplemental Methods Table 1 gives specific details on the data levels we have defined. Supplemental Methods Table 2 describes examples of these data levels in action with each experimental assay our program collects.

The AALS data program prefixes all data products in a systematic manner. The prefix consists of the following components: whether the sample is from a disease patient or healthy control patient, the de-identified patient GUID, the sample vial ID and the assay type abbreviation. An example of this is the raw Transcriptomics FASTQ file CASE-NEUAA599TMX-5310-T\_P10\_1.fastq.gz. The first underscore separates the prefix from any supplementary 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.

**Supplemental Methods Table 7: Examples of Data Levels for each assay**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assay** | **Level 1** | **Level 2** | **Level 3** | **Level 4** |
| Genomics | FASTQ | BAM | VCF | Joint VCF |
| Epigenomics | FASTQ | BAM | Peaks | Differential Peaks |
| Transcriptomics | FASTQ | BAM | Counts | Differential Counts |
| Proteomics | WIFF | MZML | Intensities | Differential Intensities |

**References: Supplemental Methods**

1 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

2 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

3 Milani, P. *et al.* Cell freezing protocol suitable for ATAC-Seq on motor neurons derived from human induced pluripotent stem cells. *Scientific reports* **6**, 25474, doi:10.1038/srep25474 (2016) 4857123

4 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008) 2592715

5 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010) 2898526

6 Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics* **10**, 48, doi:10.1186/1471-2105-10-48 (2009) 2644678

7 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, doi:10.1016/j.cell.2010.10.016 (2010) PMC3003590

8 Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat Methods* **9**, 676-682, doi:10.1038/nmeth.2019 (2012) PMC3855844

9 Yoneyama, T., Watanabe, T., Tamai, S., Miyashita, K. & Nakada, M. Bright spot analysis for photodynamic diagnosis of brain tumors using confocal microscopy. *Photodiagnosis Photodyn Ther* **25**, 463-471, doi:10.1016/j.pdpdt.2019.02.005 (2019)

10 Schwartzman, A., Jaffe, A., Gavrilov, Y. & Meyer, C. A. Multiple Testing of Local Maxima for Detection of Peaks in Chip-Seq Data. *Ann Appl Stat* **7**, 471-494, doi:10.1214/12-aoas594 (2013) PMC4233463

11 Atta-Fosu, T. *et al.* 3D clumped cell segmentation using curvature based seeded watershed. *J. Imaging* **2**, doi:10.3390/jimaging2040031 (2016) PMC5340274

12 Squarcina, L. *et al.* Fractal analysis of MRI data for the characterization of patients with schizophrenia and bipolar disorder. *Phys Med Biol* **60**, 1697-1716, doi:10.1088/0031-9155/60/4/1697 (2015)

13 Zhou, C., Wang, G., Huang, H., Song, L. & Xue, K. Edge detection based on joint iteration ghost imaging. *Opt Express* **27**, 27295-27307, doi:10.1364/OE.27.027295 (2019)

14 Kaso, A. Computation of the normalized cross-correlation by fast Fourier transform. *PLoS One* **13**, e0203434, doi:10.1371/journal.pone.0203434 (2018) PMC6147431

15 Dreiseitl, S. & Ohno-Machado, L. Logistic regression and artificial neural network classification models: A methodology review. *J. Biomed. Inform.* **35**, 352–359, doi:10.1016/s1532-0464(03)00034-0 (2002)

16 Rebentrost, P., Mohseni, M. & Lloyd, S. Quantum support vector machine for big data classification. *Phys Rev Lett* **113**, 130503, doi:10.1103/PhysRevLett.113.130503 (2014)

17 Gramatikov, B. I. Detecting central fixation by means of artificial neural networks in a pediatric vision screener using retinal birefringence scanning. *Biomed Eng Online* **16**, 52, doi:10.1186/s12938-017-0339-6 (2017) PMC5408446

18 Cao, Y., Hu, Z. D., Liu, X. F., Deng, A. M. & Hu, C. J. An MLP classifier for prediction of HBV-induced liver cirrhosis using routinely available clinical parameters. *Dis Markers* **35**, 653-660, doi:10.1155/2013/127962 (2013) PMC3834663

19 Breiman, L. Bagging predictors. *Machine Learning* **24**, 123-140, doi:10.1007/BF00058655 (1996)

20 Chen, T. & Guestrin, C. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, San Francisco, California, USA, 2016).

21 Littlestone, N. & Warmuth, M. K. The Weighted Majority Algorithm. *Information and Computation* **108**, 212-261, doi:<https://doi.org/10.1006/inco.1994.1009> (1994)

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

23 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

24 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164, doi:10.1093/nar/gkq603 (2010) PMC2938201

25 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010) 2938201

26 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) PMC3683745

27 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 Res* **43**, D789-798, doi:10.1093/nar/gku1205 (2015) PMC4383985

28 Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* **44**, D862-868, doi:10.1093/nar/gkv1222 (2016) PMC4702865

29 Green, R. C. *et al.* ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med* **15**, 565-574, doi:10.1038/gim.2013.73 (2013) PMC3727274

30 Richards, C. S. *et al.* ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genet Med* **10**, 294-300, doi:10.1097/GIM.0b013e31816b5cae (2008)

31 Kazazian, J., Boehm, C. D., & Seltzer, W. K. . ACMG recommendations for standards for interpretation of sequence variations. Genetics in Medicine. *Genetics in Medicine,* **2(5)**, 302-303., doi:doi.org/10.1097/00125817-200009000-00009 (2000)

32 Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17**, 405-424, doi:10.1038/gim.2015.30 (2015) PMC4544753

33 Li, Q. & Wang, K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *Am J Hum Genet* **100**, 267-280, doi:10.1016/j.ajhg.2017.01.004 (2017) PMC5294755

