nature genetics

Article

https://doi.org/10.1038/s41588-024-01685-y

Cell subtype-specific effects of genetic variation in the Alzheimer's disease brain

Received: 16 December 2022

Accepted: 8 February 2024

Published online: 21 March 2024

Check for updates

Masashi Fujita[1,19], Zongmei Gao[1,19], Lu Zeng[1,19], Cristin McCabe[2], Charles C. White[1], Bernard Ng[3], Gilad Sahar Green[4], Orit Rozenblatt-Rosen[2,18], Devan Phillips[2,18], Liat Amir-Zilberstein[2], Hyo Lee[5,6], Richard V. Pearse II[5,6], Atlas Khan[7], Badri N. Vardarajan[8,9,10], Krzysztof Kiryluk[7], Chun Jimmie Ye[11,12,13,14], Hans-Ulrich Klein[1], Gao Wang[9], Aviv Regev[2,15,18], Naomi Habib[4], Julie A. Schneider[3], Yanling Wang[3], Tracy Young-Pearse[5,6], Sara Mostafavi[16,17], David A. Bennett[3], Vilas Menon[1,20] & Philip L. De Jager[1,20]

The relationship between genetic variation and gene expression in brain cell types and subtypes remains understudied. Here, we generated single-nucleus RNA sequencing data from the neocortex of 424 individuals of advanced age; we assessed the effect of genetic variants on RNA expression in *cis* (*cis*-expression quantitative trait loci) for seven cell types and 64 cell subtypes using 1.5 million transcriptomes. This effort identified 10,004 eGenes at the cell type level and 8,099 eGenes at the cell subtype level. Many eGenes are only detected within cell subtypes. A new variant influences *APOE* expression only in microglia and is associated with greater cerebral amyloid angiopathy but not Alzheimer's disease pathology, after adjusting for *APOEε4*, providing mechanistic insights into both pathologies. Furthermore, only a *TMEM106B* variant affects the proportion of cell subtypes. Integration of these results with genome-wide association studies highlighted the targeted cell type and probable causal gene within Alzheimer's disease, schizophrenia, educational attainment and Parkinson's disease loci.

Gene discovery studies have created a critical new foundation for the study of neurodegenerative and neuropsychiatric diseases, consisting of a growing list of validated susceptibility loci and a much larger set of loci with suggestive levels of evidence. Reference epigenomic atlases generated from small numbers of individuals or cell lines have been helpful in prioritizing variants within loci and in suggesting the relevant tissue or cell type for a given variant[1–3]. However, these reference data are insufficient: 'quantitative trait locus' (QTL) studies are needed to map the functional consequences of variants. Until now, QTL studies were largely limited to tissue-level molecular profiles[4,5] and the imperfect estimates of cell types from bulk deconvolution efforts[6,7]. However, cytometry enabled the measurement of targeted proteins on individual cells[8,9], while cell sorting approaches enabled the large-scale profiling of rare cell types[10,11].

An early, modest-sized effort to map brain expression QTLs (eQTLs) collated different small single-nucleus RNA sequencing (snRNA-seq) datasets from different brain regions and showed that such mapping efforts were likely to be fruitful[12]. In this study, we analyzed snRNA-seq data derived from the dorsolateral prefrontal cortex (DLPFC)—a hub for cognitive and mood circuits—of 424 older participants. The relationship of these data to phenotypes of Alzheimer's disease (AD) is reported elsewhere[13]. Our effort uncovered eQTLs in the seven major neocortical cell types and 64 of their subtypes[13], discovering many *cis*-eQTLs active in only one subtype of cell. This importance of nuanced gene expression is critical to understand the functional consequences of disease-associated variation and the transition of insights to induced pluripotent stem cell

A full list of affiliations appears at the end of the paper. e-mail: pld2115@cumc.columbia.edu

https://doi.org/10.1038/s41588-024-01685-y

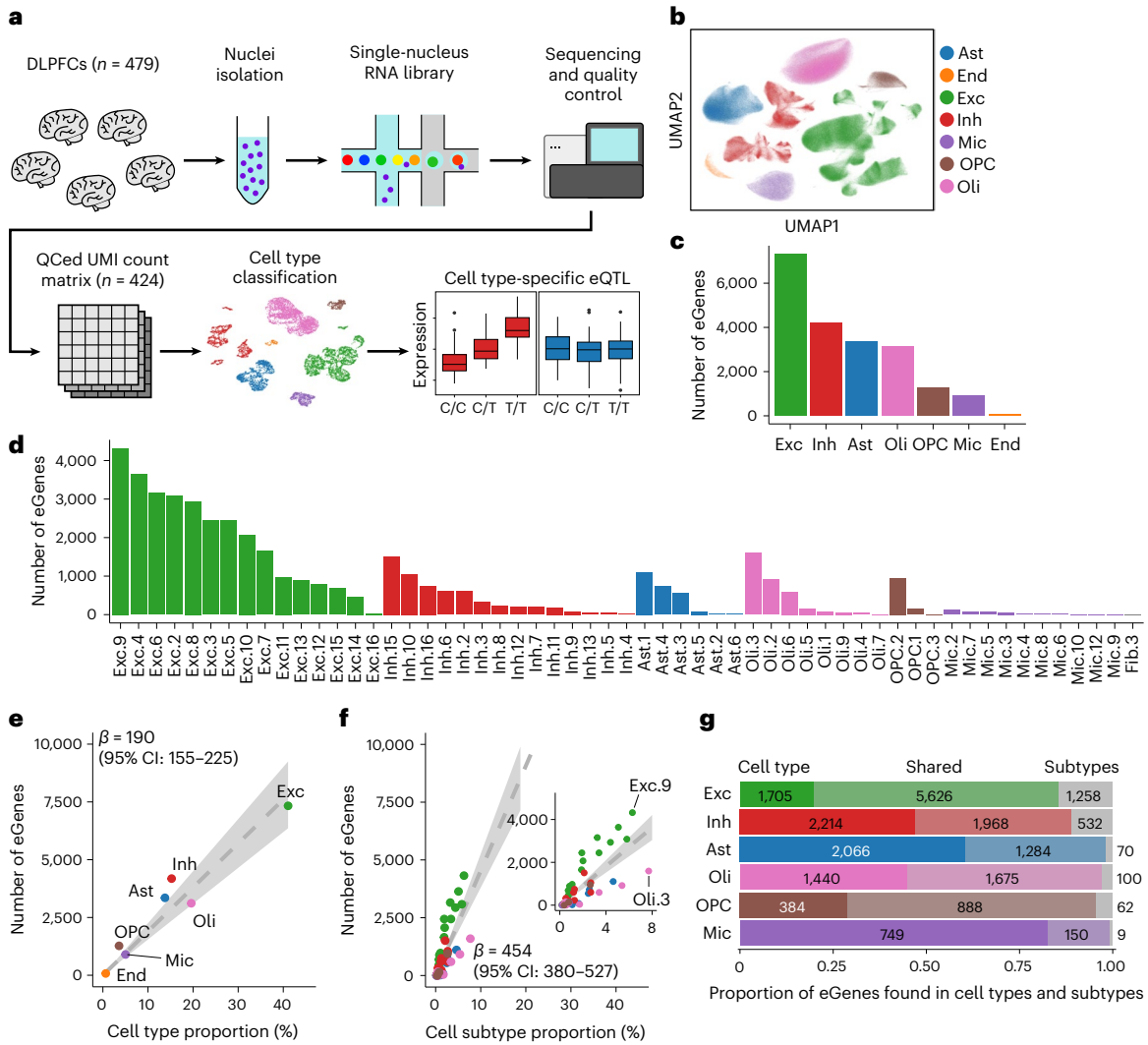

Fig. 1 | Study design and summary of cell type-specific and subtype-specific *cis*-eQTL. a, Schema of our study. **b**, Uniform manifold approximation and projection (UMAP) visualization of 1,509,626 nuclei from 424 donors. Each of the seven major cell types is labeled with a different color. Ast, astrocyte; End, endothelial cell; Exc, excitatory neuron; Inh, inhibitory neuron; Mic, microglia; Oli, oligodendrocyte. **c**, Number of eGenes (genes targeted by a *cis*-eQTL effect) detected within each of the seven cell types. **d**, Number of eGenes detected in each of the 64 cell subtypes that were retained for analysis. **e**,**f**, Relationship between cell (sub)type proportions and number of eGenes detected. The dashed line shows a least-squares fit with zero intercept; β is its slope. The shaded area represents the 95% confidence interval (CI). **e**, Cell type proportions. **f**, Cell subtype proportions; the slope is much steeper than for the cell types, as illustrated by the inset, which enlarges the plotting of the data near the origin. **g**, Number of eGenes that are unique to the analysis of cell subtypes: for each cell type, we present a bar chart summarizing the extent to which cell type-level eGenes were found once the cells assigned to a given cell type were partitioned into the subtypes of that cell type; the six most common cell types are shown. For each cell type, the set of eGenes identified in all subtypes of a given cell type are shown in gray; in each cell type, a subset of these cell subtype eGenes were not recovered in the cell type-level analysis, suggesting that they may be specific to a cell subtype context.

(iPSC)-derived neurons and astrocytes, which exhibit certain eQTLs. We also mapped the effects of genetic variants on and the heritability of the frequency of cell subtypes (fraction QTLs, (fQTLs)). Integrating our results with those of gene discovery studies, we prioritized (1) cell types in which individual susceptibility loci appear to be having their effect, (2) putative causal variants in each risk haplotype and (3) the target gene of each variant.

Results

Description of the dataset

Figure 1a shows the schema of our study. Leveraging our previous work[14,15], snRNA-seq data were generated from frozen samples of DLPFC obtained from the brains of participants in two longitudinal studies of cognitive aging with prospective autopsies: the Religious Order Study (ROS) and the Memory Aging Project (MAP)[16]. All participants were without known dementia at baseline (Methods). The demographic and diagnostic details of the 424 participants are presented in Supplementary Table 1. At the time of death, 34% of participants were cognitively nonimpaired, 26% were mildly impaired and 40% had dementia. Of the 424 participants, 68% were female and 63% fulfilled a pathological diagnosis of AD by the National Institutes of Health (NIH) Reagan Criteria. After preprocessing, 424 participants with both snRNA-seq and whole-genome sequencing (WGS) data were retained for analysis. Participants had a median of 3,824 nuclei. We used a stepwise clustering approach (Methods) to identify first the major cell types of the DLPFC and then subtypes in each cell type (Fig. 1b and Supplementary Figs. 1 and 2). In the end, we organized our data into eight major cell types (seven of which had enough data for QTL mapping), which were

further subdivided into 95 cell subtypes found in the human DLPFC (64 of which had sufficient data for QTL mapping)[13].

Mapping *cis*-eQTLs at the cell type level

To map *cis*-eQTLs, we created pseudobulk RNA expression measures for each major cell type by collapsing unique molecular identifier (UMI) counts from all nuclei assigned to a given cell type in each individual (Methods). We repeated the process at the cell subtype level. We used the Matrix eQTL v2.3 software[17] adjusting for 30 expression principal components, as we empirically determined this number as maximizing eQTL discovery (Methods). After inclusion of these covariates, clinicopathological traits and postmortem interval did not meaningfully influence eQTL discovery. Supplementary Fig. 3a shows the number of genes tested for eQTLs in each cell type. Supplementary Table 2 summarizes the results of *cis*-eQTL mapping at the cell type level. In Fig. 1c, we summarize the number of eGenes detected, where an 'eGene' is a gene for which an SNP–gene QTL pair exceeded our threshold of significance, a two-step false discovery rate (FDR) lower than 0.05 (Methods). Using a similar approach, Fig. 1d and Supplementary Table 3 present the number of eGenes detected in the 64 cell subtypes retained for eQTL analysis.

