scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Multi-omic atlas of the parahippocampal gyrus in Alzheimer's disease

Claire Coleman[1,2,3,4,5,14], Minghui Wang[1,6,14], Erming Wang[1,6,14], Courtney Micallef[1,2,3,4,5], Zhiping Shao[1,2,3,4,5], James M. Vicari[1,2,3,4,5], Yuxin Li[7,8,9], Kaiwen Yu[7,8,9], Dongming Cai[5,10,11,12], Junmin Peng[7,8,9], Vahram Haroutunian[2,11,13] ✉, John F. Fullard[1,2,3,4,5,15], Jaroslav Bendl[1,2,3,4,5,15], Bin Zhang[1,6,15] ✉ & Panos Roussos[1,2,3,4,5,10,11,15] ✉

Alzheimer's disease (AD) is the most common form of dementia worldwide, with a projection of 151 million cases by 2050. Previous genetic studies have identified three main genes associated with early-onset familial Alzheimer's disease, however this subtype accounts for less than 5% of total cases. Next-generation sequencing has been well established and holds great promise to assist in the development of novel therapeutics as well as biomarkers to prevent or slow the progression of this devastating disease. Here we present a public resource of functional genomic data from the parahippocampal gyrus of 201 postmortem control, mild cognitively impaired (MCI) and AD individuals from the Mount Sinai brain bank, of which whole-genome sequencing (WGS), and bulk RNA sequencing (RNA-seq) were previously published. The genomic data include bulk proteomics and DNA methylation, as well as cell-type-specific RNA-seq and assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) data. We have performed extensive preprocessing and quality control, allowing the research community to access and utilize this public resource available on the Synapse platform at https://doi.org/10.7303/syn51180043.2.

Background & Summary

Alzheimer's disease (AD) is the most common form of dementia worldwide, with a projection of 151 million cases by 2050, owing in part to our aging global population[1]. Since Dr. Alois Alzheimers' seminal discovery over a century ago, when he described "A peculiar severe disease process of the cerebral cortex", there have been a plethora of theories. However, clinical trials of disease-modifying treatments have been largely unsuccessful[2]. While forgetfulness and the loss of memory were always considered the first disease symptoms, spatial navigation and orientation deficits have been increasingly shown in preclinical AD as emerging cognitive biomarkers[3]. The parahippocampal gyrus (PHG) was reported as critical in spatial memory[4] and demonstrates various effects in Alzheimer's disease, including delay-dependent inaccuracy of memory-guided eye movements and poor

[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA. [2]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA. [3]Center for Disease Neurogenomics, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA. [4]Icahn Institute of Genomics, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA. [5]Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA. [6]Mount Sinai Center for Transformative Disease Modeling, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA. [7]Department of Structural Biology, St. Jude Children's Research Hospital, Memphis, TN, 38105, USA. [8]Department of Developmental Neurobiology, St. Jude Children's Research Hospital, Memphis, TN, 38105, USA. [9]Center for Proteomics and Metabolomics, St. Jude Children's Research Hospital, Memphis, TN, 38105, USA. [10]Department of Neurology, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA. [11]James J Peters VA Medical Center, Research & Development, Bronx, NY, 10468, USA. [12]Alzheimer Disease Research Center, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA. [13]Nash Family Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA. [14]These authors contributed equally: Claire Coleman, Minghui Wang, Erming Wang. [15]These authors jointly supervised this work: John F. Fullard, Jaroslav Bendl, Bin Zhang, Panos Roussos. ✉e-mail: bin.zhang@mssm.edu; panagiotis.roussos@mssm.edu

long-term spatial memory[5,6]. A number of studies have suggested a link between the PHG and AD, with the potential for MRI-measured atrophy of the PHG functioning as a biomarker for preclinical AD[7–9], while others propose that such cognitive impairments may not yet be present in cases of preclinical AD[10].

Next-generation sequencing (NGS) is an example of non-clinical research that has enhanced our understanding of AD. Since the development of NGS, multiple genes implicated in AD risk and pathogenesis have been identified. Early NGS studies focused on performing deep resequencing of established early-onset AD genes, namely amyloid precursor protein (APP), presenilin 1 (PSEN1), and presenilin 2 (PSEN2), all extremely rare and accounting for less than 5% of cases[11]. More recent whole genome sequencing (WGS) studies focused on both rare and common risk variants for the more complex, and also more common, late-onset type of AD[12–15]. Due to the inherent difficulty in obtaining fresh specimens, most molecular studies of the human brain are restricted to frozen post-mortem samples. Working with frozen material is not without its challenges, including the loss of cytoplasm (and, with it, many cell-specific antigens) as a consequence of freeze-thawing. Nevertheless, an increasing number of human brain studies have employed cell-type specific nuclear markers to isolate nuclei of interest via Fluorescence-Activated Nuclear Sorting (FANS)[16–18]. In our recent study, FANS was utilized to isolate neuronal and non-neuronal samples (using an Anti-NeuN antibody) from AD cases and controls to identify cell-specific epigenetic changes associated with AD progression[19]. Here, we enlarged the panel of antibodies used by including SOX10, to further sort non-neuronal samples into oligodendrocytes (NeuN-/Sox10 +) and microglia/astrocytes (NeuN-/Sox10-). In total, we have generated 124 cell-specific transcriptome samples (FANS-sorted RNA-seq), 110 cell-specific epigenome samples (FANS-sorted ATAC-seq) as well as 196 bulk DNA methylome samples and 185 bulk proteome samples. The newly generated data sets expand the panel of genomic assays in the Mount Sinai Brain Bank AD cohort (MSBB-AD)[20] and increase cell-specific resolution (Fig. 1).

Methods

Cohort data collection. In this study, we generated a multi-omics data set from the frozen brain tissue of 201 subjects of predominantly European ancestry (Fig. 2a) obtained from the Mount Sinai NIH Neurobiobank[20]. All neuropsychological, diagnostic and autopsy protocols were approved by the Mount Sinai and JJ Peters VA Medical Center Institutional Review Boards[20]. Extensive cognitive and neuropathological assessment as well as demographic data were already available for all donors[21]. For bulk proteome and DNA methylome data, we generated the samples for all 201 donors. For RNA-seq transcriptome and ATAC-seq epigenome data, we selected only 42 donors (21 AD cases and 21 controls with either no discernable neuropathology or cognitive complaints) and performed FANS to generate 3 cell type specific samples per donor. The donors were selected to represent the full spectrum of clinical and pathological severity based on the following phenotypes: (1) case–control status defined using the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) criteria[22], i.e. 1 = normal, 2 = definite AD, 3 = probable AD, and 4 = possible AD; (2) Braak AD-staging score for the progression of neurofibrillary neuropathology (Braak and Braak score[23,24]); (3) mean density of neuritic plaques (PlaqueMean); and (4) assessment of dementia based on the Clinical Dementia Rating scale (CDR)[25].

Bulk proteomics data. *Tandem mass tag assays.* The post-mortem brain samples were lysed in the fresh lysis buffer (50 mM HEPES, pH 8.5, 8 M urea, and 0.5% sodium deoxycholate) with the established protocol[26]. Protein concentrations of the lysates were measured by the BCA assay (Thermo Fisher Scientific) and further validated by short SDS Coomassie-stained gels[27]. Approximately 0.1 mg of quantified proteins in the lysis buffer with 8 M urea were digested in two steps: first with Lys-C (Wako, 1:100 w/w) at 21 °C for 2 h, and then with 4-fold dilution to reduce urea to 2 M followed by trypsin digestion (Promega, 1:50 w/w) at 21 °C overnight. The resulting peptide samples were acidified, desalted with Sep-Pak C18 cartridge (Waters), and then dried. These samples were re-dissolved in 50 mM HEPES (pH 8.5) for TMT reaction for 30 mins, and equally pooled. The pooled samples were desalted and fractionated by offline basic pH reverse phase LC (an XBridge C18 column of 3.5 μm particle size, 4.6 mm × 25 cm, Waters), and each collected fraction was then analyzed by the acidic pH reverse phase LC coupled with MS/MS analysis[28]. The fractions were analyzed sequentially on a C18 column (75 μm × 15–30 cm, 1.9 μm resin from Dr. Maisch GmbH, 65 °C to reduce backpressure) coupled with a Q Exactive HF Orbitrap mass spectrometer (Thermo Fisher Scientific). In mass spectrometer (MS) settings, positive ion mode and data-dependent acquisition were applied with one full MS scan followed by 20 MS/MS scans. MS1 scans were acquired at a resolution of 60,000, 1E6 AGC and 50 ms maximal ion time. After ion fragmentation with higher energy collision-induced dissociation (HCD, ~35% normalized collision energy and ~1.0 m/z isolation window with 0.3 m/z offset), MS2 spectra were acquired at a resolution of 60,000, fixed first mass of 120 m/z, 410–1600 m/z, 1E5 AGC, 100–150 ms maximal ion time, and ~15 sec of dynamic exclusion.

