ANALYSIS
https://doi.org/10.1038/s41588-018-0238-1

nature genetics

Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility

Towfique Raj[1,2]*, Yang I. Li[3], Garrett Wong[1,2], Jack Humphrey[4,5], Minghui Wang[2], Satesh Ramdhani[1,2], Ying-Chih Wang[2], Bernard Ng[6,7], Ishaan Gupta[1], Vahram Haroutunian[8,9], Eric E. Schadt[2], Tracy Young-Pearse[10], Sara Mostafavi[6,7], Bin Zhang[2], Pamela Sklar[2,8], David A. Bennett[11] and Philip L. De Jager[12,13]*

Here we use deep sequencing to identify sources of variation in mRNA splicing in the dorsolateral prefrontal cortex (DLPFC) of 450 subjects from two aging cohorts. Hundreds of aberrant pre-mRNA splicing events are reproducibly associated with Alzheimer's disease. We also generate a catalog of splicing quantitative trait loci (sQTL) effects: splicing of 3,006 genes is influenced by genetic variation. We report that altered splicing is the mechanism for the effects of the *PICALM, CLU* and *PTK2B* susceptibility alleles. Furthermore, we performed a transcriptome-wide association study and identified 21 genes with significant associations with Alzheimer's disease, many of which are found in known loci, whereas 8 are in novel loci. These results highlight the convergence of old and new genes associated with Alzheimer's disease in autophagy–lysosomal-related pathways. Overall, this study of the transcriptome of the aging brain provides evidence that dysregulation of mRNA splicing is a feature of Alzheimer's disease and is, in some cases, genetically driven.

Alternative splicing is an important posttranscriptional regulatory mechanism through which pre-mRNA molecules can produce multiple distinct mRNAs. Alternative splicing affects over 95% of human genes[1], contributing to the functional diversity and complexity of proteins expressed in tissues[2]. Alternative splicing is abundant in human nervous system tissues[3] and contributes to phenotypic differences within and between individuals: at least 20% of disease-causing mutations may affect pre-mRNA splicing[4]. Mutations in RNA-binding proteins (RBPs) involved in splicing regulation and aberrant splicing have been linked to amyotrophic lateral sclerosis[5] and autism[6]. Furthermore, disruptions in RNA metabolism, including mRNA splicing, are associated with age-related disorders, such as frontotemporal lobar dementia[7], Parkinson's disease[8] and Alzheimer's disease[9,10]. These previous studies have mostly focused on alternative splicing of selected candidate genes, including the amyloid precursor protein (*APP*)[8] and microtubule-associated protein tau (*MAPT*)[8,9,11]. However, proteomic profiles of brains from patients with Alzheimer's disease identified increased aggregation of insoluble U1 snRNP, a small nuclear RNA component of the spliceosomal complex, suggesting that the core splicing machinery may be altered in Alzheimer's disease[12]. Apart from these studies, there have been few investigations of the possibility that more widespread splicing disruption affects brain transcriptomes in Alzheimer's disease[13]. However, a comprehensive study of cis- and trans-acting genetic factors that regulate alternative splicing in aging brains is lacking.

Over 24 genetic loci have now been associated with Alzheimer's disease susceptibility by genome-wide association studies (GWAS)[14], and these variants are enriched for associations with gene expression levels in peripheral myeloid cells and often lie within cis-regulatory elements[15]. For example, we reported that one of these variants influences splicing of *CD33*[16]. Given the high abundance of alternative splicing in the brain, we hypothesized that other genetic variants associated with Alzheimer's disease could also affect pre-mRNA splicing, possibly by disrupting efficient binding of splicing factors.