The large difference in the number of eQTLs between cell types is explained in part by cell type proportions (Fig. 1e). Given that neurons, particularly the excitatory type but also true for inhibitory neurons, are the most abundant cell types in the neocortex and also have, on average, the most RNA, it is not surprising that they drive the largest proportion of eGene discovery. Less common cells such as microglia (approximately 5.0% of nuclei) return a substantial (899 eGenes) but smaller number of eGenes when compared to excitatory neurons (7,331 eGenes). Because we excluded genes that have few UMI counts from the pseudobulk computation, abundant cell types had higher UMI counts and thus had more genes tested for eQTLs (Supplementary Fig. 3a). Endothelial cells had fewer participants in its pseudobulk expression (Supplementary Fig. 3b), further deteriorating the statistical power for eQTL detection. Moreover, higher UMI counts resulted in more accurate estimate and higher heritability of gene expression (Supplementary Fig. 3c). The correlation between cell population frequency and eGene discovery also held among cell subtypes (Fig. 1f). Unexpectedly, the slope of eGenes per cell population proportions was much steeper among cell subtypes than in cell types ($\beta = 454$ and $\beta = 190$ per 1% increase in cell subtype or type frequency, respectively, which is significantly different; $P = 6.9 \times 10^{-9}$), suggesting that cell subtypes may be a better target for eGene discovery in future studies.

To evaluate the extent to which cell subtype analysis enhanced *cis*-eQTL discovery, we compared eGenes between each cell type and its subtypes (Fig. 1g). In excitatory neurons, 1,258 unique eGenes were only detected in excitatory neuron subtypes but not in the cell type-level analysis in which all excitatory neuron subtypes were pooled. There was also a nonnegligible gain of eGenes when analyzing subtypes for all other cell types, including astrocytes, microglia, oligodendrocytes and oligodendroglial progenitor cells (OPCs) (Fig. 1g).

Similarity and specificity of neocortical eGenes

Among the 10,004 eGenes that we detected across all cell types, a substantial minority (4,598 eGenes; 46%) was cell type-specific (Fig. 2a). Figure 2b illustrates the extent to which eGenes were shared among the different cell types. Cell type specificity of eQTLs can be explained either by (1) a target gene expressed only in one cell type, or (2) a target gene expressed in multiple cell types but with genetic association in only one cell type. Interestingly, a large fraction of cell type-specific eGenes fitted the latter pattern (blue bars in Fig. 2b). These eGenes may have multiple enhancers, one of which is specific for a given cell type. A similar trend was observed in the cell subtype analyses (Supplementary Fig. 4).

To assess the extent of eQTL sharing between cell types, we computed a π_1 statistic (Fig. 2c). Aside from endothelial cells, which had a limited number of eGenes given their low frequency among the nuclei, the six other cell types had a high degree of eGene sharing ($\pi_1 = 0.53$–0.94). For *cis*-eQTLs shared and significant in two cell types, Fig. 2d shows that the vast majority of shared eQTLs had the same direction of effect and similar effect sizes (Supplementary Fig. 5a,b). However, there were cell type-specific effects: *APP*, an important gene related to AD, was associated with rs128648 only in oligodendrocytes (Fig. 2e) and *APOE*, expressed in six cell types, showed a new *cis*-eQTL effect unique to microglia (Fig. 2f). *APOE* is one of the key genes upregulated in disease-associated microglia found in amyloid proteinopathy models[18]. This *APOE* locus variant, rs2288911, is associated with AD in published GWAS ($P = 1.0 \times 10^{-12}$ and $P = 6.0 \times 10^{-10}$)[19,20]. However, this variant was not associated with amyloid or tau proteinopathy ($P > 0.05$) in ROS/MAP. Nonetheless, it was associated with the burden of cerebral amyloid angiopathy (CAA) ($P = 1.18 \times 10^{-7}$) (Supplementary Fig. 6); this association persisted even after accounting for the effect of *APOEε4* ($P = 9.9 \times 10^{-6}$). There was no evidence of an interaction between the effects of rs2288911 and *APOEε4* on CAA burden ($P > 0.05$). Therefore, this microglial eQTL was independent of the *APOEε4* haplotype, which alters the coding sequence of *APOE* and does not affect the *APOE* expression level. Thus, increased expression of *APOE* in microglia leads to more CAA but not AD pathology, causing microhemorrhages that contribute to dementia, potentially explaining the association to AD dementia. The rs2288911 variant may also be relevant for risk stratification in current anti-amyloid antibody treatment protocols, as amyloid-related imaging abnormality (ARIA) is an adverse event of these therapies and is already reported to be influenced by the *APOEε4* allele[21].

Comparison to existing datasets

To assess the robustness of our eQTL results, we expanded an earlier effort[5] and mapped *cis*-eQTLs from bulk DLPFC RNA-seq data from 1,092 ROS/MAP participants (Supplementary Table 4); 408 participants were shared between the bulk and snRNA-seq analyses. Figure 3a shows that most eGenes were detected in both datasets. The bulk-specific effects are not surprising given the greater sample size and greater complexity of the bulk transcriptome (which contained cytoplasmic RNA) relative to nuclear data. However, 40% of snRNA-seq-derived eGenes were not discovered in the larger bulk cortical RNA-seq dataset from the same brain region, confirming the importance of mapping eGenes at the single-cell level. As expected, sharing (π_1 statistic) with the bulk results was greatest in those more abundant cell types (Fig. 3b and Supplementary Fig. 7).

We also compared our microglial results from snRNA-seq to a recently published set of eQTLs from bulk microglia isolated from autopsy tissue[11]: there was modest overlap between the two datasets (Supplementary Fig. 8). The limited overlap may be partly explained by the single-nucleus nature of our study, which cannot detect cytoplasmic RNA, the different brain regions profiled and the small sample size of the bulk microglial profiles ($n = 77$).

Finally, we also compared our results to those recently generated from the merger of several small snRNA-seq datasets from different brain regions[12]. In these analyses, we validated many results of the earlier, smaller effort: 80% of cell type–eGene pairs in the previous study are included in our results (Fig. 3c) ($\pi_1 = 0.90$ when the results of the previous manuscript were used as reference). However, our larger, coherent dataset uncovered 15,335 new eGene–cell type pairs; importantly, we mapped eGenes at the cell subtype level, which the earlier effort did not.

Translating results in vitro to guide functional studies

To evaluate how well our eGenes translate to model systems, we repurposed bulk RNA-seq data from iPSC-derived neurons ($n = 44$) and

https://doi.org/10.1038/s41588-024-01685-y

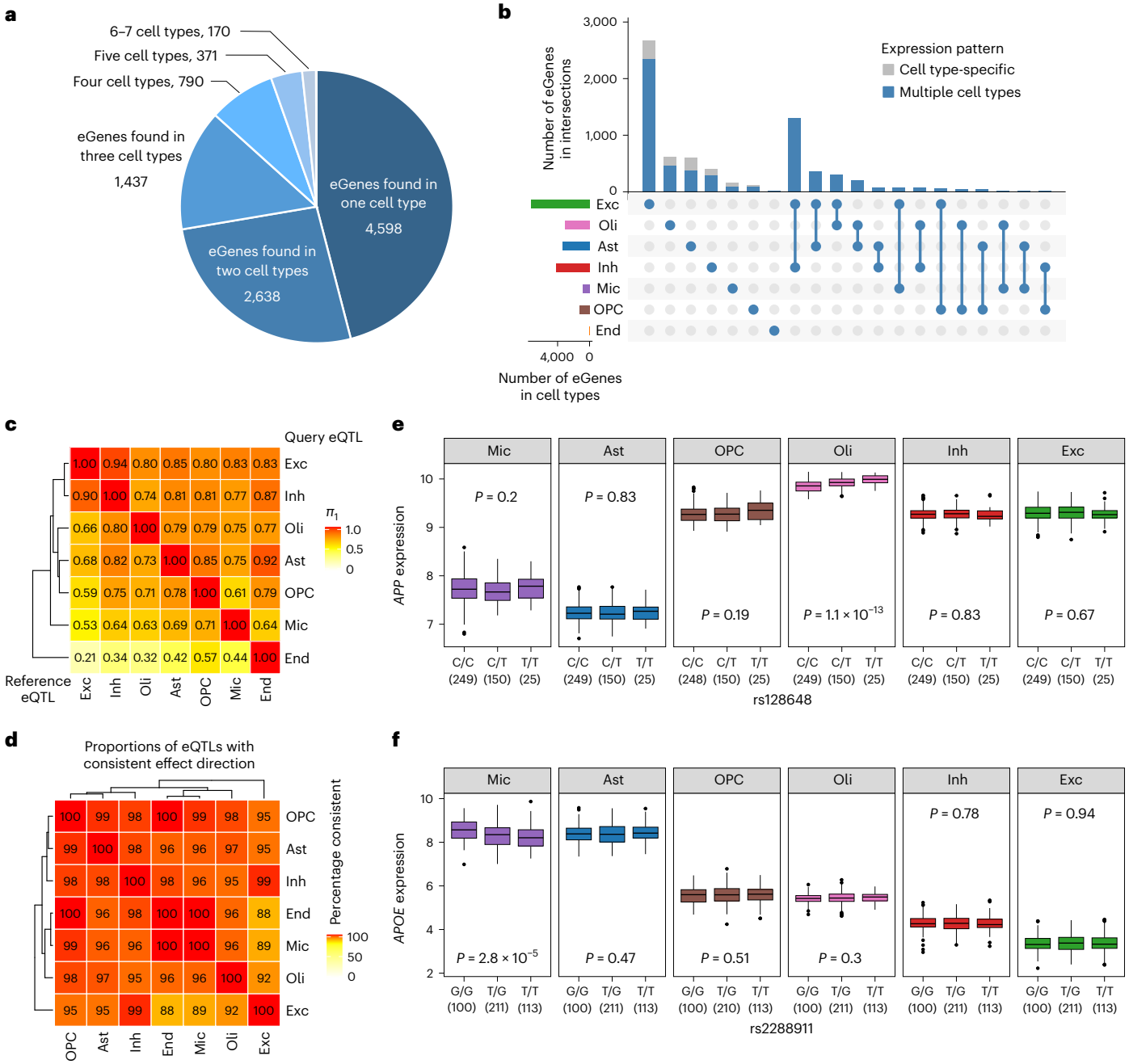

Fig. 2 | Similarities and differences of cell type-specific eQTLs. a, Number of cell type-specific and nonspecific eGenes. **b**, Number of eGenes that were unique to or shared between cell types. Only the top 20 intersections are shown. In the vertical bar chart, the proportion of eGenes specifically expressed in one cell type is colored gray and that expressed in two or more cell types is colored blue: most genes are expressed in more than one cell type. **c**, π_1 statistic to quantitate the extent of eQTL sharing between each pair of cell type. **d**, Proportion of shared eQTLs that had consistent direction of effect in each pair of cell types. **e**,**f**, Examples of cell type-specific eQTLs. *P* values were computed using a simple linear regression between allele dosage and cell type-level gene expression. The number of participants is shown in parentheses. **e**, Oligodendrocyte-specific eQTL between rs128648 and *APP* gene expression. **f**, Microglia-specific eQTLs between rs2288911 and *APOE* gene expression. Elements of the boxes show the following statistics: center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range (IQR); points, outliers. Measurements were taken from distinct samples in **e** and **f**.