Computational processing. Tandem Mass Tag (TMT)-based proteomics analysis was utilized to profile protein expression abundance. Proteomics data analysis was performed as previously described[29]: the JUMP search engine[30] was used to search MS/MS raw data against a composite target/decoy database[31] to evaluate FDR. The protein database was generated by combining downloaded SwissProt, TrEMBL, and UCSC databases and removing redundancy (83,955 entries for human proteins), followed by concatenation with a decoy database. Major parameters included 15 ppm mass tolerance for precursor ions and 10 ppm for product ions, full tryptic-ity, static modification of the TMT tags (+229.162932 Da) on Lys residues and peptide N termini and carbamidomethyl modification on cysteine (+57.02146 Da), dynamic modification for Met oxidation (+15.99492 Da), maximal miscleavage sites ($n = 2$), and maximal modification sites ($n = 3$). The resulting PSMs were filtered by precursor ion mass accuracy and minimal search score, and then grouped by peptide length, tryptic ends, modifications, miscleavage sites, and precursor ion charge state followed by the cutoffs of JUMP-based matching scores (Jscore and ΔJn) to reduce FDR below 1% for proteins. If one peptide could be generated from multiple

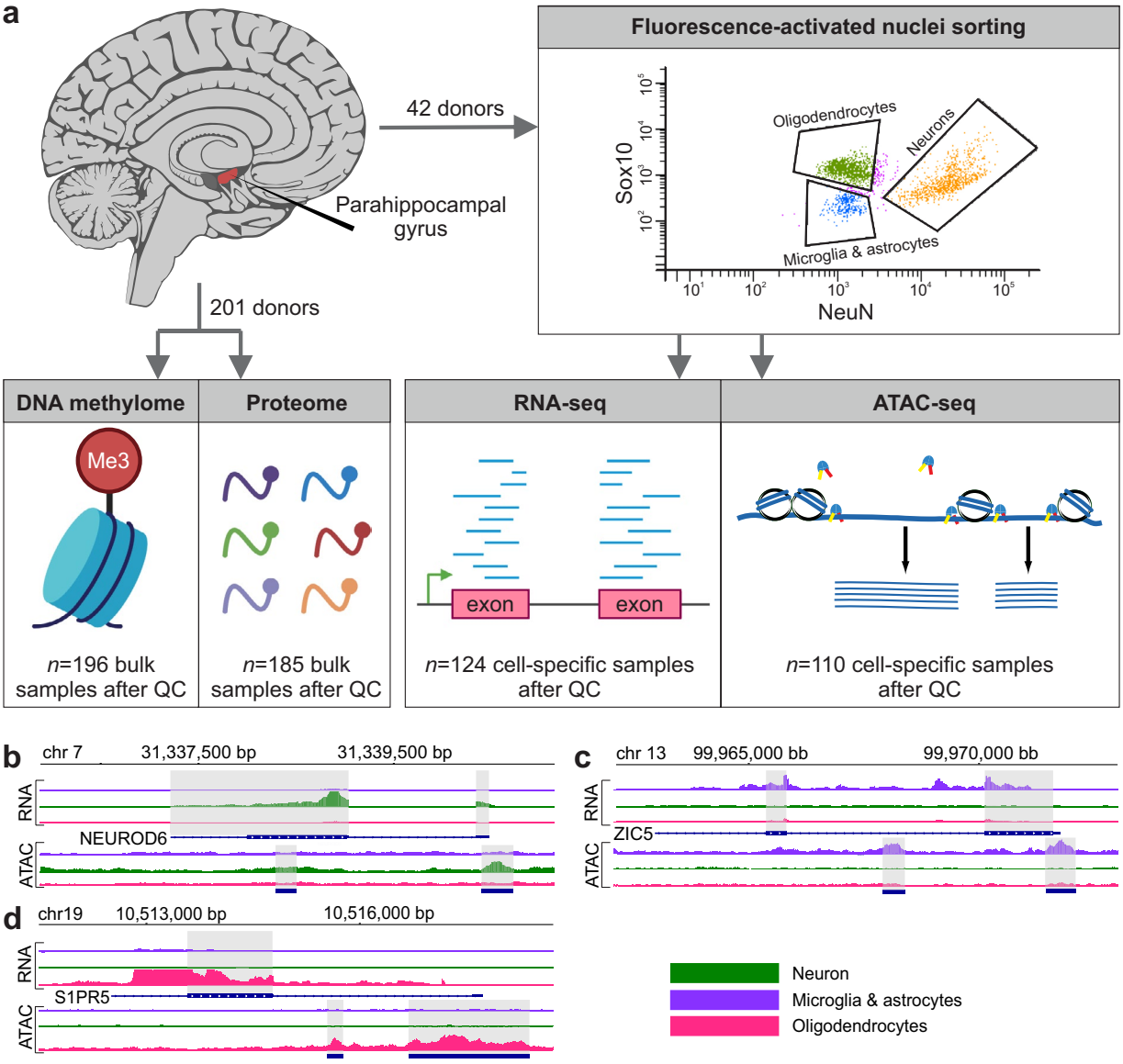

Fig. 1 Study design and examples of transcriptome and epigenome landscape around three selected cell type markers. (**a**) Study design: Dissections from PHG brain region of AD case and control subjects were obtained from frozen human postmortem tissue. Nuclei were subjected to fluorescence-activated nuclear sorting to yield three cell populations, followed by RNA-seq and ATAC-seq profiling and subsequent downstream analyses to perform quality control and identify cell type-specific open chromatin regions. (**b**–**d**) Examples of cell-specific chromatin accessibility and gene expression for three cell type markers, i.e. (**b**) NEUROD6 (neurons), (**c**) ZIC5 (microglia & astrocytes) and (**d**) S1PR5 (oligodendrocytes).

homologous proteins, based on the rule of parsimony, the peptide was assigned to the canonical protein form in the manually curated SwissProt database. If no canonical form was defined, the peptide was assigned to the protein with the highest PSM number.

Proteins were quantified in the following steps, similar to previous reports[28,29,32]: (i) TMT reporter ion intensities of each PSM were extracted; (ii) the raw intensities were corrected according to isotopic distribution of each labeling reagent; (iii) PSMs with very low reporter ion intensities were excluded (e.g. minimum intensity < 1,000 and median intensity < 5,000); (iv) sample loading bias was corrected by normalization with the trimmed median intensity of all PSMs; (v) the mean-centered intensities across samples were calculated; (vi) protein relative intensities were summarized by averaging related PSMs; (vii) protein absolute intensities were derived by multiplying the relative intensities by the grand-mean intensity of the top three most highly abundant PSMs. In addition, we performed y1-ion based correction of TMT data[28]. Data QC based on protein quantification identified one batch ("batch 20") as an outlier batch and was thus discarded. To generate a combined quantification table for multiple batches, a common sample (mixture of multiple samples) was included in each batch as an internal standard. MS intensities from different batches were normalized according to this internal standard.