Here, by applying state-of-the-art analytic methods, we generated a comprehensive genome-wide map of splicing variation in the aging prefrontal cortex. We use this map to identify: (1) aberrant mRNA splicing events related to Alzheimer's disease; (2) a new reference of thousands of genetic variants influencing local mRNA splicing in the brain; (3) trans-acting splicing factors that are involved in intron excision; and (4) association of GWAS findings to specific genes within each Alzheimer's disease susceptibility locus. Overall, we deepen our understanding of genetic regulation in the transcriptome of the aging brain and provide a foundation for

[1]Ronald M. Loeb Center for Alzheimer's Disease, Department of Neuroscience and Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [2]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [3]Section of Genetic Medicine, University of Chicago, Chicago, IL, USA. [4]Genetics Institute, University College London, London, UK. [5]Department of Neurodegenerative Disease, Institute of Neurology, University College London, London, UK. [6]Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada. [7]Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada. [8]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [9]James J. Peters VA Medical Center, New York, NY, USA. [10]Ann Romney Center for Neurologic Diseases, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [11]Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL, USA. [12]Center for Translational and Computational Neuroimmunology, Department of Neurology, Columbia University Medical Center, New York, NY, USA. [13]The Broad Institute of MIT and Harvard, Cambridge, MA, USA. *e-mail: towfique.raj@mssm.edu; pld2115@cumc.columbia.edu

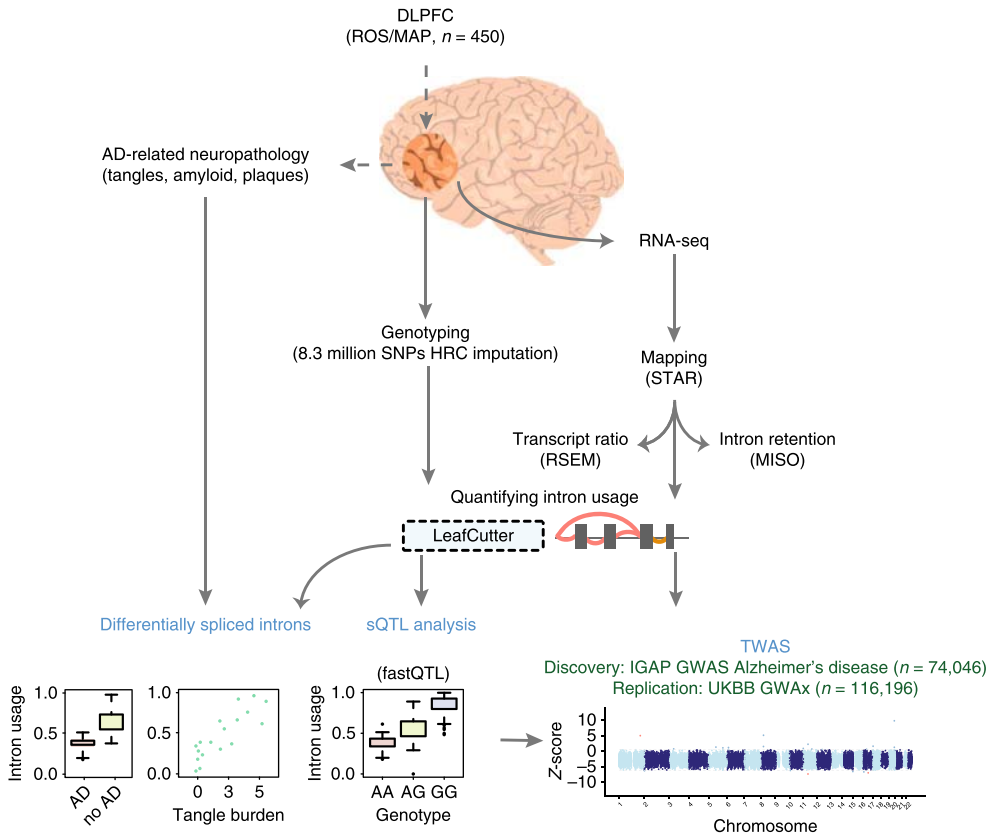

Fig. 1 | Overview of the study. RNA was sequenced from the gray matter of the DLPFC of 542 samples (450 remained after quality control and matching for genotype data) from the ROS/MAP cohort. RNA-seq data were processed, aligned and quantified by our parallelized pipeline. The intronic usage ratios for each cluster were then computed using LeafCutter[20], standardized (across individuals) and quantile normalized. The intronic usage ratios were used for differential splicing analysis, for calling sQTLs and for TWAS. TWAS was performed on summary statistics from IGAP Alzheimer's disease (AD) GWAS of 74,046 individuals[14].

the formulation of mechanistic hypotheses for Alzheimer's disease and other neurodegenerative diseases.