astrocytes ($n = 38$) derived from ROS/MAP participants[22]. Given the modest iPSC sample size, we limited our evaluation to replicating the excitatory neuron and astrocytic eGene results (Fig. 3d). A total of 6,414 genes were expressed in both iPSC-derived neurons and excitatory neuron nuclei; 234 genes (3.6% of the assessed eGenes) were associated with the predicted expression SNPs (eSNPs) in iPSC-derived neurons (FDR < 0.05). This is significantly more frequent than random pairs of genes and SNPs ($P < 1 \times 10^{-6}$). For astrocytes, 121 of 2,529 (4.8%) eGenes were reproduced in iPSC-derived astrocytes ($P < 1 \times 10^{-6}$). This analysis shows that many eQTL effects are reproduced in an artificial, in vitro context. Effect sizes were largely similar and most were in the same direction (snRNA-seq versus iPSC-derived neuron and astrocyte data) (Fig. 3e,f). However, there were some interesting exceptions, including the *MAPT* gene, which encodes the tau protein. rs11100065 is an eSNP in all four contexts, but the direction of the effect on *MAPT* expression is inverted in the iPSC-derived contexts (Fig. 3g). Given the strength

https://doi.org/10.1038/s41588-024-01685-y

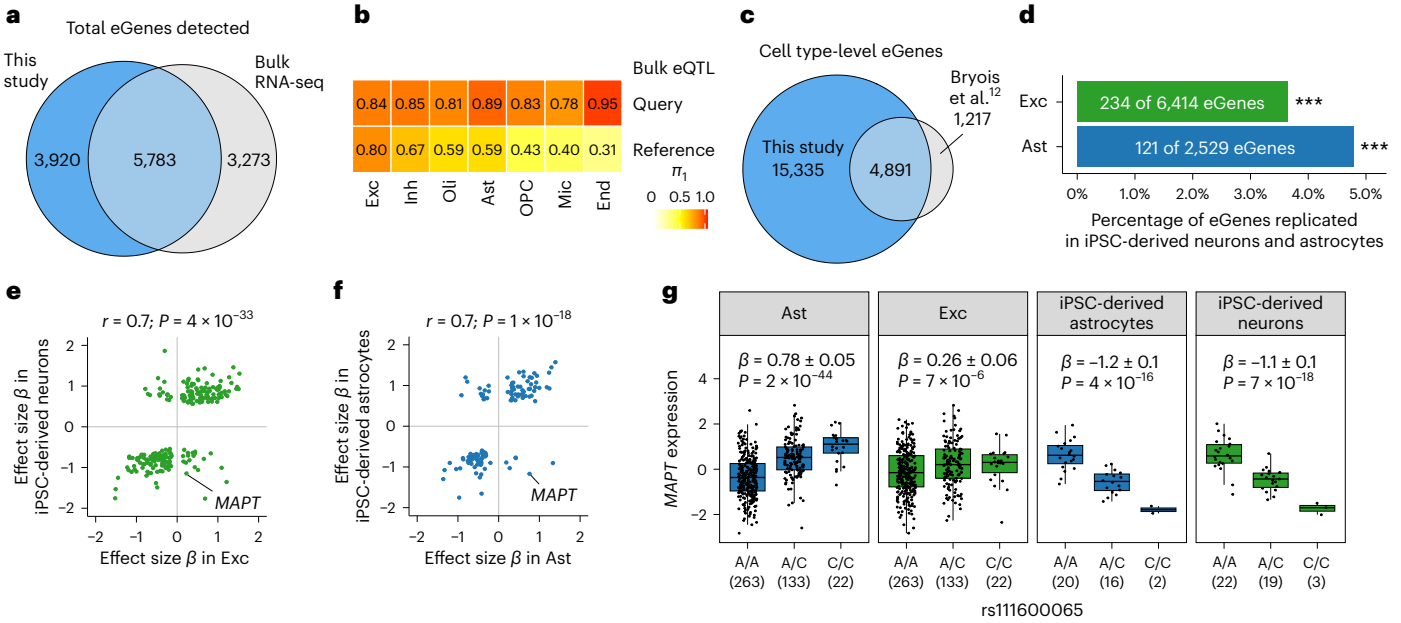

Fig. 3 | Replication of eQTL using bulk cortical RNA and RNA from induced pluripotent stem cell lines. a, Number of eGenes detected in our single-nucleus eQTL study versus a bulk cortical RNA eQTL study using the same brain region (DLPFC). For the single-nucleus eQTL results, unique eGenes of the seven cell types were combined. For bulk cortical eQTLs, we mapped *cis*-eQTLs using 1,092 individuals from the ROS/MAP studies. **b**, π_1 statistic of single-nucleus and bulk eQTL. The top and bottom rows used bulk eQTL as query and reference, respectively. **c**, Comparison of our single-nucleus-derived eGenes and those from an earlier study[12] that had a smaller sample size and combined a mixture of brain tissues to produce results only at the cell type level. **d**, Number of cell type-level eGenes that were replicated in 44 neuronal and 38 astrocyte cell lines derived from the iPSCs of ROS/MAP participants. The lead eSNP of each eGene was tested to see whether its allele dosage was associated with the eGene expression level in the corresponding iPSC-derived data with an FDR < 0.05. We replicated more of the eGenes than expected by chance. $^{***}P < 1 \times 10^{-6}$ using a one-sided permutation test. **e**,**f**, Effect size β of eQTLs shared between single-nucleus and iPSC-derived cells. *r*, Pearson's correlation coefficient. *P* values were computed using two-sided *t*-tests. **e**, Excitatory neurons and iPSC-derived neurons. 234 eGenes; 95% CI of $r = 0.60–0.74$. **f**, Astrocytes and iPSC-derived astrocytes. 121 eGenes; 95% CI of $r = 0.59–0.78$. **g**, Opposite direction of effect for the *MAPT* eQTL at rs111600065. A linear regression was applied to *MAPT* expression and allele dosage. Effect size β and its s.e. are shown in the figure. The major 'A' allele tags the H1 haplotype of *MAPT*, and the minor 'C' allele tags the H2 haplotype, which is associated with PD, parkinsonism and perhaps AD[39,40]. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5× IQR; points, outliers. Measurements were taken from distinct samples. The number of participants is shown in parentheses.

of the effect in vitro, its presence in both iPSC-derived neurons and astrocytes and the fact that most eGenes had an effect in the same direction (snRNA-seq versus iPSC-derived), this is unlikely to be a statistical fluctuation. This observation needs further validation, particularly because two other studies of bulk neocortical data reported a direction of effect similar to iPSC-derived cells[23,24]. Nonetheless, it is a cautionary tale that even well-studied but genetically complex loci, such as *MAPT*, can harbor substantial surprises in model systems.

Chromatin annotation of *cis*-eQTLs

We annotated our results with available epigenomic features from bulk DLPFC[1] and relevant cell types[25]. Figure 4a illustrates the expected enrichment of eSNPs within approximately 100 kb of the target gene's transcription start site (TSS)[26,27]. Furthermore, these eSNPs were enriched in DLPFC euchromatin (transcriptionally active regions and enhancers) and relatively depleted in heterochromatic regions (Fig. 4b). We also saw larger enrichment among TSS (Fig. 4c) and enrichment among enhancers annotated in the corresponding reference cell type. eSNPs for microglial eGenes were more likely found in the enhancer regions found in bulk microglia than in bulk neurons, oligodendroglia or astrocytes. In Fig. 4d, we show the distribution of peaks from chromatin immunoprecipitation followed by sequencing (ChIP–seq) in the *APOE* locus relative to the rs2288911 SNP that drives a microglial-specific *cis*-eQTL of *APOE* (Fig. 2f). This SNP is 40.2 kb from *APOE* (near another gene, *APOC2*) but in a chromosomal segment decorated, only in microglia, with H3K27ac and H3K4me3, two marks associated with active enhancers and promoters, respectively. Furthermore, proximity ligation-assisted chromatin immunoprecipitation sequencing (PLAC-seq) data suggested that this chromosomal segment was in physical proximity to the *APOE* gene[25], which was also in an active conformation. Thus, it is plausible that this variant may be driving the observed effect on microglial *APOE* expression, while the astrocytic *APOE* locus (which is also in a transcriptionally active conformation) is unaffected by rs2288911, suggesting that the risk allele at this SNP may have a key role in the accumulation of CAA through an effect on microglia as described above.

Variants influencing the proportions of cell subtypes

Figure 5a and Supplementary Table 5 present the results of fraction quantitative trait locus (fQTL) analyses. Assessing the heritability of each subtype frequency, only committed oligodendrocyte precursors (COPs) showed modest evidence of heritability (Supplementary Table 6). The statistical power of our dataset may be insufficient to evaluate heritability, but these results suggest that the frequency of many of these subtypes may not be strongly influenced by genetic variation.

We then examined the relevance of fQTLs to AD risk. Among the AD risk SNPs we tested[28], rs5011436 coincided with the fQTL of excitatory neuron 3 (Exc.3) (Fig. 5b). This *TMEM106B* locus SNP was the sole fQTL for Exc.3 (Fig. 5c). Using the CelMod method[15], we inferred the proportion of all cell subtypes in bulk DLPFC RNA-seq data from an independent set of 602 ROS/MAP participants (Methods). The Exc.3 fQTL with rs5011436 was replicated in these imputed cell type frequencies ($P = 1.5 \times 10^{-7}$). The SNP not only was an fQTL but also had a *cis*-eQTL effect on *TMEM106B* expression in bulk cortical RNA-seq data (Fig. 5d–f). The major allele rs5011436^{A} is a risk allele for AD[20,28],

https://doi.org/10.1038/s41588-024-01685-y

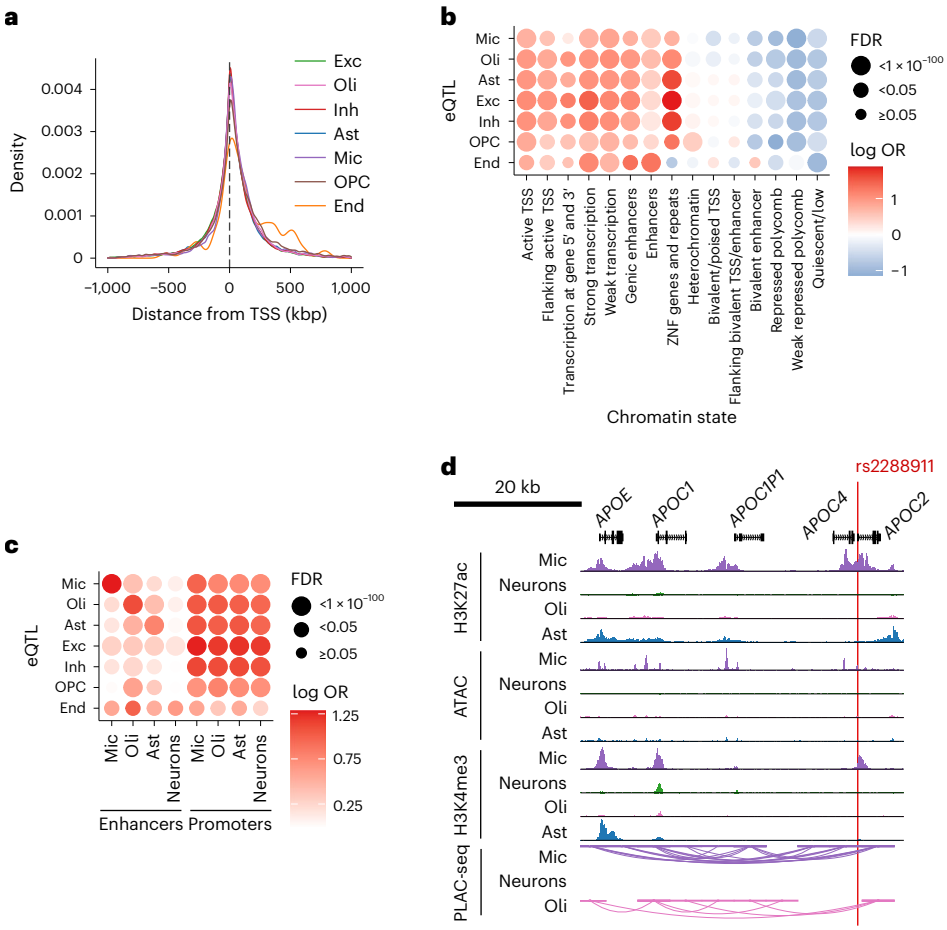

Fig. 4 | Chromatin states of cell type-level eQTLs. a, Distance between eSNPs and TSS of eGenes. **b**, Enrichment of eSNP in chromatin states. Chromatin states of a human DLPFC tissue were obtained from sample E073 of the Roadmap Epigenomics[1]. **c**, Enrichment of eQTL in cell type-specific enhancers and promoters. Enhancers and promoters of four brain cell types were obtained from a published report[25]. **d**, Microglia-specific eQTL for *APOE* (rs2288911) in the context of microglial-specific chromatin conformation data. These data were repurposed from an earlier report[25]. The *x* axis denotes the physical position along a segment of chromosome 19 containing the *APOE* gene and several related genes; their exon structure is presented in the top horizontal track. The next four tracks report ChIP–seq data against the H3K27ac epitope, a mark found in active TSS and enhancers; each track presents data from a different cell type, isolated as purified nuclei. The peaks denote segments that were in a transcriptionally active conformation. The next four tracks present data from the same samples using the assay for transposase accessible chromatin (ATAC), which denotes chromosomal segments in an open conformation and accessible for transcription. The four H3K4me3 tracks present ChIP–seq data for that epitope, which is also correlated with active promoter regions of a chromosome. PLAC-seq data are presented in the last four tracks and denote pairs of chromosomal segments that are in physical proximity to one another as the chromatin loops.