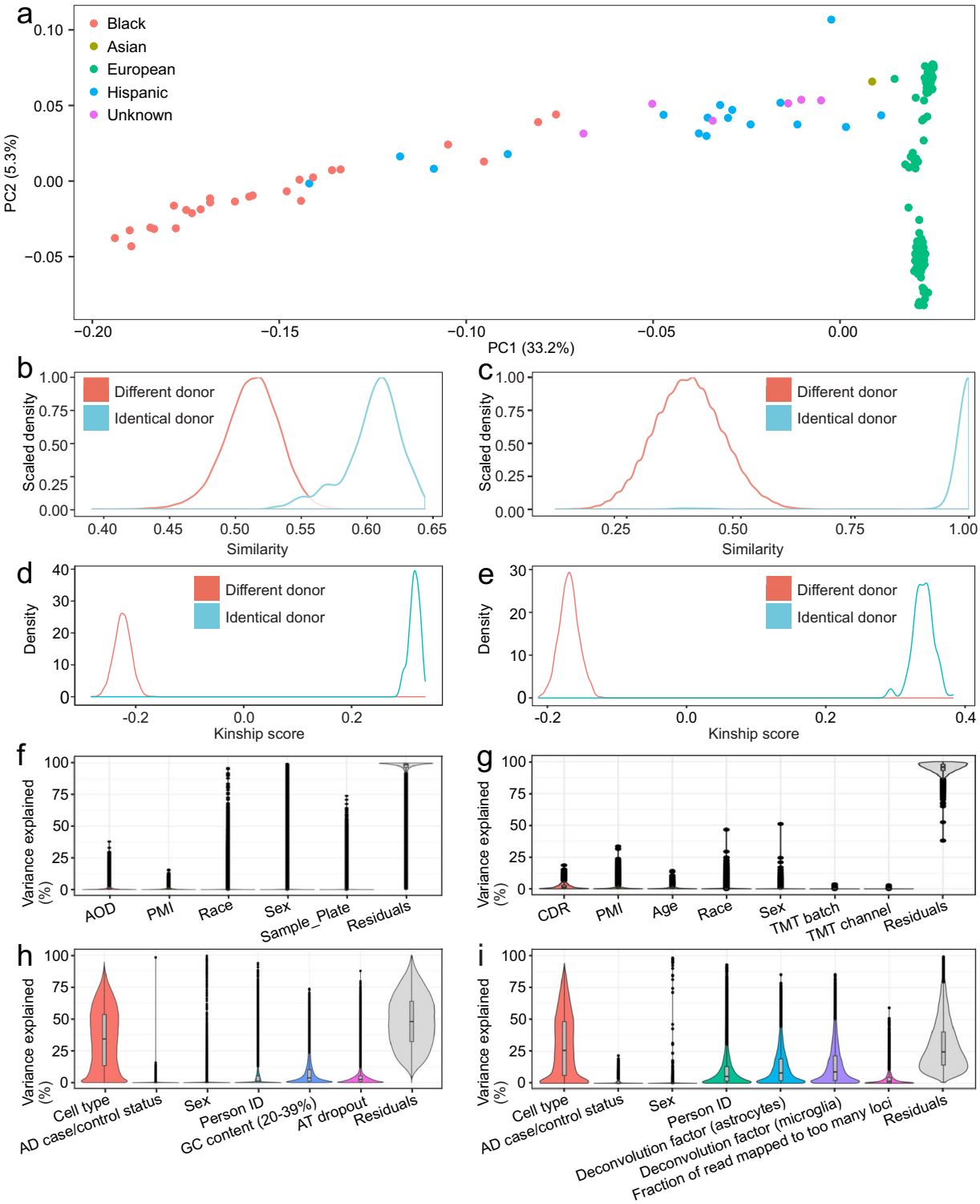

Fig. 2 Analysis of cohort ancestry, genetic similarity and assay variance. (**a**) Top two principal components of per-sample genetic ancestry estimation. (**b**–**e**) Distribution of genetic similarities estimated between WGS samples and proteomics samples (**b**), DNA methylation samples (**c**), RNA-seq samples (**d**), ATAC-seq samples (**e**). Colors denote whether the sample pairs are originating from the same or different brain donors. (**f**–**i**) Variance explained by biological and technical covariates for DNA methylation sites (**f**), proteins (**g**), ATAC-seq genes (**h**) and RNA-seq peaks (**i**).

Bulk DNA methylation. *Methylation array assays.* Total genomic DNA was isolated from 10 mg of post-mortem brain tissue dissected from the PHG region, using the Qiagen All Prep DNA/RNA Mini Kit, according to the manufacturer's instructions (Qiagen, catalog# 80204). Tissues were first homogenized using Qiagen's

TissueLyzer LT (Qiagen, catalog# 69980) combined with 5 mm stainless steel beads (Qiagen, catalog# 69989). Next, the lysed tissues were loaded onto QIAmp spin columns to wash off any impurities. Purified DNA was eluted off from the columns using a low salt buffer. 200–500 ng DNA per sample were used for bisulfite conversion using the EZ-96 DNA Methylation-Lightning Kit (Zymo, catalog# D5033). Next, DNA samples were fragmented and hybridized to Infinium MethylEPIC BeadChips[33] (Illumina, catalog# WG-317-1001). Lastly, hybridization signals were obtained through the Illumina iScan microarray scanner (Illumina, catalog# SY-202-1001).

Methylation data preprocessing. We streamlined a workflow to preprocess, normalize, and quality check (QC) Illumina 850 K methylation array data. We essentially followed the pipeline as described by Maksimovic *et al.*[34]. In brief, the R package "minfi" was utilized to preprocess and normalize the raw array data in IDAT format. We utilized the functions "read.metharray.sheet" and "read.metharray.exp" to import the sample metadata and the raw methylomic data into R, respectively, and the function "preprocessQuantile" for data normalization, and the functions "getBeta" and "getM" to calculate and output both β and M values of the 866,029 probes on the platform, respectively. For further quality control, we discarded CpG probes that were either internal control, or with low quality (detection p-value < 0.05)[34], or known to overlap with common SNPs at the same CpG sites[34]. The genic annotation of the CpG sites was obtained from the annotation package for Illumina's EPIC methylation arrays, i.e., IlluminaHumanMethylationEPICanno.ilm10b2.hg19. After removing 5 potentially mismatched samples, as described in the Technical Validation section, the β and M values of the QCed CpG probes were corrected for co-variables including batch.

Cell-type-specific RNA sequencing. *Fluorescence-activated nuclei sorting.* From each dissection, 250 mg of frozen brain tissue was homogenized in a cold lysis buffer (0.32 M Sucrose, 5 mM CaCl2, 3 mM Magnesium acetate, 0.1 mM, EDTA, 10 mM Tris-HCl, pH8, 1 mM DTT, 0.1% Triton X-100) and filtered through a 40 µm cell strainer. The flow-through was underlaid with sucrose solution (1.8 M Sucrose, 3 mM Magnesium acetate, 1 mM DTT, 10 mM Tris-HCl, pH 8) and subjected to ultracentrifugation at 24,000 rpm for 1 hour at 4 °C. Pellets were resuspended in 500 µl DPBS and incubated in BSA (final concentration 0.1%) and anti-NeuN antibody (1:600, Alexa 647 conjugated, abcam Cat #ab190565) and anti-SOX10 (1:300, Alexa488 conjugated, R&D Systems Cat #IC28642G). Prior to FANS sorting, DAPI (Thermoscientific) was added to a final concentration of 1 µg/ml. Neuronal (DAPI + NeuN + SOX10-), oligodendrocytes (DAPI + NeuN- SOX10 +) and microglia/astrocytes (DAPI + NeuN- SOX10-) nuclei were sorted into individual tubes (pre-coated with 5% BSA) using a FACSAria flow cytometer (BD Biosciences).

Library preparation and sequencing. For RNA-seq, nuclei were sorted into 1.5 ml low-bind microfuge tubes containing Extraction buffer, a component of the PicoPure RNA Extraction kit (Arcturus, Ca t# KIT0204). RNA was isolated in accordance with the PicoPure RNA Isolation kit's manufacturer's instructions. This included an RNase-free DNase treatment step (Qiagen, Cat # 79254). Samples were eluted in RNase-free water and stored at −80 °C until preparation of RNA-Sequencing libraries using the SMARTer Stranded Total RNA-Seq Pico Kit v1 or v2 (Takara Clontech Laboratories, Cat # 635005 or 634414, respectively), according to the manufacturer's instructions. Following the construction of the RNA-seq libraries, libraries were analyzed on a TapeStation using a High Sensitivity D1000 ScreenTape (Agilent, Cat # 5067-5584) and quantification of the libraries was performed using the KAPA Library Quantification Kit. Libraries that passed QC were sequenced on Hi-Seq2500 (Illumina) obtaining 2×50 paired-end reads.