Results

Aberrant mRNA splicing in Alzheimer's disease. We deeply sequenced RNA from frozen DLPFC samples obtained at autopsy from 450 participants of either the Religious Order Study (ROS) or the Memory and Aging Project (MAP), two prospective cohort studies of aging that include brain donation. All subjects were without known dementia at study entry. During the study, some subjects experienced cognitive decline, and, at autopsy, they displayed a range of amyloid-β and tau pathology, with 60% of subjects having a pathologic diagnosis of Alzheimer's disease[17,18] (Supplementary Table 1). We have previously reported changes in RNA expression level in relation to Alzheimer's disease in these data[19].

Following alignment and quantification of RNA-sequencing (RNA-seq) reads, LeafCutter[20] was applied to estimate the 'percentage spliced-in' values (PSI) for local alternative splicing events (Fig. 1). We identified 53,251 alternatively spliced intronic excision clusters in 16,557 genes. We report more alternatively spliced intron clusters in the cortex than any other previously analyzed tissues or brain regions[21]. To identify aberrant splicing events, we analyzed the association between the PSI of each intron excision event and a pathologic diagnosis of Alzheimer's disease or quantitative analyses of neuropathology, including neuritic plaques, neurofibrillary tangles and amyloid-β burden, while accounting for confounding factors. At a false-discovery rate (FDR) < 0.05, we identified a total of 82 differentially spliced introns in 67 genes associated with different neuropathologies, including 5 associated with neuritic plaques, 20 associated with amyloid-β burden and 48 with neurofibrillary tangles (Supplementary Table 2). A heat map of the top differentially spliced introns associated with neurofibrillary tangles is shown in Fig. 2a. On average, these differentially excised introns explain around 2–13% of total variation in neuropathologic burden after accounting for biological and technical covariates (Fig. 2b and Supplementary Fig. 1).

To test for association with the clinical diagnosis of Alzheimer's disease, we used LeafCutter[20] to identify differentially spliced introns by jointly modeling intron clusters using a Dirichlet-multinomial generalized linear model (Methods). At a Bonferroni-corrected rate of $P < 0.05$, we identified a total of 87 intron clusters (corresponding to 84 genes) that displayed altered splicing in relation to Alzheimer's disease (Supplementary Table 3). Of these, 11 genes were also differentially expressed, suggesting that our splicing analysis is identifying novel associations that had been missed in conventional approaches evaluating gene expression levels alone. For example, the most significant differentially excised intron (chr10:3,147,351–3,147,585) is found in the phosphofructokinase gene (*PFKP*): the frequency of this splicing event was associated with Alzheimer's disease ($P < 4.9 \times 10^{-24}$; $\beta = -0.27$) and all pathologic analyses that were tested in this study. Similarly, the next most differentially excised intron (chr14:21,490,656–21,491,400) associated with Alzheimer's disease is found in the α/β-hydrolase fold protein gene *NDRG* family member 2 (*NDRG2*) ($P < 5.6 \times 10^{-19}$; $\beta = -0.058$) and is also associated with measures of both amyloid-β and tau pathology (Fig. 2c). Differential splicing of both *PFKP* and