but this association may derive from its effect on susceptibility for frontotemporal dementia (FTD)[29], which may be why this locus emerges in certain AD genome-wide association study (GWAS). In addition, we ran a meta-phenome-wide association study (PheWAS) for this variant using the UK Biobank and the eMERGE-III data[30,31]; interestingly, the two significant results out of 1,817 clinical traits tested were related to diabetes (Supplementary Fig. 9 and Supplementary Table 7); three of the top five suggestive results ($P < 0.001$) were related to atherosclerosis and included cerebrovascular disease. This is notable because both diabetes and vascular disease influence the risk of AD and dementia. Therefore, this *TMEM106B* variant may influence certain neuronal proportions through vascular and metabolic effects.

Colocalization with disease susceptibility

An important use of our *cis*-eQTL results involves aligning them with the results of gene discovery studies to assess whether altered gene expression may be the mechanism for a particular risk allele. Using Coloc (v5.1.0), a GWAS of AD and dementia[28] and our eGenes, we found evidence of colocalization (posterior probability of the H4 hypothesis (PP.H4) > 0.8) for 21 eGenes among the 20 AD loci that we interrogated (Fig. 6a). We confirmed some of the well-validated results, such as the *BIN1* risk haplotype tagged by rs4663105, the susceptibility haplotype with the largest effect size for AD after *APOE*, which drives *BIN1* expression only in microglia[11]; other cell types express *BIN1* but do not exhibit this effect. More interesting is that we found four new colocalizations with our data: AC004797.1, AL596218.1, AP001439.1 and *ITGA2B*. As expected with AD, we found that microglia harbored the most implicated target genes, although all other neocortical cell types were also implicated by our colocalization analysis (Fig. 6a). Endothelial cells had few nuclei per participants and thus have very few eGenes so far. Thus, we cannot interpret the lack of colocalization so far in this cell type. While many loci had unambiguous cell type-specific effects (that is, *CASS4* or *ACE*), one had an effect in three cell types (*APH1B*: astrocytes, excitatory neurons and oligodendroglial cells) and another had distinct target genes in two cell types (*CCDC6*: astrocytes and microglia). While strong posterior probabilities were seen in most loci, some loci had poor (*ZCWPW1*) or muddled evidence of colocalization that will require further dissection (*SCIMP1*). As reported previously, we found colocalization in the *GRN* locus; however, interestingly, it was found in both oligodendroglial cells and excitatory neurons. Furthermore,

https://doi.org/10.1038/s41588-024-01685-y

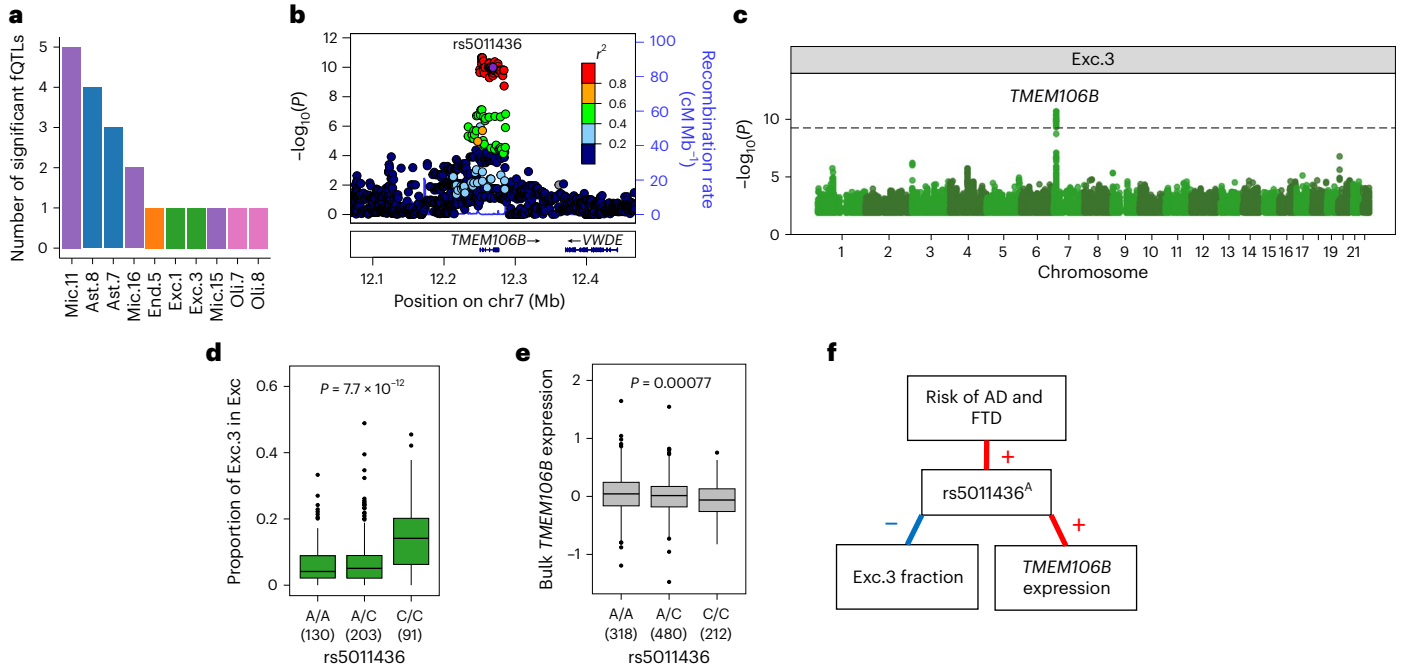

Fig. 5 | fQTLs between SNPs and cell subtype proportions. a, Number of independent significant fQTLs per cell subtype. **b**, Locus zoom plot around the *TMEM106B* locus. A lead risk SNP of AD in the locus (rs5011436) was used as the reference SNP given that it tagged a risk haplotype that contains many other SNPs with similar statistical properties given their strong LD[28]. The *x* axis shows the position of the SNPs along chromosome 7. The *y* axis shows the $-\log_{10}$ of *P* values for the excitatory neuron subtype 3 (Exc.3) cell subtype fQTL. **c**, Manhattan plot for the genome-wide fQTL results for the Exc.3 neuronal subtype. The dashed line shows the significance threshold $P < 5.6 \times 10^{-10}$ (accounting for all tested fQTL GWAS). **d**, Genotypes of rs5011436 and proportion of Exc.3 among excitatory neurons. The number of participants is shown in parentheses. **e**, Genotypes of rs5011436 and gene expression levels of *TMEM106B* in bulk DLPFC tissues, illustrating the local *cis*-eQTLs. The number of participants is shown in parentheses. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5× IQR; points, outliers. Measurements were taken from distinct samples in **c**–**e**. **f**, Schematic summary of the major allele of rs5011436 and its relationship to three phenotypes. The red and blue lines show positive and negative correlations, respectively. In **b**–**e**, *P* values were computed using two-sided *t*-tests.

there was some evidence that another gene, *ITGA2B*, may also be implicated (although modestly) in excitatory neurons. To gauge gains from single-nucleus analysis, we repeated the AD colocalization analysis using bulk eQTLs and subtype-level eQTLs. Compared to 21 colocalized eGenes found in cell type-level eQTLs, bulk eQTLs and subtype-level eQTLs had eight and 11 colocalized eGenes (Supplementary Fig. 10a). This demonstrates that cell type-level analysis substantially improved colocalization of AD GWAS signals over the analysis of bulk data and that subtype-level analysis can further augment the yield of such analyses.

While data from the frontal cortex from aging individuals and individuals with AD may be most relevant to interpreting the results of AD GWAS, we conducted similar analyses for schizophrenia (SCZ), a neuropsychiatric disease with frontal cortex involvement, and identified 57 loci colocalized with 75 eGenes (Fig. 6c). In Parkinson's disease (PD), we identified 11 loci mapping to 13 eGenes (Fig. 6b). For ALS, Coloc identified one locus and one eGene (Supplementary Fig. 10b). For both PD and SCZ, we found that excitatory neurons achieved the largest number of colocalized loci. The results of our colocalization analyses for these and other central nervous system-related traits, such as educational attainment and brain volumetric measures, are summarized in Supplementary Tables 8 and 9.

Gene prioritization for AD and neuropsychiatric diseases

In our data, we found that the expression level of thousands of genes was strongly heritable in all cell types except endothelial cells, which were undersampled in our dataset (Supplementary Fig. 10c and Supplementary Table 10). Heritability was assessed using genomic relatedness-based restricted maximum-likelihood (GREML) with a threshold of $P \leq 0.05$ (ref. 32). This enabled us to perform a 'transcriptome-wide association study' (TWAS) for each cell type in which we inferred gene-level association statistics by combining the GWAS results for a trait and a gene's model to infer RNA expression. This was complementary to colocalization studies and was particularly helpful in prioritizing genes for further evaluation in loci where no individual variant reached a threshold of genome-wide significance. In Fig. 6d, we display summaries of the TWAS for different diseases, confirming the large excess of microglial genes involved in AD relative to other neuropsychiatric diseases. Figure 6e and Supplementary Fig. 11 highlight *FNIP2* and *RAPGEF2*, two of 24 microglial genes that have not previously been associated with AD in similar analyses using other RNA data. Supplementary Tables 11 and 12, and Supplementary Figs. 12 and 13, report the TWAS results for other neuropsychiatric diseases, educational attainment and brain volumetric measures.

The TWAS analysis was also helpful in supporting the colocalization results. For example, all nine microglial colocalized effects in AD were also found in the TWAS; when the lead SNP (and those in strong linkage disequilibrium (LD)) were removed, the TWAS analysis was still positive, suggesting that there may be additional significant associations beyond the lead haplotype in those susceptibility loci.