Computational processing. The raw reads were trimmed with Trimmomatic (v0.36)[35] and then mapped to human reference genome hg38 using STAR (v2.5.3a)[36]. Following read alignment, expression quantification was performed at the transcript isoform level using RSEM (v1.3.0)[37] and then summarized at the gene level. Gene quantifications correspond to GENCODE (v30)[38]. Gene count matrix was normalized by the trimmed mean of M-values (TMM)[39] and filtered to keep only genes with over 1 count per million in at least 30% of the samples. RNA-SeqQC (v1.1.7)[40] and Picard (v2.2.4) were used to generate quality control metrics. Quality control processes (described in Technical validation) removed 2 samples, resulting in a final count matrix of 124 samples by 21,383 genes. To correct for unwanted technical variance, we applied the step-wise covariate analysis based on the Bayesian information criterion (BIC)[41]. As a starting point for this analysis, a base model was chosen with the variables "cell_type by diagnosis_status" and "sex". Then, it was tested, for each additional covariate, how many genes showed an improved BIC score minus how many showed a worse BIC score when the covariate was included in the linear regression model compared to when it wasn't. A covariate was then required to improve the mean BIC per gene by at least 5 for it to be included in the final model. This model selected 3 covariates: "reads_mapped_to_too_many_loci" (i.e. fraction of discarded reads by STAR aligner; this serves as a proxy to the technical quality of the sample) and two deconvolution metrics, i.e. predicted proportion of microglia and astrocytes in each sample, thus compensating for limitations of our experimental design that sorted cells from microglia and astrocytes together. The effect of those three technical covariates was regressed out to generate the normalized count matrices.

Cell-type-specific ATAC sequencing. *Generation of ATAC-seq libraries and sequencing.* ATAC-seq libraries were generated from cell-sorted brain tissue dissection (see the section "Fluorescence Activated Nuclei Sorting" for RNA-seq) using an established protocol[42] with minor modifications. In brief, 100,000 sorted nuclei were centrifuged at 500 g for 10 min at 4 °C. Pellets were resuspended in transposase reaction mix (25 µL 2x TD Buffer (Illumina Cat # FC-121-1030) 2.5 µL Tn5 Transposase (Illumina Cat # FC-121-1030) and 22.5 µL Nuclease Free H_2O) on ice. Reactions were incubated at 37 °C for 30 min and then purified using the MinElute Reaction

Cleanup kit (Qiagen Cat # 28204), eluting in 10 µL of buffer EB. Following purification, library fragments were amplified using the Nextera index kit (Illumina Cat # FC-121-1011) under the following cycling conditions: 72 °C for 5 minutes, 98 °C for 30 seconds, followed by thermocycling at 98 °C for 10 seconds, 63 °C for 30 seconds, and 72 °C for 1 minute for a total of 5 cycles. To prevent saturation due to over-amplification, a 5 µl aliquot was then removed and subjected to qPCR for 20 cycles to calculate the optimal number of cycles needed for the remaining 45 µL reaction. The additional number of cycles was determined by first plotting linear Rn vs. Cycle and secondly calculating the cycle number corresponding to 1⁄4 of maximum fluorescence intensity. In general, adding 4-6 cycles to this estimate was found to yield optimal libraries, as determined by analysis on Tapestation D5000 ScreenTapes (Agilent technologies Cat # 5067-5588). Libraries were then resolved on 2% agarose gels and fragments ranging in size from 100-1000 bp were excised and purified (Qiagen Minelute Gel Extraction Kit – Qiagen Cat # 28604). Prior to sequencing, libraries were quantified with the Qubit dsDNA HS assay kit (Invitrogen Cat # Q32851) and by quantitative PCR (KAPA Biosystems Ca # KK4873), and fragment sizes estimated using Tapestation D5000 ScreenTapes (Agilent technologies Cat # 5067-5588). Libraries that passed QC were normalized for concentration and sequenced on Hi-Seq2500 (Illumina), obtaining 2 × 50 paired-end reads.

Computational processing. Similar to RNA-seq processing, trimming of low-quality base pairs and adapter sequences was performed by Trimmomatic (v0.36)[35]. Then, reads were mapped to hg38 by STAR (v2.7.0)[36]. Reads mapped to multiple loci, mitochondrial genome or duplicate reads were removed by samtools (v0.1.19)[43]. To increase the sequencing depth for peak calling, all samples were downsampled to the same size and, then, merged into three separate BAM files by their cell type identity. Cell-specific peaks were called by MACS2[44] and merged into a final consensus of 263,265 peaks. Peaks overlapping ENCODE blacklisted regions[45] were removed. The peak count matrix was normalized by the trimmed mean of M-values (TMM)[39] and filtered to keep only peaks with over 1 count per million in at least 20% of the samples. Picard (v2.2.4) and phantompeakqualtools (v2.0) were used to generate quality control metrics. Quality control processes (described in Technical validation) removed 16 samples, resulting in a final count matrix of 110 samples by 257,336 peaks. Then, we applied the same covariate selection model utilizing repeated BIC model as for RNA-seq, This model selected two covariates: "GC_coverage_20-39" (i.e., normalized read coverage over each quintile of GC content ranging from 20 – 39%) and "AT_dropout" that improved a net of 68.5% and 21.5% of peaks. The effect of those two technical covariates was regressed out to generate the normalized count matrices.

Data Records

All data described herein are available for use by the research community and have been deposited in the AMP-AD Knowledge Portal in study-specific folder[46]. These include sample metadata[21], as well as raw and processed sequencing data for ATAC-seq, RNA-seq, proteome and DNA methylation[46].

Technical Validation

Bulk proteomics data quality control. We performed sample alignment between proteomics data and matched WGS[47] data from the same cohort using two different strategies. In the first strategy, we utilized a proteogenomics approach to first identify sample-specific peptides with mutations, followed by proteogenomics-based genotype inference and sample alignment using the SMAP software[48]. Briefly, by constructing a customized protein database using SNVs detected from WGS data, peptides with sample-specific mutations were identified using the JUMPg software[49]. The resulting peptides were quantified and processed by SMAP for sample alignment with two steps: (i) inference of sample-specific genotype based on TMT-based quantification while taking the genotype dosage information in the WGS data as prior knowledge; and (ii) sample verification and correction by comparing the inferred genotypes versus the mutation profiles of the matched WGS sample. Five proteomics samples were identified to be potentially mislabeled. In the second strategy, we utilized the software MODMatcher[50]. Briefly, the normalized proteomics data were corrected for TMT batch using a random effect regression model by R package variancePartition[51], and subsequently corrected for covariates including PMI, age, race, and sex using linear regression (Fig. 2g). Then protein quantitative trait loci (pQTLs) were computed with R package MatrixEQTL[52] by integrating covariates-corrected proteomics data with the WGS data. Genotypes at the most significant cis-pQTL of each cis-pQTL bearing protein were imputed from the protein expression data using an algorithm developed in the software MODMatcher[50]. Genotype consistency was computed for all possible sample pairs between the imputed genotype data from proteomics and the observed genotype data from WGS (Fig. 2b). Following MODMatcher[50], a proteomics sample was considered self-aligned with the corresponding WGS sample if its same-donor WGS sample was among the top 3 matches ranked by the genotype consistency score. Meanwhile, best-matched proteomics samples for each WGS sample were also identified based on the genotype consistency score. As a result, 183 proteomics samples were self-aligned. Among the 7 proteomics samples that were not self-aligned with WGS, 4 were considered potentially mislabeled as each showed a reciprocal best match with a WGS sample from a different donor. Notably, all these 4 proteomics samples were among the mislabeled samples detected by JUMPg. Therefore, we corrected the donor identifiers for the 4 mislabeled samples and discarded the remaining problematic samples (total 4) detected by either JUMPg or MODMatcher. One mislabeled sample became a duplication after label correction and hence was discarded as well. Lastly, the retained normalized proteomics data ($n = 185$) with properly matched WGS data were corrected for covariates including TMT batch, PMI, age, race, and sex.