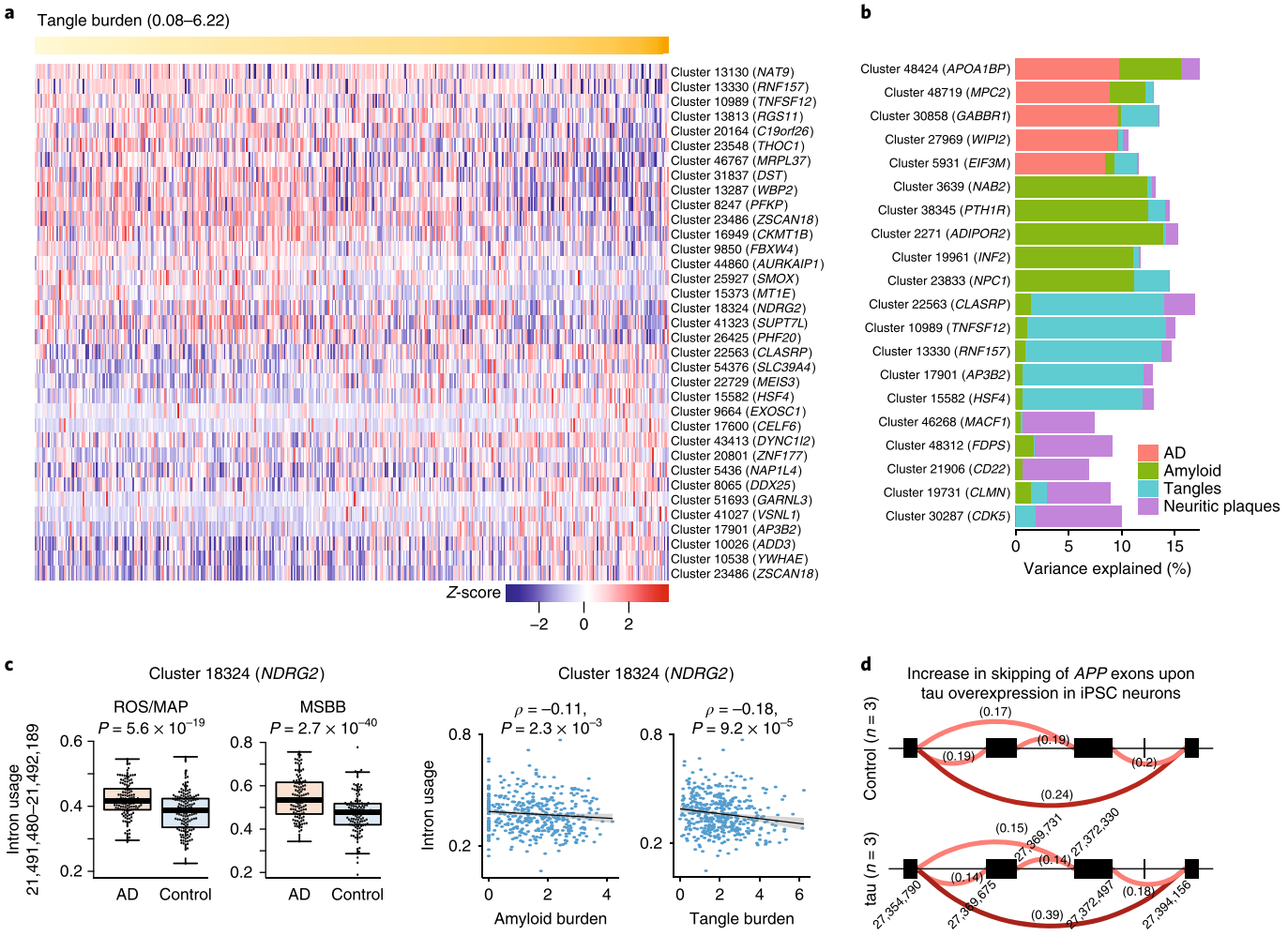

Fig. 2 | Differential splicing analysis in relation to Alzheimer's disease diagnosis and neuropathology. a, Heat map of top 35 differently excised intron association with burden of tangles in ROS/MAP. Each column is one subject; subjects are ordered by their burden of tangles (yellow row at the top of the panel). The *Z*-score strength and direction of the associations are denoted using the key at the bottom of the panel. **b**, Percentage of explained variance of the top five differently excised introns association for four different traits. **c**, The left two panels present the mean and distribution of intron usage for differently excised introns in *NDRG2* in relation to a clinical diagnosis of Alzheimer's disease (AD) in ROS/MAP and in MSBB. The right two panels display the association of amyloid-β or tangle burden to intron usage in *NDRG2*. **d**, Differentially excised intron in *APP* upon tau overexpression in iPSC neurons.