Discussion

In this study, we constructed a single-nucleus-based eQTL resource for the aging neocortex and cataloged the genetic regulation of gene expression in each cell type and subtype. This provides a substantial advance over the previous existing effort, which identified only a third of our eGenes, had a much smaller sample size, collated data from different brain regions and only evaluated eQTLs at the cell type level[12]. One of our principal findings is that many eGenes were only discovered when analyses were conducted at the cell subtype level (Fig. 1g), highlighting an important strategic choice for future study design, that is,

https://doi.org/10.1038/s41588-024-01685-y

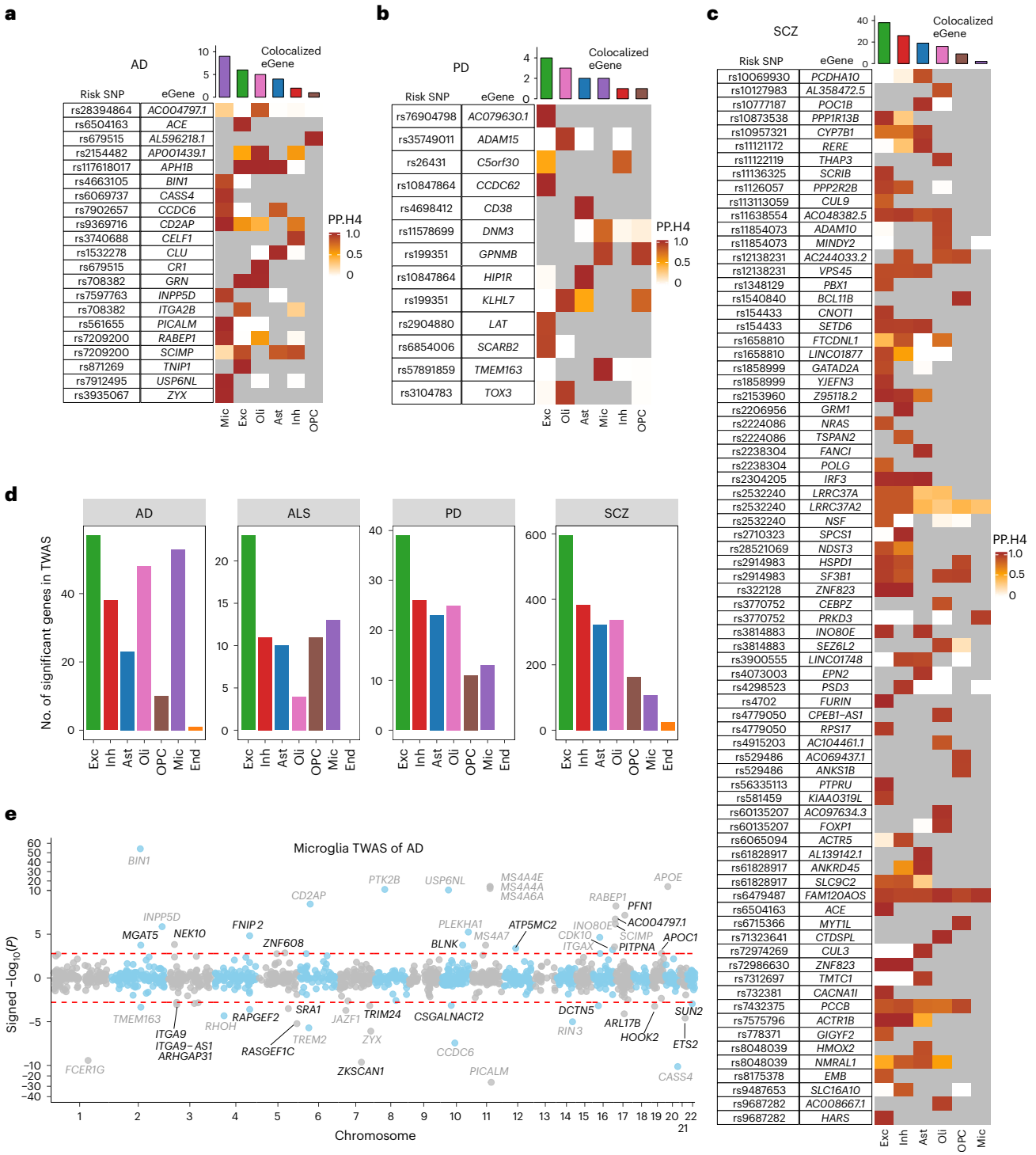

Fig. 6 | Overlap of the results from our eQTL and GWAS of selected neurodegenerative and neuropsychiatric diseases. a–**c**, Colocalization of cell type-specific eQTLs with risk SNPs of AD GWAS[28] (**a**), PD GWAS[41] (**b**) and SCZ GWAS[42] (**c**). Each of the heatmaps reports the PP.H4 of the Coloc method[43], which assumes that GWAS and eQTLs share a single causal SNP. The rows report the overlap for individual gene and SNP pairs; the columns report the PP.H4 score in each of our cell types. The color of each square is based on the code found to the right of each panel; the darker color denotes higher confidence that the same variant influences susceptibility and gene expression in that cell type. Gray cells indicate that the gene was not an eQTL target in that cell type. The top bar chart shows the number of colocalized eGenes with high confidence (PP.H4 > 0.8) in each cell type. **d**, Cell type-level TWAS. Using the FUSION method, we deployed instruments inferring the expression of 28,305 genes across all cell types in the summary statistics for AD, PD, amyotrophic lateral sclerosis (ALS) and SCZ. The count of genes meeting a transcriptome-wide threshold of significance in each cell type is presented for each disease, with the expected excess of microglial genes in AD and an intriguing number of oligodendroglial genes in PD. Cell types are in order of descending expression heritability. **e**, Illustration of the TWAS result from one cell type in one disease: the statistical significance and effect direction of all inferred microglial genes are presented, with the physical position along the chromosome being presented on the *x* axis and the significance on the *y* axis. Each dot represents a gene. The positive and negative *y* coordinates show that transcript abundance was associated with increased and decreased risk of AD, respectively. The *y* axis between −10 and 10 has been enlarged to enhance visibility. New and known candidates for AD risk genes in microglia are colored black and gray, respectively. The red dashed lines highlight the threshold of FDR = 0.05. *P* values were determined using two-sided *z*-tests.

sequencing a larger number of nuclei per participant may be more important than increasing sample size.

Our analysis detected many eQTLs not found in bulk brain tissues (Fig. 3a), despite analyzing less than half the specimens used in the bulk analysis, reflecting cell type-specific and cell subtype-specific regulation of gene expression (Fig. 4c). As exemplified by *APOE*, we found that eGenes were often expressed in multiple brain cell types but genetically regulated in only one cell type, reflecting perturbation of context-specific enhancer elements. The significantly greater slope for eGene discovery in cell subtypes (Fig. 1e,f) may arise because cell subtype-specific variation in enhancer function may be better tolerated than similar effects at the cell type level, where its effect would affect all cells of a certain type. Genetic variation influencing gene expression at the cell subtype level may thus be less likely to be selected against over evolutionary time scales. The microglial-specific *APOE* eQTL is probably driven by a variant in a microglial enhancer that is brought into contact with *APOE* by a chromatin loop; interestingly, this variant also influences the accumulation of CAA (an amyloid-driven vasculopathy) but not parenchymal amyloid plaques or tau tangles. On the other hand, the well-known *APOEε4* haplotype influences all three AD-related pathologies. This provides an important mechanistic insight. *APOE* RNA expression levels secreted by microglia exert a causal role in CAA while they do not affect parenchymal amyloid proteinopathy or taupathy. Microglia are certainly involved in both AD and CAA processes, but only the coding variants that define the AD susceptibility haplotype, which are not strongly related to gene expression, influence AD parenchymal pathology. This result is consistent with reports of increased *APOE* expression at the site of CAA[33] and of reduced CAA accumulation[34] in mice treated with an anti-*APOE* antibody. This new *APOE* variant may have clinical relevance because the presence of CAA is a risk factor for ARIA after anti-amyloid antibody treatment, and *APOEε4* is a major risk factor for ARIA[21].

Mapping the eQTLs of rare cell types, such as microglia, has been a technical challenge. Recent efforts to map microglia eQTLs relied on dissociation and purification of microglia[10,11,35]. However, enzymatic dissociation of brain tissues can induce artifactual gene expression changes, especially in microglia[36]. On the other hand, our single-nucleus eQTL analysis suffered from low detection power in less common cell types, including microglia; as such, cells and cell subtypes had fewer genes detected and tested for eQTLs. One solution to this issue would be to sequence individuals more deeply using single-nucleus analysis. As library preparation costs drop, this approach may become more feasible than cell purification, which is arduous for most neocortical cell types and could lead to exclusion of certain cell subtypes. Sample multiplexing approaches can substantially reduce costs and batch effects, as we demonstrated in this study, which relied only on genetic demultiplexing without antibody-based tags; this minimizes manipulation of the samples (versus nuclear hashing) and reduces the quantity of sequencing needed[13].

iPSC-derived neurons and astrocytes are being used to explore the functional consequences of genetic variants[22,37]; we confirmed that a significant fraction of our brain-derived eGenes was replicated in both iPSC-derived neurons and astrocytes, despite a small sample size and thus limited statistical power. Most eQTLs had the same effect direction in vivo and in vitro. However, some eQTLs, including the one affecting the *MAPT* gene, had opposite effect directions (Fig. 3e,f), suggesting that genetic effects are highly context-specific and that caution is needed when extrapolating brain-derived genetic effects to this model. This type of context-specific inverse eQTL directionality has been noted previously in a minority of genes, such as *CD52*, the target of alemtuzumab, when comparing monocyte and $CD4^+$ T cell eQTLs derived from the same blood samples[38].

Functional follow-up of disease variants identified in GWAS has lagged behind gene discovery efforts. Our colocalization analysis revealed cell types and even subtypes where risk variants exert their effect, guiding the design of experiments in the proper context. Furthermore, TWAS results clearly indicate that, while microglia in AD and neurons in SCZ, PD and educational attainment are targets of predilection for genetic susceptibility, all cell types harbor the primary effect of some disease susceptibility variants: these genetically complex traits are cellularly distributed, highlighting that it is the summary of perturbations across pathogenic cellular communities that leads to disease[13]. We can now begin to identify points of convergence in such communities to develop new therapeutic interventions.

Overall, we systematically mapped the effect of genetic variation across a wide variety of cellular contexts in the aging brain. While we characterized many disease loci, it is clear that many more eGenes remain to be discovered. We also highlighted deeper sequencing to better resolve cell subtypes as an important aspect of the path forward.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-024-01685-y.

References

1. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
2. Stunnenberg, H. G. et al. The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. *Cell* **167**, 1145–1149 (2016).
3. Abascal, F. et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
4. Aguet, F. et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
5. Ng, B. et al. An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat. Neurosci.* **20**, 1418–1426 (2017).
6. Patrick, E. et al. Deconvolving the contributions of cell-type heterogeneity on cortical gene expression. *PLoS Comput. Biol.* **16**, e1008120 (2020).
7. Donovan, M. K. R., D'Antonio-Chronowska, A., D'Antonio, M. & Frazer, K. A. Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. *Nat. Commun.* **11**, 955 (2020).
8. Chan, G. et al. CD33 modulates TREM2: convergence of Alzheimer loci. *Nat. Neurosci.* **18**, 1556–1558 (2015).
9. Roederer, M. et al. The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis. *Cell* **161**, 387–403 (2015).
10. Young, A. M. H. et al. A map of transcriptional heterogeneity and regulatory variation in human microglia. *Nat. Genet.* **53**, 861–868 (2021).
11. Lopes, K. P. et al. Genetic analysis of the human microglial transcriptome across brain regions, aging and disease pathologies. *Nat. Genet.* **54**, 4–17 (2022).
12. Bryois, J. et al. Cell-type-specific *cis*-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. *Nat. Neurosci.* **25**, 1104–1112 (2022).
13. Green, G. S. et al. Cellular dynamics across aged human brains uncover a multicellular cascade leading to Alzheimer's disease. Preprint at *bioRxiv* https://doi.org/10.1101/2023.03.07.531493 (2023).
14. Habib, N. et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955–958 (2017).
15. Cain, A. et al. Multicellular communities are perturbed in the aging human brain and Alzheimer's disease. *Nat. Neurosci.* **26**, 1267–1280 (2023).