Bulk DNA methylation data quality control. To assure the high quality of the DNA methylation data, we first evaluated if the DNA methylation samples can be properly aligned to their corresponding WGS samples. For this purpose, we carried out the genotype inference on 59 control probes querying high-frequency SNPs by

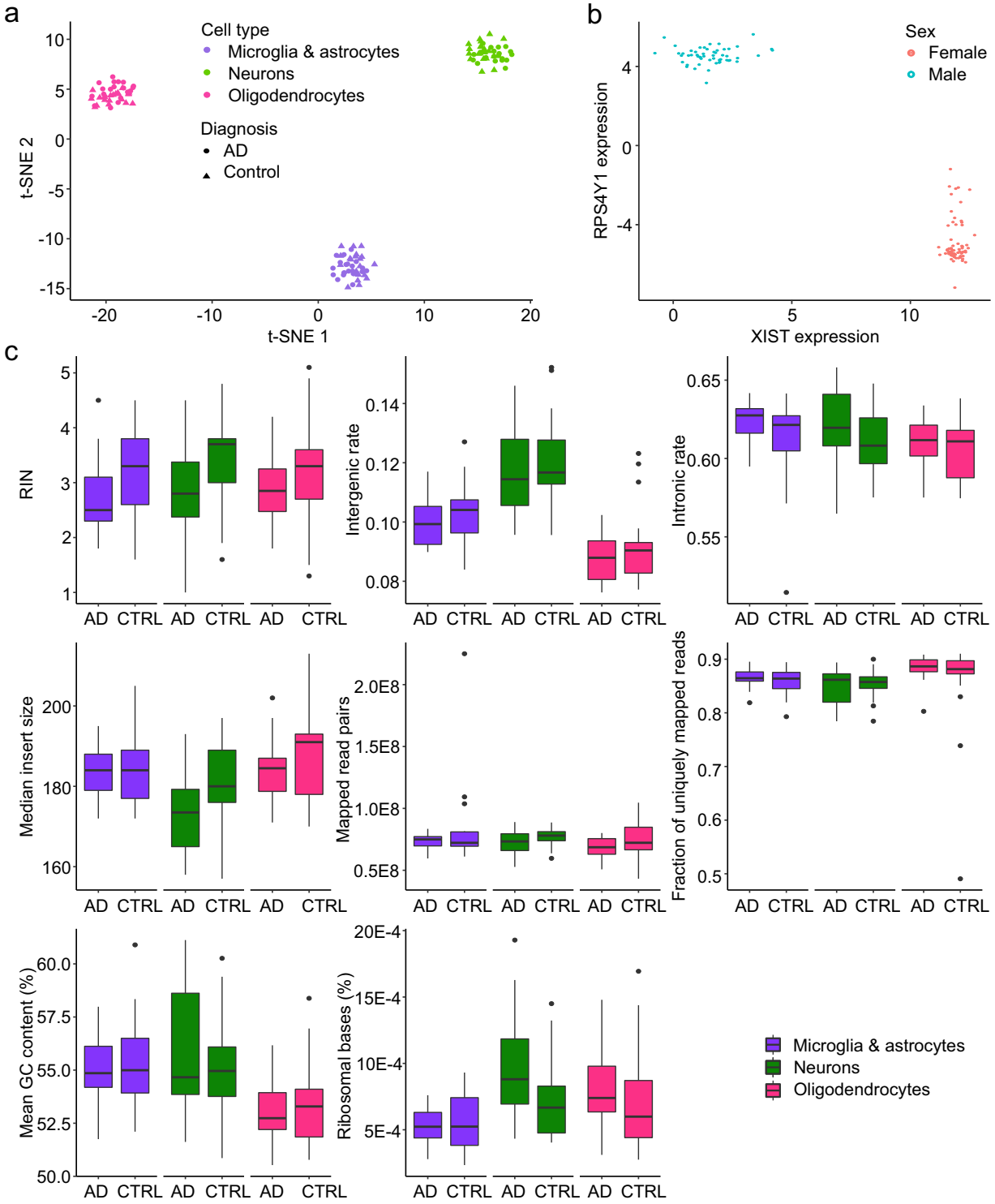

Fig. 3 Analysis of RNA-seq dataset. Diagnosis groups are defined by CERAD metrics, i.e. AD: CERAD = (2-4), CTRL: CERAD = 1. **(a)** t-SNE clustering. **(b)** Sex check based on quantification of the expression of male- (RPS4Y1) and female-specific genes (XIST). **(c)** RNA-seq quality control metrics stratified by disease status and cell subtype: RNA integrity number (RIN), intergenic rate, intronic rate, median insert size, counts of mapped read pairs, percentage of uniquely mapped reads, mean GC content and percentage of ribosomal bases. t-test comparison on the distributions of values of AD cases and controls for all QC metrics in three cell types revealed that only 2 metrics in 2 cell types are different before correction for multiple testing but not after FDR correction, i.e. the percentage of ribosomal bases in neurons (p-value = 0.030, FDR-corrected q-value = 0.090) and the number of mapped reads in oligodendrocytes (p-value = 0.048, FDR-corrected q-value = 0.144). Box plots are centered on median, bounds defined between the 25th and 75th percentile with minimum and maximum defined as median ± 1.5 × interquartile range, and whiskers extending to the lowest/highest value within this range.

Illumina's EPIC chip. Following a prior practice[53], a mixed model assuming distinct hybridization signal distribution for different genotypes was trained to predict sample genotypes for each of these probes. Subsequently, a genotype concordance score was computed by comparing the inferred genotypes with the WGS-based genotypes. While the majority of the DNA methylation samples showed a high genetic concordance with their corresponding WGS samples (genotype similarity score close to 1), 5 methylation samples exhibited a low genotype concordance with their respective WGS counterparts (genotype concordance score < 0.9) and were hence labeled as mismatched samples and discarded from the analysis (Fig. 2c).

RNA-seq quality control. A total of 126 RNA-seq samples were integrated into a single analysis across all cell types and AD case/control status in order to perform joint quality control. Dimensionality reduction techniques calculated on gene count matrix were used to confirm the successful clustering of samples by cell types, with the exception of two outlying samples that were excluded (laboratory notes indicated that those two samples yielded very low concentrations of RNA, indicative of low tissue quality) (Fig. 3a; removed samples not shown). The remaining 124 samples had acceptable values for the following RNA-seq quality control metrics: RNA integrity number (RIN) (mean 3.1, sd $\pm$ 0.84), intergenic rate (mean 10.3%, sd $\pm$ 1.6%), intronic rate (mean 61.4%, sd $\pm$ 2.2%), ribosomal RNA rate (mean 0.07%, sd $\pm$ 0.03%), mapped read pairs (mean 75×10^6, sd $\pm 1.7 \times 10^6$), percentage of uniquely mapped reads (mean 86.1%, sd $\pm$ 4.5%), median insert size (mean 182 bp, sd $\pm$ 10 bp) and mean GC content (mean 54.6%, sd $\pm$ 2.3%) (Fig. 3c). In order to confirm that the expression of genes on sex chromosomes is consistent with the reported sex, RPS4Y1 and XIST were selected as representatives of sex-specific genes, and all samples showed distinct clustering by reported sex (Fig. 3b). To verify donor identity of all samples, we used kinship coefficient from KING v1.913[54] to compare the per-sample variants called from raw sequencing reads to the variants from existing WGS reference[47]. All detected swaps between samples were corrected by sample re-labeling, however, two potentially contaminated samples with low similarity to all genotypes, including the expected genotype, were excluded. After performing all steps of genotype concordance analysis, we observed clear and unambiguous separation of $n = 124$ samples from 21 donors (Fig. 2d).

ATAC-seq quality control. Similar to RNA-seq quality control, we performed a joint analysis of all 122 sequenced ATAC-seq samples to detect outlying and low-quality samples. All samples passed our QC metrics criteria for minimum mappability (more than 50% required), minimum fraction of reads in peaks (more than 4% required) and maximum fraction of reads mapped to the mitochondrial genome (less than 3% required). However, 6 samples were removed due to the low signal-to-noise ratio as we required more than 3,000 narrow peaks per sample. An additional 6 samples were removed due to low cell type specific signal detected by clustering analysis and visually confirmed in IGV by looking at open chromatin accessibility signal within promoters of cell-specific genes. After completion of QC steps, the remaining samples showed clear cell type separation (Fig. 4a). The remaining 110 samples had acceptable values for the following ATAC-seq quality control metrics: number of narrow peaks called per sample (mean 32,906, sd $\pm$ 18,170), the fraction of reads in peaks (mean 13%, sd $\pm$ 2.9%), the fraction of reads mapping to the mitochondrial genome (mean 0.98%, sd $\pm$ 0.4%), median insert size (mean 116 bp, sd $\pm$ 24 bp), the fraction of reads that were uniquely mapped (mean 0.865, sd $\pm$ 0.029), mean GC content (mean 46.4%, sd $\pm$ 1.2%), the number of uniquely mapped reads (mean 59×10^6, sd $\pm 9.3 \times 10^6$) and the fraction of duplicated reads (mean 0.117, sd $\pm$ 0.029%) (Fig. 4c). We also carried out sex check by comparing per-sample numbers of all mapped reads versus chromosome Y reads, confirming distinct clustering by reported sex (Fig. 4b). Lastly, we checked the identity of a final set of $n = 110$ samples from 21 donors using the same approach as explained for RNA-seq data and corrected all swaps and mislabelings (Fig. 2e).