NDRG2 in human brains has been previously shown to be associated with Alzheimer's disease pathogenesis[22,23], offering a measure of replication. Other genes with differentially excised introns associated with Alzheimer's disease at a Bonferroni-corrected $P<0.05$ include *APP* ($P<1.6\times10^{-3}$; $\beta=-0.003$) and genes in known GWAS loci such as *PICALM* ($P<0.02$; $\beta=0.005$) and *CLU* ($P<3.2\times10^{-4}$; $\beta=-0.019$). These differentially spliced genes are not necessarily expressed in a single cell type (for example, neurons), but are expressed across many cell types, including astrocytes (Supplementary Figs. 2, 3 and Supplementary Table 4). Moreover, co-splicing network analysis using WGCNA[24] suggests that the differentially spliced genes are enriched in specific functional modules and are part of a coherent biological process (Supplementary Figs. 2, 3, Supplementary Tables 5–7 and Supplementary Note).

Next, to assess the robustness of our results, we performed a replication analysis using RNA-seq data from the Mount Sinai Brain Bank (MSBB)[25], which includes 301 brain samples from patients with Alzheimer's disease and controls (Supplementary Note). Of the 84 genes with differentially spliced intron clusters in ROS/MAP, 52 (including *APP*, *PFKP* and *NDRG2*) were significant at a Bonferroni-corrected $P<0.05$ threshold in the MSBB data and the effect sizes are highly correlated between the two datasets (Pearson's $r=0.35$; $P<6.75\times10^{-14}$) (Fig. 2c, Supplementary Fig. 4 and Supplementary Table 8). This constitutes an independent replication of specific, aberrant splicing alterations in brains from patients with Alzheimer's disease. Finally, to further validate and explore the mechanism of our observations, we analyzed RNA-seq data derived from induced pluripotent stem cell (iPSC)-derived neurons and neurons from the same line overexpressing tau: differential intron excision was noted for 42 genes (FDR<0.05), of which 11 genes overlap with the Alzheimer's disease-associated splicing in the cortex including *APP*, *PICALM* and *NDRG2* (Fig. 2d and Supplementary Table 9). Despite the small sample size, these in vitro data suggest that tau accumulation in neurons—at a stage in which neurons are accumulating phosphorylated tau but are not apoptotic—may be sufficient to induce some of the splicing alterations that we observed in cortical tissue of human subjects; this in vitro validation of disease-related splicing changes suggests that this Alzheimer's disease-altered splicing (1) is unlikely to be related to confounding factors from autopsy or the agonal state and (2) has specific target RNAs that can be modeled in vitro.

Genetic effects on pre-mRNA splicing in aging brains. We next performed a sQTL study to identify local genetic effects that drive