16. Bennett, D. A. et al. Religious Orders Study and Rush Memory and Aging Project. *J. Alzheimers Dis.* **64**, S161–S189 (2018).
17. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
18. Li, Y., Macyczko, J. R., Liu, C.-C. & Bu, G. ApoE4 reduction: an emerging and promising therapeutic strategy for Alzheimer's disease. *Neurobiol. Aging* **115**, 20–28 (2022).
19. Kunkle, B. W. et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).
20. Bellenguez, C. et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat. Genet.* **54**, 412–436 (2022).
21. Joseph-Mathurin, N. et al. Amyloid-related imaging abnormalities in the DIAN-TU-001 trial of gantenerumab and solanezumab: lessons from a trial in dominantly inherited Alzheimer disease. *Ann. Neurol.* **92**, 729–744 (2022).
22. Lagomarsino, V. N. et al. Stem cell-derived neurons reflect features of protein networks, neuropathology, and cognitive outcome of their aged human donors. *Neuron* **109**, 3402–3420 (2021).
23. Allen, M. et al. Association of MAPT haplotypes with Alzheimer's disease risk and MAPT brain gene expression levels. *Alzheimers Res. Ther.* **6**, 39 (2014).
24. Valenca, G. T. et al. The role of *MAPT* haplotype H2 and isoform 1N/4R in parkinsonism of older adults. *PLoS ONE* **11**, e0157452 (2016).
25. Nott, A. et al. Brain cell type-specific enhancer–promoter interactome maps and disease-risk association. *Science* **366**, 1134–1139 (2019).
26. Dimas, A. S. et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325**, 1246–1250 (2009).
27. Gamazon, E. R. et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* **50**, 956–967 (2018).
28. Wightman, D. P. et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat. Genet.* **53**, 1276–1282 (2021).
29. Van Deerlin, V. M. et al. Common variants at 7p21 are associated with frontotemporal lobar degeneration with TDP-43 inclusions. *Nat. Genet.* **42**, 234–239 (2010).
30. eMERGE Consortium. Lessons learned from the eMERGE Network: balancing genomics in discovery and practice. *HGG Adv.* **2**, 100018 (2021).
31. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
32. Yang, J., Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. GCTA-GREML accounts for linkage disequilibrium when estimating genetic variance from genome-wide SNPs. *Proc. Natl Acad. Sci. USA* **113**, E4579–E4580 (2016).
33. Matsuo, K. et al. Complement activation in capillary cerebral amyloid angiopathy. *Dement. Geriatr. Cogn. Disord.* **44**, 343–353 (2018).
34. Xiong, M. et al. APOE immunotherapy reduces cerebral amyloid angiopathy and amyloid plaques while improving cerebrovascular function. *Sci. Transl. Med.* **13**, eabd7522 (2021).
35. Kosoy, R. et al. Genetics of the human microglia regulome refines Alzheimer's disease risk loci. *Nat. Genet.* **54**, 1145–1154 (2022).
36. Marsh, S. E. et al. Dissection of artifactual and confounding glial signatures by single-cell sequencing of mouse and human brain. *Nat. Neurosci.* **25**, 306–316 (2022).
37. Sullivan, S. E. & Young-Pearse, T. L. Induced pluripotent stem cells as a discovery tool for Alzheimer's disease. *Brain Res.* **1656**, 98–106 (2017).
38. Raj, T. et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science* **344**, 519–523 (2014).
39. Zabetian, C. P. et al. Association analysis of *MAPT* H1 haplotype and subhaplotypes in Parkinson's disease. *Ann. Neurol.* **62**, 137–144 (2007).
40. Pastor, P. et al. *MAPT* H1 haplotype is associated with late-onset Alzheimer's disease risk in *APOEε4* noncarriers: results from the Dementia Genetics Spanish Consortium. *J. Alzheimers Dis.* **49**, 343–352 (2016).
41. Nalls, M. A. et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **18**, 1091–1102 (2019).
42. Trubetskoy, V. et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).
43. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).

[1]Center for Translational and Computational Neuroimmunology, Department of Neurology, Columbia University Irving Medical Center, New York, NY, USA. [2]Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [3]Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL, USA. [4]Edmond & Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel. [5]Ann Romney Center for Neurologic Diseases, Brigham and Women's Hospital, Boston, MA, USA. [6]Harvard Medical School, Boston, MA, USA. [7]Department of Medicine, Vagelos College of Physicians & Surgeons, Columbia University, New York, NY, USA. [8]Taub Institute for Research on Alzheimer's Disease and the Aging Brain, College of Physicians and Surgeons, Columbia University, New York, NY, USA. [9]Department of Neurology, College of Physicians and Surgeons, Columbia University and the New York Presbyterian Hospital, New York, NY, USA. [10]The Gertrude H. Sergievsky Center, College of Physicians and Surgeons, Columbia University, New York, NY, USA. [11]Institute for Human Genetics, University of California, San Francisco, CA, USA. [12]Department of Epidemiology and Biostatistics, University of California, San Francisco, CA, USA. [13]Parker Institute for Cancer Immunotherapy, San Francisco, CA, USA. [14]Chan Zuckerberg Biohub, San Francisco, CA, USA. [15]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [16]Department of Statistics, Centre for Molecular Medicine and Therapeutics, British Columbia Children's Hospital, University of British Columbia, Vancouver, British Columbia, Canada. [17]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA. [18]Present address: Genentech, South San Francisco, CA, USA. [19]These authors contributed equally: Masashi Fujita, Zongmei Gao, Lu Zeng. [20]These authors jointly supervised this work: Vilas Menon, Philip L. De Jager. ✉e-mail: pld2115@cumc.columbia.edu

Methods

Study participants

All brain specimens were derived from two longitudinal clinico-pathological cohort studies, that is, the ROS and MAP[16]. In both cohorts, participants did not have known dementia at the time of enrollment. Participants agreed to receive clinical evaluation each year and to donate their brain at the time of death. Because ROS and MAP are prospective cohorts, participants with incident dementia after enrollment are characterized before death. The two studies were designed and are run by the same group of investigators, with a large core of identical antemortem and postmortem phenotypic data collection. Thus, they are designed to be analyzed jointly[44] and are referred to as 'ROS/MAP'. Each study was approved by the institutional review board of Rush University Medical Center. All participants signed a written informed consent, Anatomical Gift Act and repository consent. For this study, we selected 479 specimens based on availability of frozen pathological material from the DLPFC; only participants with a postmortem interval less than 48 h were considered, as in our previous studies[45]. At the end of our data preprocessing and quality control analyses (described below), we excluded 19 individuals without whole-genome sequence data, leaving 424 participants retained for the genetic analyses. Their demographic and clinicopathological characteristics are described in Supplementary Table 1.

WGS

The WGS of ROS/MAP participants were performed as described previously[46]. Briefly, DNA was extracted from brain or blood samples. WGS libraries were prepared using the KAPA Hyper Library Preparation Kit and sequenced on an Illumina HiSeq X sequencer as 150-bp paired-end reads. Reads were mapped to the reference human genome GRCh37 using the Burrows–Wheeler Aligner-MEM (v.0.7.8); variants were called using the GATK HaplotypeCaller (v.3.4.0). For this study, the variant call format (VCF) file was lifted over to the GRCh38 using Picard's LiftoverVcf (v.2.18.14); only variants that passed the GATK filter were used. These WGS VCF files are available from the AD Knowledge Portal (https://www.synapse.org/#!Synapse:syn11724057).

Library preparation and sequencing of single nuclei

Each batch of samples for library construction consisted of eight participants, except for batch B63, which consisted of seven participants. Batches were designed to balance clinical and pathological diagnosis and sex as much as possible (Supplementary Fig. 14). DLFPC tissue specimens were received frozen from the Rush Alzheimer's Disease Center. We observed variability in the morphology of these tissue specimens with differing amounts of gray and white matter and presence of attached meninges. Working on ice throughout, we carefully dissected to remove white matter and meninges, when present. The following steps were also conducted on ice: about 50–100 mg of gray matter tissue was transferred into the dounce homogenizer (catalog no. D8938, Sigma-Aldrich) with 2 ml of NP-40 lysis buffer (0.1% NP-40, 10 mM Tris, 146 mM NaCl, 1 mM $CaCl_2$, 21 mM $MgCl_2$, 40 U ml^{-1} of RNase inhibitor (catalog no. 2313B, Takara Bio)). Tissue was gently dounced while on ice 25 times with Pestle A followed by 25 times with Pestle B, then transferred to a 15-ml conical tube. Then, 3 ml PBS + 0.01% BSA (catalog no. B9000S, New England Biolabs) and 40 U ml RNase inhibitor were added for a final volume of 5 ml and then immediately centrifuged with a swing bucket rotor at 500*g* for 5 min at 4 °C. Samples were processed two at a time, the supernatant was removed and the pellets were set on ice to rest while processing the remaining tissues to complete a batch of eight samples. The nuclear pellets were then resuspended in 500 ml of PBS + 0.01% BSA and 40 U ml^{-1} of RNase inhibitor. Nuclei were filtered through 20-µm pre-separation filters (catalog no. 130-101-812, Miltenyi Biotec) and counted using the Nexcelom Cellometer Vision and a 2.5 µg $µl^{-1}$ DAPI stain at 1:1 dilution with the cellometer cell counting chamber (CHT4-SD100-002, Nexcelom Bioscience). Five thousand nuclei from each of eight participants were then pooled into one sample and the 40,000 nuclei in around 15–30-µl volume were loaded into two channels on the 10X Single Cell RNA-seq Platform using the Chromium Single Cell 3′ Reagent Kits v.3. Libraries were made according to the manufacturer's protocol. Briefly, single nuclei were partitioned into nanoliter-scale Gel Bead in Emulsion (GEM) in the Chromium controller instrument where complementary DNA (cDNA) shares a common 10X barcode from the bead. Amplified cDNA was measured using the Qubit HS DNA assay (catalog no. Q32851, Thermo Fisher Scientific) and quality-assessed by BioAnalyzer (catalog no. 5067-4626, Agilent Technologies). This whole-transcriptome-amplified material was diluted to less than 8 ng ml^{-1} and processed through v.3 library construction; the resulting libraries were quantified again using Qubit and BioAnalyzer. Libraries from four channels were pooled and sequenced on one lane of the Illumina HiSeq X by the Broad Institute's Genomics Platform for a target coverage of around one million reads per channel. The same libraries of batches B10–B63 were resequenced at the New York Genome Center using an Illumina NovaSeq 6000 system. The sequencing data of both the Broad Institute and New York Genome Center were used for analysis.

Processing of snRNA-seq reads

For each batch of snRNA-seq FASTQ files, the CellRanger software v.6.0.0 (10X Genomics) was used to map reads onto the reference human genome GRCh38, to collapse reads according to UMI and to count the UMI per gene per droplet. The 'GRCh38-2020-A' file set distributed by 10X Genomics was used as a transcriptome model. The --include-introns option was set to incorporate reads mapped to the intronic region of nuclear pre-mRNA into UMI counts. To call cells among the entire droplets, the remove-background module of CellBender (https://github.com/broadinstitute/CellBender) was applied to raw UMI count matrices. The admixture of ambient RNA was estimated and subtracted from the UMI counts using CellBender. These filtered UMI count matrices were used in the subsequent analyses. All raw and processed data are available through the AD Knowledge Portal (https://www.synapse.org/#!Synapse:syn31512863).