Usage Notes

As a usage example, here we summarize the analytic flow and the key findings from our recent publication[55] in which we integrated the multi-omics data in the MSBB-AD cohort which was developed through a previous study[55] and this current study (termed as the discovery cohort) and the Religious Orders Study and Memory and Rush Aging Project[56,57] (ROSMAP) cohort with multiomics data from the dorsolateral prefrontal cortex (the validation cohort). The discovery cohort includes matched epigenomic (ATAC-seq), methylomic, transcriptomic (RNA-seq) and proteomic data from the PHG, as described in the previous sections in this paper, while the validation cohort includes methylomic, transcriptomic (RNA-seq) and proteomic data, along with ATAC-seq data and H3K9ac domain atlas in the prefrontal cortex (PFC) region[58] (Fig. 5a).

As shown in Fig. 5b, we first identified AD-associated methylomic changes by computing differentially methylated probes and differentially methylated regions (DMRs). In the MSBB AD, 270 DMRs were found to be not only associated with AD clinical and pathological traits cohort and the expression levels of many genes and proteins differentially expressed between AD and controls but also enriched for known AD GWAS risk genes and the Aβ pathways. To model and quantify the overall effect of DNA methylation on individual genes and proteins, we developed a novel statistic, termed overall methylation score (OMS)[55], and revealed that in the gene or protein co-expression network modules which were most strongly associated with AD, their member genes or proteins generally had a high amplitude of OMS that was also correlated with the respective gene/protein expression changes between AD and controls. We also found that, in the Bayesian causal networks, the top-ranked key drivers tended to be regulated by methylation. Finally, to investigate the causal relationship between DMRs and ATAC peaks on gene expression, the causal inference test (CIT)[59] was performed on DMRs, ATAC peaks, and associated genes or proteins. Our analysis identified thousands of significant causal chains with a relationship of DMR → ATAC → gene/protein, but none of the relationship of ATAC → DMR → gene/protein, suggesting that DMRs likely influenced gene/protein expression via ATAC peak domains in AD, rather than ATAC peak domains influenced gene/protein expression via DNA methylation. In summary, our integrative analysis of the

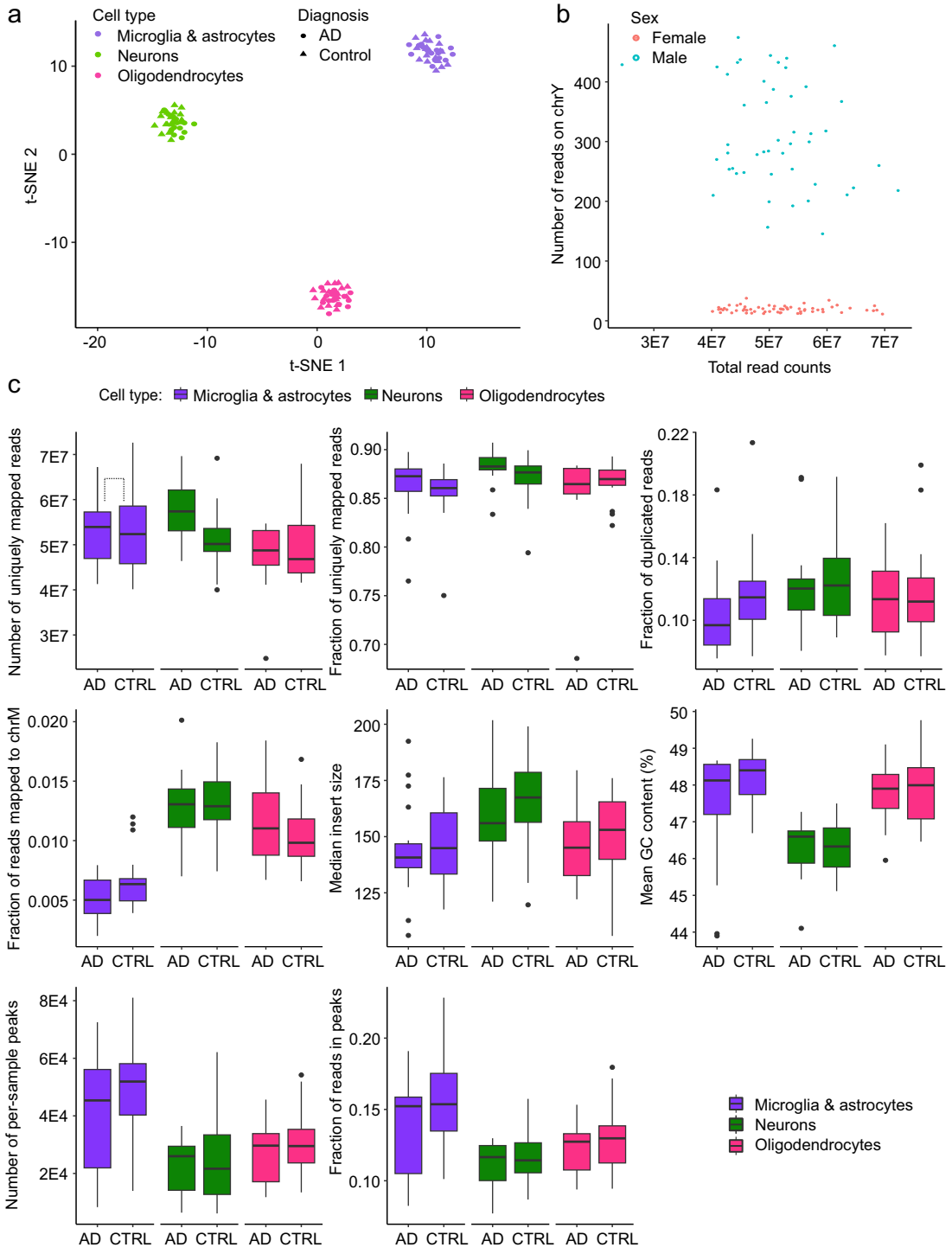

Fig. 4 Analysis of ATAC-seq dataset. Diagnosis groups are defined by CERAD metrics, i.e. AD: CERAD = (2-4), CTRL: CERAD = 1. **(a)** t-SNE clustering. **(b)** Sex check based on quantification of the number of reads on chromosome Y (outside the pseudoautosomal region). **(c)** ATAC-seq quality control metrics stratified by cell subtype and AD disease status: counts of uniquely mapped reads, fraction of uniquely mapped reads, fraction of duplicated reads, fraction of reads mapping to the mitochondrial genome, median insert size, mean GC content, number of per-sample peaks and FRiP (fraction of reads in peaks). t-test comparison on the distributions of values of AD cases and controls for all QC metrics in three cell types revealed that only 1 metrics in 1 cell type is different before correction for multiple testing (the fraction of reads mapped on chrM: p-value = 0.041, FDR-corrected q-value = 0.123) and 1 metrics in 1 cell type is statistically significantly different after correction for multiple testing. (the number of uniquely mapped reads in neurons: p-value 0.008, FDR-corrected q-value = 0.234). Box plots are centered on median, bounds defined between the 25th and 75th percentile with minimum and maximum defined as median ± 1.5 × interquartile range, and whiskers extending to the lowest/highest value within this range.