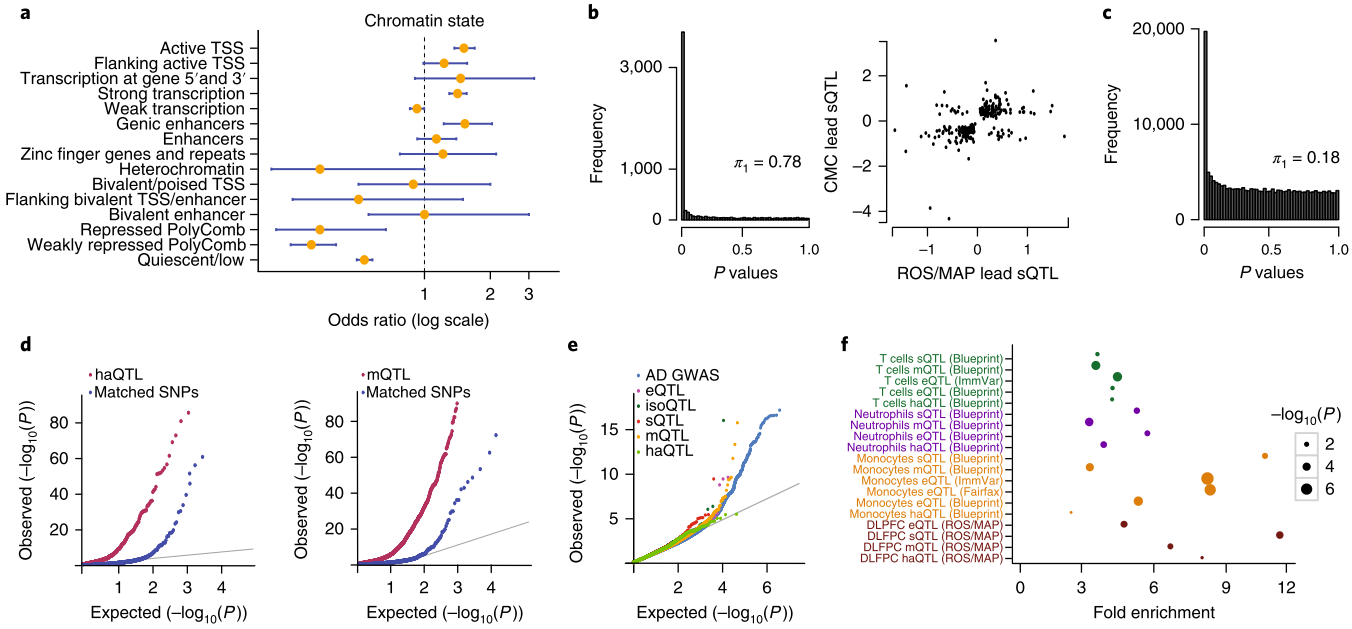

Fig. 3 | Enrichment of sQTLs in epigenomic marks and in Alzheimer's disease GWAS. a, sQTLs are enriched in regions (or chromatin states) associated with active transcription and genic enhancers, and they are depleted in polycomb regions that are transcriptionally repressed in the DLPFC. TSS, transcription start site. **b**, Left, *P*-value distribution of ROS/MAP sQTLs that are significant in CMC (FDR < 0.05). The majority (78%) of sQTLs in ROS/MAP are also discovered in CMC. Right, the direction of effect is consistent for the majority (93%) of the significant (FDR < 0.05) lead sQTLs in CMC and ROS/MAP. **c**, *P*-value distribution of ROS/MAP eQTLs that are significant sQTLs (FDR < 0.05). **d**, SNPs that drive QTLs in H3K9ac and DNA methylation data in the same ROS/MAP brains are more likely to be sQTLs than matched SNPs within H3K9ac domains (left) and near DNA methylated CG (right). **e**, *Q*–*Q* plot for Alzheimer's disease GWAS suggests that sQTLs are enriched among Alzheimer's disease GWAS (IGAP study[14]) compared to other types of QTLs. **f**, Fold enrichment of Alzheimer's disease GWAS SNPs (GWAS $P < 10^{-5}$) among QTL SNPs driving variation in gene expression, splicing, histone acetylation and DNA methylation in primary monocytes[15,31,50], T cells[15,31] or DLPFC[29].