Demultiplexing

Because our snRNA-seq library consisted of nuclei from eight individuals, each nucleus was assigned back to its participant of origin using each nucleus's genotype data obtained from the snRNA-seq reads. We used two different procedures, depending on whether all eight individuals had been genotyped with WGS. When eight individuals were genotyped, we used the demuxlet software[47]. From the WGS-based VCF file of 1,196 ROS/MAP individuals, we extracted SNPs that were in the transcribed regions, passed a filter of GATK and at least one of the eight individuals had its alternate allele. The extracted SNP genotype data were fed to demuxlet along with BAM files generated by CellRanger. When fewer than eight individuals were genotyped, we used freemuxlet (https://github.com/statgen/popscle), which clusters droplets based on SNPs in snRNA-seq reads and generates a VCF file of snRNA-seq-based genotypes of the clusters. The number of clusters was specified to be eight. The snRNA-seq-based VCF file was filtered for genotype quality greater than 30 and compared with the available WGS genotypes using the BCFtools gtcheck command (v.1.9). Each WGS-genotyped individual was assigned to one of the droplet clusters by visually inspecting a heatmap of the number of discordant SNP sites between snRNA-seq and WGS. The above two procedures converged to a table that mapped droplet barcodes onto inferred individuals. Each BAM file generated by CellRanger was split into eight per-individual BAM files, each of which contained reads from distinct individuals, using subset-bam (v.1.1.0) (https://github.com/10XGenomics/subset-bam). The UMI count matrices filtered by CellBender were split into eight per-individual UMI count matrices.

Quality control

Among 479 specimens analyzed by snRNA-seq, 19 specimens were excluded from our analyses because they did not have WGS genotypes. Also, three specimens were excluded at the stage of freemuxlet-based demultiplexing because they had ambiguity in the assignment of droplet clusters to individual genotypes. To identify and exclude potential sample swaps in the remaining 457 specimens, we assessed the concordance of genotypes between snRNA-seq and WGS. Logarithm of the odds (LOD) scores, a metric of genotype concordance, were computed by comparing the per-individual BAM files with the WGS genotypes of matched individuals using Picard CrosscheckFingerprints (v.2.25.4). We used a haplotype map downloaded from https://github.com/naumanjaved/fingerprint_maps. After inspecting a histogram of LOD scores, ten specimens whose LOD scores were less than 50.0 were filtered out. These specimens received few cells by the demultiplexing procedure and were set aside from future reprocessing. As another measure to detect sample swaps, we checked RNA expression levels of the *XIST* gene and confirmed that they were consistent with the clinical sex of the remaining 447 specimens. Three individuals were further excluded because they failed the quality control of the WGS; two were marked as potential sample swaps among WGS and the other was marked as an outlier based on genotype principal component analysis. The latter individual was discarded given the preference to have a genetically homogeneous set of individuals for QTL mapping.

Four specimen-level sequencing metrics were computed from the per-specimen UMI count matrices: (1) estimated number of cells; (2) median UMI counts per cell; (3) median genes per cell; and (4) total genes detected. After inspecting these metrics, eight specimens whose median UMI counts per cell were less than 1,500 were excluded. Among the remaining 436 specimens, 12 individuals were sequenced twice in distinct batches. After comparing sequencing metrics, one of these duplicates was excluded from further analyses. After these quality control processes and matching to whole-genome sequence data was available, 424 individuals remained.

Classifications of cell types

In this section, we outline our classification procedure of cell types. The details can be found in our accompanying paper[13]. The classification procedure of cell subtypes is shown in the Supplementary Methods.

Normalization and clustering pipeline. The following pipeline was executed on the RNA count matrix: normalization and scaling by SCTransform method (with variable.features.n = 2,000, conserve.memory = T; Seurat package v.4 (ref. 48)), dimensionality reduction by principal component analysis (Seurat RunPCA, npcs = 30), construction of *k*-nearest neighbor graph (Seurat FindNeighbors, dims = 1:30) and Louvain community detection clustering (Seurat FindClusters, resolution = 0.2, algorithm = 1).

Automatic classification of cell types. We automatically classified nuclei into one of the following eight major cell types: excitatory neurons; inhibitory neurons; astrocytes; microglia; oligodendrocytes; OPCs; endothelial cells; and pericyte cells. The automatic annotations of nuclei, was done using a weighted elastic-net-regularized logistic regression classifier, fitted over our previous atlas of human aging DLPFC from 24 donors[15] with a total of 182,739 nuclei. The gene count matrix of the previous atlas[15] was log-normalized (Seurat NormalizeData) and scaled (Seurat ScaleData, method = vst) over the top 700 variable features (Seurat FindVariableFeatures, excluding noncoding, nonannotated loci).

To select the optimal regularization parameter, we applied ten-fold cross-validation (cv.glmnet method, glmnet package[49,50]) over a randomly selected 75% of the data. To ensure capture of rare cell types, such as pericytes, we weighted samples as for the number of nuclei of cell types present in the training set. We selected the elastic-net mixing parameter (to increase the sparsity of the fitted model) by evaluating test accuracy over the remaining 25% of the data. The fitted model used only 121 features and achieved a test accuracy of 99.95%.

Removal of low-quality cells. Low-quality nuclei were identified by the total number of UMIs and the number of unique genes. Because different brain cell types have inherently different RNA quantities, we learned cell type-specific thresholds over these parameters. Thresholds were optimized based on hand annotation of ten pooled libraries and applied to all 128 libraries to classify low-quality cells, and then removed from the downstream analysis. The clusters of the ten pooled libraries were manually curated to low-quality and high-quality clusters based on the total number of UMIs and unique gene distributions (Seurat VlnPlot). Then, we selected the cell type-specific thresholds as the median of all optimal total number of UMIs and unique gene parameter pairs, scored using the harmonic mean of precision and recall. The total number of UMIs and unique gene thresholds were: excitatory neurons 2,232 and 1,916; inhibitory neurons 800 and 100; astrocytes 800 and 616; microglia 400 and 253; oligodendrocytes 400 and 253; OPCs 695 and 253; vascular cells 400 and 253; and pericyte cells 400 and 100, respectively. Low-quality clusters were also removed using a soft support vector machine classifier fitted over the ten pooled libraries and using the (1) proportion of nuclei annotated as low quality (by the total number of UMIs and unique gene threshold); (2) the average entropy of cell type prediction; and (3) the proportion of doublets using the demuxlet algorithm.

Doublet detection. Doublets were identified using the demuxlet algorithm, based on the sample barcodes. The additional doublets within a sample were predicted in silico based on their RNA profiles. To predict doublets, we ran DoubletFinder[51] (DoubletFinder_v3 method, pN = 0.5, pK = 75/(1.5 × (no. of nuclei in the library)), nExp = 0, sct = T) over each of the libraries. The doublet threshold for the DoubletFinder predictions was determined for each library, based on the maximal Matthew's correlation coefficient compared to the demuxlet-identified doublets. Furthermore, because DoubletFinder is not designed to identify doublets of the same cell type, we modified it to simulate doublets of parent nuclei of different cell types, inferred based on the cell type classification. Using high-resolution clustering of the nuclei (Seurat FindClusters, resolution = 1.5), we expanded and marked as a doublet any nuclei predicted to be a demultiplexed doublet, a DoubletFinder doublet or belonging to a cluster consisting of more than 70% DoubletFinder doublets.

Pseudobulk expression quantification

As we mentioned, our snRNA-seq libraries were prepared in duplicate. The UMI counts of the two replicates from the same individual were aggregated together. For each cell type, individuals who had fewer than ten cells were excluded from expression quantification for that cell type. Rare cell types were excluded from subsequent analysis if fewer than ten individuals were available for expression quantification. We generated a pseudobulk UMI count matrix for each cell type by extracting the UMI counts of the cell type and by aggregating the counts per gene per individual. Low-expression genes were filtered out using the filterByExpr function of edgeR (v.3.30.3) with its default parameters. The pseudobulk counts were normalized using the trimmed mean of *M*-values method of edgeR; the $\log_2$ of counts per million mapped reads (CPM) was computed using the voom function of limma (v.3.44.3). Low-expression genes whose $\log_2$ CPM was less than 2.0 were filtered out. Batch effects were corrected using the ComBat function of sva (v.3.36.0). Expression levels were quantile-normalized. Pseudobulk expression of cell subtypes was quantified using the same method.

Mapping of *cis*-eQTL

We used the Matrix eQTL v.2.3 method[17] to identify *cis*-eQTLs with 1 Mb of the TSS of each measured gene (gene expression derived using $\log_2$

CPM). All bi-allelic SNPs with a minor allele frequency greater than 0.05, a call rate greater than 95% and Hardy–Weinberg $P > 10^{-6}$ were retained for analysis. We used a linear model for gene expression whose explanatory variables were the allele counts of SNPs and several covariates. Statistical significance was determined from the *t*-statistic. Genotype principal components (PCs) were calculated using PLINK (v.1.90)[52]; we included the top three genotype PCs to account for residual population structure among these individuals of European ancestry. We also calculated the expression PCs based on the RNA expression data within each cell type to identify the number of expression PCs that optimized *cis*-eQTL discovery by regressing out the nongenetic structure in the data; the results of these evaluations are shown in Supplementary Fig. 15. While there was some variation in the optimal number of PCs to include in each cell type, differences were small, and we opted to be consistent and to include the top 30 expression PCs as covariates. We also examined the postmortem interval and clinical traits of individuals as covariates, but they had little impact on the number of eQTLs detected (Supplementary Fig. 16). The final set of covariates were the top three genotype PCs, the top 30 expression PCs, age, sex, the postmortem interval, the study (ROS or MAP) and the total number of genes detected in each participant. Multiple hypothesis correction was performed using a two-step method. Gene-wise *P* values were determined by applying Bonferroni correction to the smallest nominal *P* value of each gene with the number of tested SNPs for the gene. The threshold for statistical significance of eGenes was set to an FDR < 5%, where the FDR was determined from gene-wise *P* values using the Benjamini–Hochberg method. The statistical significance of eSNPs was judged using nominal *P* values and its threshold was set to the largest nominal *P* value of the gene–SNP pair with an FDR < 5%. The lead eSNP was selected to have the smallest *P* value for each eGene. The mapping procedures for the fQTLs and bulk RNA-seq *cis*-eQTLs are described in the Supplementary Methods.

Meta-PheWAS

The Phenome-wide association analysis was performed using the eMERGE-III and UK Biobank datasets with the PheWAS R package. For details, see the Supplementary Methods.

Induced neurons and astrocytes

The iPSC lines were generated from ROS/MAP participants. The iPSC lines were differentiated into excitatory neurons and astrocytes. Bulk RNA-seq was performed on day 21 for iPSC-derived neurons and on day 28 for iPSC-derived astrocytes. For details, see the Supplementary Methods.

Chromatin states

Fifteen class chromatin states of a bulk DLPFC tissue (E073) were downloaded from the FTP site of the Roadmap Epigenome project. Cell type-specific enhancers and promoters of brain cells were downloaded from the UCSC genome browser (as of 11 March 2022; https://genome.ucsc.edu/). The SNPs tested for the eQTLs were categorized into four groups based on whether they were eSNPs and whether they were in the chromatin state of interest. A contingency table was constructed using the number of SNPs in the four categories. The log odds ratio (OR) and *P* values of a two-sided Fisher's exact test were determined with the R package epitools (v.0.5-10.1).

Coloc

The Coloc package (v.5.1.0) was used to apply the approximate Bayes factor (ABF) colocalization hypothesis, which was conducted using the coloc.abf() function, which is under a single causal variant assumption. Under the ABF analysis, the association of the trait with SNPs can be achieved by calculating the posterior probability (from 0 to 1), with 1 indicating the causal SNP. In addition, the ABF analysis has five hypotheses, where PP.H0.abf indicates that there is neither an eQTL nor a GWAS signal at the loci; PP.H1.abf indicates that the locus is only associated with the GWAS; PP.H2.abf indicates that the locus is only associated with the eQTL; PP.H3.abf indicates that both the GWAS and eQTL are associated but to a different genetic variant; and PP.H4.abf indicates that the eQTL and GWAS are associated to the same genetic variant. With the posterior probability of each SNP and aiming to find the causal variants between the GWAS and eQTL, we focused on extracting the PP.H4 value for each SNP in our study.