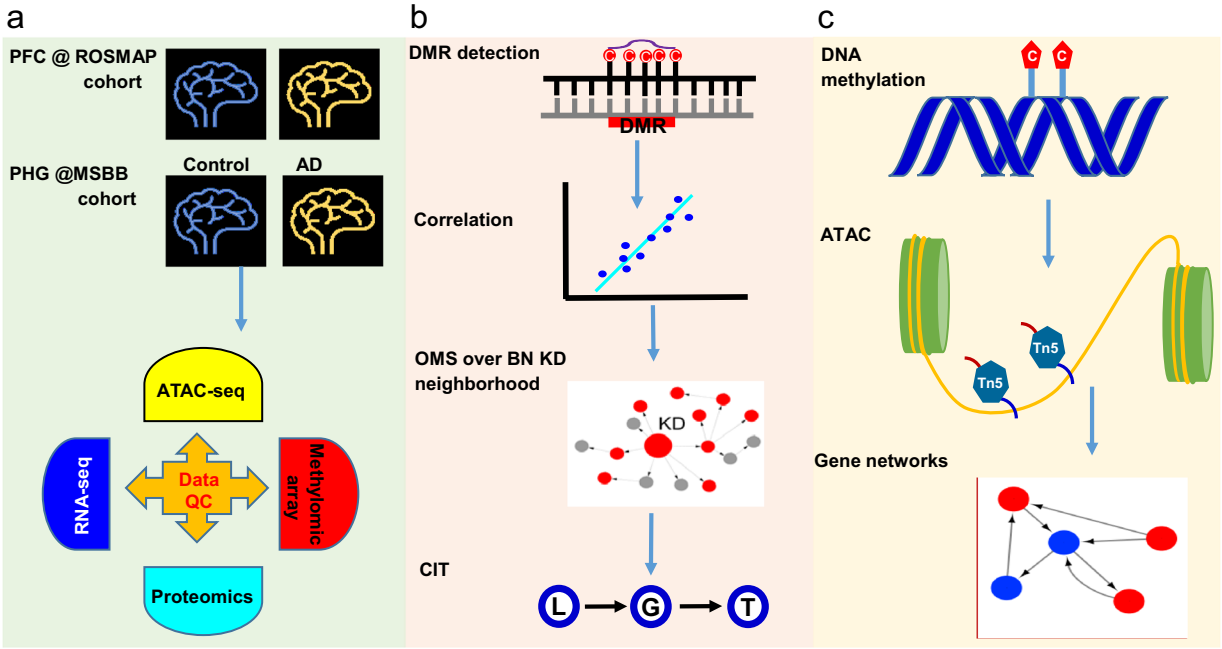

Fig. 5 Integration of large-scale multi-omics data in AD to develop predictive molecular network models for identifying key driving factors of AD. (**a**) Overview of the study design, multi-omics data collection and quality check. (**b**) DNA methylation regulates gene and protein expression as well as their coexpression networks. DMRs were first identified, followed by correlation/association analyses of DMRs with AD clinical traits, network module relevance to AD, gene/protein network connectivity, AD risk genes, and Aβ gene signatures. Overall methylation score (OMS) was calculated to model the net effects of DMRs over gene expression, and was further investigated for their relevance to network metrics; Last, the causal relationship of methylation and ATAC domain to gene expression was evaluated by cit (causal inference test). BN: Bayesian network, KD: key driver; OMS: overall methylation score, DMR: differentially methylated region, CIT: causal inference test. (**c**) Methylomic regulatory cascade. The results suggest that DMRs are likely causal to ATAC domain activity in regulating the expression of genes in the networks.

multi-omics data reveals a detailed signaling map of the regulatory cascade among DNA methylation, epigenomic chromatin accessibility, transcription and translation in AD (Fig. 5).

Code availability

The source code demonstrating the work with the dataset is available at https://doi.org/10.5281/zenodo.7818443[60].

Received: 16 May 2023; Accepted: 29 August 2023;
Published online: 08 September 2023

References

1. Livingston, G. *et al.* Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *Lancet* **396**, 413–446, https://doi.org/10.1016/s0140-6736(20)30367-6 (2020).
2. Moutinho, S. The long road to a cure for Alzheimer's disease is paved with failures. *Nat. Med.* **28**, 2228–2231, https://doi.org/10.1038/s41591-022-02062-0 (2022).
3. Karran, E. & De Strooper, B. The amyloid hypothesis in Alzheimer disease: new insights from new therapeutics. *Nat. Rev. Drug Discov.* **21**, 306–318, https://doi.org/10.1038/s41573-022-00391-w (2022).
4. Bohbot, V. D. *et al.* Spatial memory deficits in patients with lesions to the right hippocampus and to the right parahippocampal cortex. *Neuropsychologia* **36**, 1217–1238, https://doi.org/10.1016/s0028-3932(97)00161-9 (1998).
5. Ploner, C. J. *et al.* Lesions affecting the parahippocampal cortex yield spatial memory deficits in humans. *Cereb. Cortex* **10**, 1211–1216, https://doi.org/10.1093/cercor/10.12.1211 (2000).
6. LaFlamme, E. M., Waguespack, H. F., Forcelli, P. A. & Malkova, L. The Parahippocampal Cortex and its Functional Connection with the Hippocampus are Critical for Nonnavigational Spatial Memory in Macaques. *Cereb. Cortex* **31**, 2251–2267, https://doi.org/10.1093/cercor/bhaa358 (2021).
7. Teipel, S. J. *et al.* Comprehensive dissection of the medial temporal lobe in AD: measurement of hippocampus, amygdala, entorhinal, perirhinal and parahippocampal cortices using MRI. *J. Neurol.* **253**, 794–800, https://doi.org/10.1007/s00415-006-0120-4 (2006).
8. Krumm, S. *et al.* Cortical thinning of parahippocampal subregions in very early Alzheimer's disease. *Neurobiol. Aging* **38**, 188–196, https://doi.org/10.1016/j.neurobiolaging.2015.11.001 (2016).
9. Echávarri, C. *et al.* Atrophy in the parahippocampal gyrus as an early biomarker of Alzheimer's disease. *Brain Struct. Funct.* **215**, 265–271, https://pubmed.ncbi.nlm.nih.gov/20957494 (2011).
10. Dickerson, B. C. *et al.* Differential effects of aging and Alzheimer's disease on medial temporal lobe cortical thickness and surface area. *Neurobiol. Aging* **30**, 432–440, https://doi.org/10.1016/j.neurobiolaging.2007.07.022 (2009).
11. Bertram, L. Next Generation Sequencing in Alzheimer's Disease. *Methods Mol. Biol.* **1303**, 281–297, https://doi.org/10.1007/978-1-4939-2627-5_17 (2016).