variation in RNA splicing in the DLPFC. First, we assessed the splice events from the LeafCutter[20] algorithm (Fig. 1); 30% of these 53,251 intron excision clusters are novel splicing events, not previously reported in other sQTL studies. The PSI values were adjusted for known and hidden factors (15 principal components) and then fit to imputed SNP data using an additive linear model implemented in fastQTL[26] (Methods and Supplementary Fig. 5). At FDR < 0.05, we found 9,028 sQTLs in 3,006 genes (Supplementary Table 10). As expected, splicing was most strongly affected by variants in the splice region itself (59.8%): 20.2% of variants are mapped to splice acceptor sites and 16.4% to splice donor sites. The remaining (23.2%) mapped to other splice regions or are found within an intron (Supplementary Fig. 6). Furthermore, sQTLs are mapped to distinct regulatory features as defined by 15 chromatin states in DLPFC[27]: sQTLs were significantly enriched in actively transcribed regions and enhancers. They are depleted in repressed chromatin marked with polycomb, heterochromatin and quiescent regions (Fig. 3a), consistent with the diminished transcription noted in these regions.

To assess the extent of sQTL replication, we compared our sQTLs to the recently published dataset from the CommonMind Consortium (CMC), consisting of DLPFC profiles from 258 individuals with schizophrenia and 279 control subjects[28] (Supplementary Note). Our sQTLs yield a Storey's $\pi_1 = 0.78$ in the CMC data, suggesting substantial sharing of sQTLs between these two different brain collections (Fig. 3b). Moreover, 93% of shared sQTLs showed the same direction of effect (Fig. 3b). The fraction of novel sQTLs deserves further evaluation to assess the extent to which they may be context-specific given that the average age at death of our participants (88 years) is significantly older than that of the CMC dataset.

In agreement with recent findings in lymphoblastoid cell lines[21], we found that a majority of sQTLs act independent of gene expression effects, as evident by the low degree of sharing between sQTLs and expression (e)QTLs[29] from the same brains ($\pi_1 = 0.18$) (Fig. 3c). Of the 9,045 lead sQTL SNPs, only 42 are also a lead eQTL, suggesting that a substantial fraction of sQTLs are unique and are not detected by standard eQTL analysis.

To further understand the mechanisms underlying sQTLs, we assessed the overlap of sQTLs with SNPs influencing epigenomic marks (xQTLs), such as DNA methylation (mQTL) and histone H3 acetylation on lysine 9 (H3K9ac, haQTL)[29], that are available from the same DLPFC samples. Indeed, we found that such xQTLs[29] are significantly enriched among sQTLs when compared to randomly selected, matched SNPs (Kolmogorov–Smirnov test $P < 0.001$): of the lead sQTL, 9% (578) were also associated with an haQTL and 19% (1,246) were also an mQTL (Fig. 3d). This suggests that an important subset of genetic variants co-influences splicing, methylation levels and histone modifications. In a complementary analysis, we found significant sharing of sQTL SNPs among SNPs that also influence histone acetylation ($\pi_1 = 0.74$) or methylation ($\pi_1 = 0.82$). These overlaps suggest that there is a contribution of epigenomic regulation in splicing.

Given prior reports[21,30], we evaluated whether our sQTLs from the aging brain were enriched for Alzheimer's disease susceptibility variants (Fig. 3e,f). We also assessed enrichment of Alzheimer's disease SNPs (GWAS $P < 1 \times 10^{-5}$) in splicing, methylation or expression QTLs from DLPFC[29], monocytes[15,31], neutrophils[31] and T cells[15,31]. We found that DLPFC sQTLs are more likely to be enriched for Alzheimer's disease GWAS SNPs, followed by sQTL and eQTL from monocytes (Fig. 3f). These findings highlight (1) the important role of RNA splicing on variation in Alzheimer's disease