34 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)

35 Abel, O. *et al.* Development of a smartphone app for a genetics website: The amyotrophic lateral sclerosis online genetics database (ALSoD). *JMIR Mhealth Uhealth* **1**, e18, doi:10.2196/mhealth.2706 (2013) PMC4114449

36 Pinero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res* **45**, D833-D839, doi:10.1093/nar/gkw943 (2017) PMC5210640

37 Sim, N. L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res* **40**, W452-457, doi:10.1093/nar/gks539 (2012) PMC3394338

38 Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7 20, doi:10.1002/0471142905.hg0720s76 (2013) PMC4480630

39 Ramensky, V., Bork, P. & Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* **30**, 3894-3900, doi:10.1093/nar/gkf493 (2002) PMC137415

40 Sunyaev, S. R. *et al.* PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng* **12**, 387-394, doi:10.1093/protein/12.5.387 (1999)

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

42 Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* **39**, e118, doi:10.1093/nar/gkr407 (2011) PMC3177186

43 Shihab, H. A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* **34**, 57-65, doi:10.1002/humu.22225 (2013) PMC3558800

44 Shihab, H. A., Gough, J., Cooper, D. N., Day, I. N. & Gaunt, T. R. Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics* **29**, 1504-1510, doi:10.1093/bioinformatics/btt182 (2013) PMC3673218

45 Shihab, H. A. *et al.* Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Hum Genomics* **8**, 11, doi:10.1186/1479-7364-8-11 (2014) PMC4083756

46 Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat* **32**, 894-899, doi:10.1002/humu.21517 (2011) PMC3145015

47 Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum Mutat* **34**, E2393-2402, doi:10.1002/humu.22376 (2013) PMC4109890

48 Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat* **37**, 235-241, doi:10.1002/humu.22932 (2016) PMC4752381

49 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 Genet* **9**, e1003709, doi:10.1371/journal.pgen.1003709 (2013) PMC3749936

50 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) PMC4640721

51 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)

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

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

54 G. TEx Consortium. Human genomics. The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (New York, N.Y.)* **348**, 648–660, doi:10.1126/science.1262110 (2015) PMC4547484

55 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016) PMC5018207

56 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) PMC3708544

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

58 Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797, doi:10.1101/gr.137323.112 (2012) PMC3431494

59 Encode Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (New York, N.Y.)* **306**, 636-640, doi:10.1126/science.1105136 (2004)

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

61 Barrett, T. *et al.* NCBI GEO: Archive for high-throughput functional genomic data. *Nucleic Acids Res* **37**, D885-890, doi:10.1093/nar/gkn764 (2009) PMC2686538

62 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) PMC4532895

63 Griffiths-Jones, S. The microRNA Registry. *Nucleic Acids Res* **32**, D109-111, doi:10.1093/nar/gkh023 (2004) PMC308757

64 Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A. & Enright, A. J. miRBase: MicroRNA sequences, targets and gene nomenclature. *Nucleic. Acids Res.* **34**, D140–144, doi:10.1093/nar/gkj112 (2006) PMC1347474

65 Griffiths-Jones, S., Saini, H. K., van Dongen, S. & Enright, A. J. miRBase: Tools for microRNA genomics. *Nucleic. Acids Res.* **36**, D154–158, doi:10.1093/nar/gkm952 (2008) PMC2238936

66 Slenter, D. N. *et al.* WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res* **46**, D661-D667, doi:10.1093/nar/gkx1064 (2018) PMC5753270

67 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000) PMC3037419

68 The Gene Ontology, C. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* **47**, D330-D338, doi:10.1093/nar/gky1055 (2019) PMC6323945

69 Setakis, E., Stirnadel, H. & Balding, D. J. Logistic regression protects against population structure in genetic association studies. *Genome Res* **16**, 290-296, doi:10.1101/gr.4346306 (2006) PMC1361725

70 Noble, W. S. What is a support vector machine? *Nat Biotechnol* **24**, 1565-1567, doi:10.1038/nbt1206-1565 (2006)

71 Brieuc, M. S. O., Waters, C. D., Drinan, D. P. & Naish, K. A. A practical introduction to Random Forest for genetic association studies in ecology and evolution. *Mol Ecol Resour* **18**, 755-766, doi:10.1111/1755-0998.12773 (2018)

72 Sylvester, E. V. A. *et al.* Applications of random forest feature selection for fine-scale genetic population assignment. *Evol Appl* **11**, 153-165, doi:10.1111/eva.12524 (2018) PMC5775496

73 Ogunleye, A. A. & Qing-Guo, W. XGBoost Model for Chronic Kidney Disease Diagnosis. *IEEE/ACM Trans Comput Biol Bioinform*, doi:10.1109/TCBB.2019.2911071 (2019)

74 Ahmed, M. S., Shahjaman, M., Kabir, E. & Kamruzzaman, M. Prediction of Protein Acetylation Sites using Kernel Naive Bayes Classifier Based on Protein Sequences Profiling. *Bioinformation* **14**, 213-218, doi:10.6026/97320630014213 (2018) PMC6077816

75 Dou, Y., Guo, X., Yuan, L., Holding, D. R. & Zhang, C. Differential Expression Analysis in RNA-Seq by a Naive Bayes Classifier with Local Normalization. *Biomed Res Int* **2015**, 789516, doi:10.1155/2015/789516 (2015) PMC4538581

76 Hosni, M., Abnane, I., Idri, A., Carrillo de Gea, J. M. & Fernandez Aleman, J. L. Reviewing ensemble classification methods in breast cancer. *Comput Methods Programs Biomed* **177**, 89-112, doi:10.1016/j.cmpb.2019.05.019 (2019)

77 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)