For the AD GWAS[28], we used the reported lead SNPs of 38 loci. For each locus, we searched for the eSNPs within 500 kb of the lead SNP and listed the eGenes that were paired with the eSNP. We then obtained the eGenes *cis*-eQTL output around the lead eSNP within a 1-Mb window size. In addition, we extracted the GWAS summary statistics around the reported 38 lead SNPs. Finally, we conducted the Coloc for the respective eGene–eQTL pair and the eSNP–GWAS for each cell type. Similarly, for the PD GWAS[41], there were 90 independent genome-wide significant risk loci. We picked one SNP with the smallest *P* value from each locus as the lead SNP for the Coloc analysis. For the SCZ GWAS[42], 270 risk loci were identified as relating to SCZ. The SNP from each locus was used for the Coloc analysis. For the ALS GWAS[53], 15 risk loci were identified as relating to ALS and the SNP from each locus was used for the Coloc analysis. Besides the neurodegenerative diseases, we also conducted colocalization analysis on other brain traits, that is, educational attainment[54] and brain volume. In terms of brain volume, we conducted Coloc using the GWAS summary statistics of intracranial volume[55], hippocampal volume[56], subcortical volume[57], and cortical surface area and thickness[58].

TWAS

We used the pseudobulk RNA-seq data and genotypes from ROS/MAP (424 individuals) to impute the *cis* genetic component of expression into multiple GWAS summary statistics as mentioned in the Coloc analysis. The complete TWAS pipeline was implemented in the FUSION (1 October 2019 version) suite of tools[59]. The steps implemented in FUSION are as follows. First, we estimated the heritability of gene expression and stopped if not significant. We estimated it using a robust version of GCTA-GREML[60], which generates heritability estimates per feature as well as the likelihood ratio test *P* value. Only features with a heritability of $P < 0.05$ were retained for the TWAS analysis. Second, the expression weights were computed by modeling all *cis*-SNPs (±1 Mb from the TSS) using the best linear unbiased prediction, or modeling SNPs and effect sizes with a Bayesian sparse linear mixed model, least absolute shrinkage and selection operator, elastic-net and top SNPs[59,61]. Cross-validation of each of the desired models was then performed. Third, a final estimate of weights for each of the desired models was performed and the results were stored. The imputed unit was treated as a linear model of genotypes with weights based on the correlation between SNPs and expression in the training data while accounting for LD among SNPs. To account for multiple hypotheses, an FDR-corrected *P* threshold (FDR ≤ 0.05) was used to define significant TWAS associations. snRNA-seq from each cell subtype (sample size 100 or greater) was also imputed for the TWAS analysis.

Statistics and reproducibility

Statistical analyses were performed using R (v.4.0.0 or v.4.2.2) and are described in the figure legends.

Reporting summary

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

Data availability

Raw sequence and processed data of single-nucleus RNA-seq are available at Synapse (https://www.synapse.org/#!Synapse:syn31512863). To preserve the anonymity of study participants, access to the datasets

is restricted and requires a data use certificate (DUC) to be submitted. To submit a DUC, please see https://adknowledgeportal.synapse.org/Data%20Access. Cell (sub)type annotation and cell (sub) type-level eQTL summary statistics are available at Synapse (https://doi.org/10.7303/syn52335732). The cell type-level eQTLs of this study are also available from https://vmenon.shinyapps.io/rosmap_snrnaseq_eqtl/. The VCF files of the WGS are available at Synapse (https://www.synapse.org/#!Synapse:syn11724057). The reference human genome GRCh37 used for WGS variant calling is available from the European Molecular Biology Laboratory-European Bioinformatics Institute (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_g1k_v37.fasta.gz).

Code availability

All software used in the study is publicly available as described in the Methods and Reporting Summary. The custom code to prepare the pseudobulk gene expression and mapping *cis*-eQTL can be found at Zenodo (https://zenodo.org/records/10472216) and GitHub (https://github.com/masashi-CU/snuc-eQTL).

References

44. Bennett, D. A. et al. Overview and findings from the rush Memory and Aging Project. *Curr. Alzheimer Res.* **9**, 646–663 (2012).
45. Mostafavi, S. et al. A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease. *Nat. Neurosci.* **21**, 811–819 (2018).
46. De Jager, P. L. et al. A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Sci. Data* **5**, 180142 (2018).
47. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
48. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
49. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* **39**, 1–13 (2011).
50. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
51. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* **8**, 329–337 (2019).
52. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
53. van Rheenen, W. et al. Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *Nat. Genet.* **53**, 1636–1648 (2021).
54. Lee, J. J. et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).
55. Adams, H. H. H. et al. Novel genetic loci underlying human intracranial volume identified through genome-wide association. *Nat. Neurosci.* **19**, 1569–1582 (2016).
56. Hibar, D. P. et al. Novel genetic loci associated with hippocampal volume. *Nat. Commun.* **8**, 13624 (2017).
57. Satizabal, C. L. et al. Genetic architecture of subcortical brain structures in 38,851 individuals. *Nat. Genet.* **51**, 1624–1636 (2019).
58. Grasby, K. L. et al. The genetic architecture of the human cerebral cortex. *Science* **367**, eaay6690 (2020).
59. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
60. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
61. Gusev, A. et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat. Genet.* **50**, 538–548 (2018).

Acknowledgements

We thank all the participants of the ROS/MAP study for their participation and generous donation of brains. This work was funded by NIH grant nos. U01AG061356 (P.L.D.J./D.A.B.), RF1AG057473 (P.L.D.J./D.A.B.) and U01AG046152 (P.L.D.J./D.A.B.) as part of the AMP-AD consortium, as well as NIH grant nos. R01AG066831 (V.M.), K25DK128563 (A.K.) and U01AG072572 (P.L.D.J./St George-Hyslop).

Author contributions

V.M. and P.L.D.J. conceptualized the study. M.F., Z.G., L.Z., C.M., C.C.W., B.N., G.S.G., O.R.-R., D.P., L.A.-Z., H.L., R.V.P., A.K., B.N.V., K.K., C.J.Y., H.-U.K., G.W., A.R., N.H., J.A.S., Y.W., T.Y.-P., S.M., D.A.B., V.M. and P.L.D.J. carried out the investigation. D.A.B., V.M. and P.L.D.J. acquired the funding. V.M. and P.L.D.J. supervised the study. M.F., Z.G., L.Z. and P.L.D.J. wrote the original manuscript draft. M.F., C.C.W., B.N., G.S.G., K.K., C.J.Y., H.-U.K., T.Y.-P., D.A.B. and P.L.D.J. reviewed and edited the manuscript draft.

Competing interests

A.R. is a cofounder and equity holder of Celsius Therapeutics, is an equity holder in Immunitas and was a scientific advisory board member of Thermo Fisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics and Asimov until 31 July 2020. Since 1 August 2020, A.R. is an employee of Genentech with equity in Roche. O.R.-R. has been an employee of Genentech since 19 October 2020. She has given many lectures on the subject of single-cell genomics to a wide variety of audiences and, in some cases, received remuneration to cover time and costs. O.R.-R. and A.R. are coinventors on patent applications filed at the Broad Institute of MIT and Harvard related to single-cell genomics. Since 3 May 2021, D.P. is an employee of Genentech with equity in Roche. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-01685-y.

Correspondence and requests for materials should be addressed to Philip L. De Jager.

Peer review information *Nature Genetics* thanks Inge Holtman and the other anonymous reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Philip L. De Jager

Last updated by author(s): Jan 10, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	BWA-mem (v0.7.8), GATK HaplotypeCaller (v3.4.0), Picard LiftoverVcf (v2.18.14), CellRanger software (v6.0.0)
Data analysis	demuxlet (v0.1-beta), freemuxlet (v0.1-beta), bcftools (v1.9), subset-bam (v1.1.0), Picard CrosscheckFingerprints (v2.25.4), Seurat (version 4), edgeR (version 3.30.3), limma (version 3.44.3), sva (versions v3.34.0 and 3.36.0), Matrix eQTL (version 2.3), PLINK (version 1.90), epitools (version 0.5-10.1), qvalue (v2.15.0), PheWAS (0.99.5.5), Kallisto (v0.43.1), Sleuth (v0.30.0), Eagle (v2.4), COLOC (version 5.1.0), FUSION (ver. Oct. 1, 2019), UCSC genome browser (as of March 11th, 2022). Custom code for preparing pseudobulk gene expression and mapping cis-eQTL can be found at Zenodo (https://zenodo.org/records/10472216) and GitHub (https://github.com/masashi-CU/snuc-eQTL).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Raw sequence and processed data of single-nucleus RNA-seq are available at Synapse (https://www.synapse.org/#!Synapse:syn31512863). To preserve anonymity of study participants, access to the datasets is restricted and requires a Data Use Certificate (DUC) to be submitted. Cell (sub)type annotation and cell (sub)type-level eQTL summary statistics are available at Synapse (https://doi.org/10.7303/syn52335732). Cell type-level eQTL of this study are also available from the website: https://nircweb.neuro.columbia.edu/snuc-eqtl/. Variant Call Format (VCF) files of WGS are available at Synapse (https://www.synapse.org/#!Synapse:syn11724057). The reference human genome GRCh37 that was used for WGS variant calling is available from EMBL-EBI (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_g1k_v37.fasta.gz)

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Most findings of this study will apply to both male and female because cis-eQTL mapping was performed using participants of both sex while adjusting for sex as a covariate. Sequencing batches were designed to balance sex as much as possible. Although sex is self-reported here, expression levels of XIST RNA was consistent with self-reported sex in all participants (n = 424). There are 136 males and 288 females. Disaggregated sex of ROSMAP participants can be found at https://www.synapse.org/#!Synapse:syn3191087.
Reporting on race, ethnicity, or other socially relevant groupings	Self-reported race of all participants (n = 424) was white.
Population characteristics	Population characteristics are described in Supplementary Table 1. All participants were free from dementia at the time of enrollment to the ROSMAP study.
Recruitment	The Religious Orders Study (ROS) enrolls Catholic nuns, priests and brothers, from more than 40 groups across the United States. The Memory and Aging Project (MAP) study primarily enrolls residents of continuous care retirement communities throughout northeastern Illinois. All participants are without known dementia and agree to annual clinical evaluation and brain donation.
Ethics oversight	Study was approved by an Institutional Review Board of Rush University Medical Center.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed. Sample size was maximized with a limited budget. Our sample size (n = 424) will be sufficient because typical eQTL study requires at least 100 samples.
Data exclusions	Single-nucleus RNA-seq libraries were constructed from 479 brain specimens. Fifty-five of them were excluded because of lack of WGS, ambiguity in sample demultiplexing, likely sample swap, shallow sequencing metrics, and duplicated participants. The exclusion criteria were not pre-established. Instead, they were chosen by inspecting actual distribution of data and spotting outliers.
Replication	To maximize the statistical power of eQTL discovery with a limited budget, replication experiment was not performed. Findings of our study was partly reproduced in a similar study of smaller sample size.
Randomization	Randomization is not relevant because this is an observational study.
Blinding	Blinding is not relevant because this study is data-generation effort and free from specific hypothesis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	iPSC were derived from 48 ROSMAP participants, of which 34 were female, and 14 were male.
Authentication	All cell lines were STR-profiled to validate identity.
Mycoplasma contamination	All cell lines tested negative for monthly mycoplasma contamination test.
Commonly misidentified lines (See ICLAC register)	Commonly misidentified lines were not used in this study.