12. Bertram, L., McQueen, M. B., Mullin, K., Blacker, D. & Tanzi, R. E. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat. Genet.* **39**, 17–23, https://doi.org/10.1038/ng1934 (2007).
13. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413, https://doi.org/10.1038/s41588-018-0311-9 (2019).
14. Prokopenko, D. *et al.* Whole-genome sequencing reveals new Alzheimer's disease-associated rare variants in loci related to synaptic function and neuronal development. *Alzheimers. Dement.* **17**, 1509–1527, https://doi.org/10.1002/alz.12319 (2021).
15. Andrews, S. J., Fulton-Howard, B. & Goate, A. Interpretation of risk loci from genome-wide association studies of Alzheimer's disease. *Lancet Neurol.* **19**, 326–335, https://doi.org/10.1016/s1474-4422(19)30435-1 (2020).
16. Kozlenkov, A. *et al.* Substantial DNA methylation differences between two major neuronal subtypes in human brain. *Nucleic Acids Res.* **44**, 2593–2612, https://doi.org/10.1093/nar/gkv1304 (2016).
17. Nott, A. *et al.* Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. *Science* **366**, 1134–1139, https://doi.org/10.1126/science.aay0793 (2019).
18. Jiang, Y., Matevossian, A., Huang, H.-S., Straubhaar, J. & Akbarian, S. Isolation of neuronal chromatin from brain tissue. *BMC Neurosci.* **9**, 42, https://doi.org/10.1186/1471-2202-9-42 (2008).
19. Bendl, J. *et al.* The three-dimensional landscape of cortical chromatin accessibility in Alzheimer's disease. *Nat. Neurosci.* **25**, 1366–1378 https://doi.org/10.1038/s41593-022-01166-7 (2022).
20. Wang, M. *et al.* The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. *Scientific Data* **5**, 1–16, https://doi.org/10.1038/sdata.2018.185 (2018).
21. Wang, M. *et al.* AD Knowledge Portal - backend. *Synapse* https://doi.org/10.7303/syn7392158 (2018).
22. Fillenbaum, G. G. *et al.* Consortium to Establish a Registry for Alzheimer's Disease (CERAD): the first twenty years. *Alzheimers. Dement.* **4**, 96–109, https://doi.org/10.1016/j.jalz.2007.08.005 (2008).
23. Braak, H., Alafuzoff, I., Arzberger, T., Kretzschmar, H. & Del Tredici, K. Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. *Acta Neuropathol.* **112**, 389–404, https://doi.org/10.1007/s00401-006-0127-z (2006).
24. Braak, H. & Braak, E. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol.* **82**, 239–259, https://doi.org/10.1007/bf00308809 (1991).
25. Morris, J. C. The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology* **43**, 2412–2414, https://doi.org/10.1212/wnl.43.11.2412-a (1993).
26. Bai, B. *et al.* Deep Profiling of Proteome and Phosphoproteome by Isobaric Labeling, Extensive Liquid Chromatography, and Mass Spectrometry. *Methods Enzymol.* **585**, 377–395, https://doi.org/10.1016/bs.mie.2016.10.007 (2017).
27. Xu, P., Duong, D. M. & Peng, J. Systematical optimization of reverse-phase chromatography for shotgun proteomics. *J. Proteome Res.* **8**, 3944–3950, https://doi.org/10.1021/pr900251d (2009).
28. Niu, M. *et al.* Extensive Peptide Fractionation and y1 Ion-Based Interference Detection Method for Enabling Accurate Quantification by Isobaric Labeling and Mass Spectrometry. *Anal. Chem.* **89**, 2956–2963, https://doi.org/10.1021/acs.analchem.6b04415 (2017).
29. Tan, H. *et al.* Integrative Proteomics and Phosphoproteomics Profiling Reveals Dynamic Signaling Networks and Bioenergetics Pathways Underlying T Cell Activation. *Immunity* **46**, 488–503, https://doi.org/10.1016/j.immuni.2017.02.010 (2017).
30. Wang, X. *et al.* JUMP: a tag-based database search tool for peptide identification with high sensitivity and accuracy. *Mol. Cell. Proteomics* **13**, 3663–3673, https://doi.org/10.1074/mcp.o114.039586 (2014).
31. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J. & Gygi, S. P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2**, 43–50, https://doi.org/10.1021/pr025556v (2003).
32. Stewart, E. *et al.* Identification of Therapeutic Targets in Rhabdomyosarcoma through Integrated Genomic, Epigenomic, and Proteomic Analyses. *Cancer Cell* **34**, 411–426.e19, https://doi.org/10.1016/j.ccell.2018.07.012 (2018).
33. Moran, S., Arribas, C. & Esteller, M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics* **8**, 389–399, https://doi.org/10.2217/epi.15.114 (2016).
34. Maksimovic, J., Phipson, B. & Oshlack, A. A cross-package Bioconductor workflow for analysing methylation array data. *F1000Res.* **5**, 1281, https://doi.org/10.12688/f1000research.8839.3 (2016).
35. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120, https://doi.org/10.1093/bioinformatics/btu170 (2014).
36. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21, https://doi.org/10.1093/bioinformatics/bts635 (2013).
37. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323, https://doi.org/10.1186/1471-2105-12-323 (2011).
38. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773, https://doi.org/10.1093/nar/gky955 (2019).
39. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140, https://doi.org/10.1093/bioinformatics/btp616 (2010).
40. DeLuca, D. S. *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532, https://doi.org/10.1093/bioinformatics/bts196 (2012).
41. Hauberg, M. E. *et al.* Common schizophrenia risk variants are enriched in open chromatin regions of human glutamatergic neurons. *Nat. Commun.* **11**, 5581, https://doi.org/10.1038/s41467-020-19319-2 (2020).
42. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1–21.29.9, https://doi.org/10.1002/0471142727.mb2129s109 (2015).
43. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079, https://doi.org/10.1093/bioinformatics/btp352 (2009).
44. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS. *Genome Biol.* **9**, R137, https://doi.org/10.1186/gb-2008-9-9-r137 (2008).
45. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci. Rep.* **9**, 9354, https://doi.org/10.1038/s41598-019-45839-z (2019).
46. Coleman, C. *et al.* AD Knowledge Portal - backend. *Synapse* https://doi.org/10.7303/syn2580853 (2023).
47. Wang, M. *et al.* AD Knowledge Portal - backend. *Synapse* https://doi.org/10.7303/syn10901600 (2018).
48. Li, L. *et al.* SMAP is a pipeline for sample matching in proteogenomics. *Nat. Commun.* **13**, 744, https://doi.org/10.1038/s41467-022-28411-8 (2022).
49. Li, Y. *et al.* JUMPg: An Integrative Proteogenomics Pipeline Identifying Unannotated Proteins in Human Brain and Cancer Cells. *J. Proteome Res.* **15**, 2309–2320, https://doi.org/10.1021/acs.jproteome.6b00344 (2016).
50. Yoo, S. *et al.* MODMatcher: multi-omics data matcher for integrative genomic analysis. *PLoS Comput. Biol.* **10**, e1003790, https://doi.org/10.1371/journal.pcbi.1003790 (2014).
51. Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, 483, https://doi.org/10.1186/s12859-016-1323-z (2016).
52. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358, https://doi.org/10.1093/bioinformatics/bts163 (2012).

53. Heiss, J. A. & Just, A. C. Identifying mislabeled and contaminated DNA methylation microarray data: an extended quality control toolset with examples from GEO. *Clin. Epigenetics* **10**, 73, https://doi.org/10.1186/s13148-018-0504-1 (2018).
54. Manichaikul, A. *et al*. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873, https://doi.org/10.1093/bioinformatics/btq559 (2010).
55. Wang, E. *et al*. Genome-wide methylomic regulation of multiscale gene networks in Alzheimer's disease. *Alzheimers. Dement.* https://doi.org/10.1002/alz.12969 (2023).
56. Bennett, D. A. & Launer, L. J. Longitudinal epidemiologic clinical-pathologic studies of aging and Alzheimer's disease. *Curr. Alzheimer Res.* **9**, 617–620, https://doi.org/10.2174/156720512801322645 (2012).
57. Bennett, D. A. *et al*. Religious Orders Study and Rush Memory and Aging Project. *J. Alzheimers. Dis.* **64**, S161–S189, https://doi.org/10.3233/jad-179939 (2018).
58. Klein, H.-U. *et al*. Epigenome-wide study uncovers large-scale changes in histone acetylation driven by tau pathology in aging and Alzheimer's human brains. *Nat. Neurosci.* **22**, 37–46, https://doi.org/10.1038/s41593-018-0291-1 (2018).
59. Millstein, J., Zhang, B., Zhu, J. & Schadt, E. E. Disentangling molecular relationships with a causal inference test. *BMC Genet.* **10**, 23, https://doi.org/10.1186/1471-2156-10-23 (2009).
60. Coleman, C. *et al*. clairecoleman1/PHG_Code_Examples: Second release. *Zenodo* https://doi.org/10.5281/zenodo.7818443 (2023).

Acknowledgements

We thank the patients and families who donated material for these studies. We thank members of the Roussos laboratory for thoughtful advice and critique. This study was supported by grants from the National Institute on Aging, the National Institutes of Health (NIH) [RF1AG057440, U01AG046170, RF1AG054014, R01AG057907, R01AG068030, RF1AG074010, R01AG065582, R01AG067025, R01AG050986]. J.B. was supported in part by Alzheimer's Association Research Fellowship AARF-21-722200. This work was supported in part through the computational and data resources and staff expertise provided by Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai and supported by the Clinical and Translational Science Awards (CTSA) grant UL1TR004419 from the National Center for Advancing Translational Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

P.R. and B.Z. perceived the concept and designed the cohort study. V.H. provided the brain samples and performed pathological analysis. J.F.F., C.M., Z.S., J.M.V., K.Y., Y.L., J.P. prepared the samples. C.C., J.B., M.W., Y.L., J.P. and E.W. performed the data analysis. C.C., J.B., J.F.F., P.R., M.W., E.W., Y.L., J.P. and B.Z. wrote and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.Z. or P.R.

Reprints and permissions information is available at www.nature.com/reprints.

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