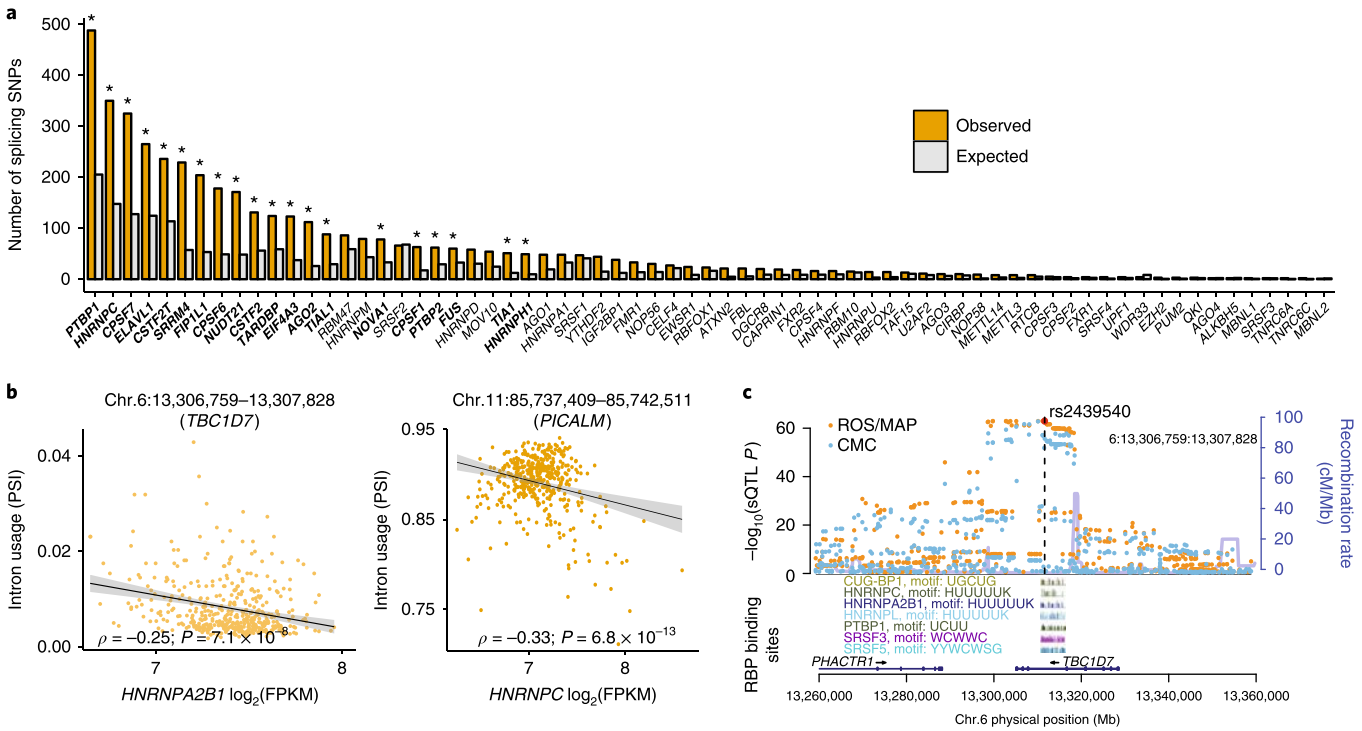

Fig. 4 | Enrichment of RBP binding sites among sQTLs. a, RBP enrichment (expected versus observed) among the lead sQTLs. Significant RBP enrichment are shown in bold. *$P < 0.001$. **b**, Association of *HNRNPA2B1* (left) and *HNRNPC* (right) gene expression levels with differential intron usage in *TBC1D7* (left) and in *PICALM* (right). **c**, Regional plot of sQTL results for SNPs in the vicinity of *TBC1D7* (chr.6:13,306,759–13,307,828). SNPs driving sQTLs for *TBC1D7* overlap CLIP binding sites (from CLIPdb[34]) for several splicing factors. The top SNP (rs2439540, red) overlaps motifs for a number of RBPs. sQTL results are highly consistent between ROS/MAP (orange) and CMC (blue) data.

susceptibility, (2) the prominent role of myeloid cells in Alzheimer's disease susceptibility[15] and (3) the fact that a number of Alzheimer's disease variants have mechanisms that may be mediated through non-myeloid effects.

Some of these effects of Alzheimer's disease variants on splicing are known, such as the eightfold increase in full-length *CD33* isoform expression[16,32], but several of these variants—in *CLU*, *PICALM* and *PTK2B*—have not been previously reported (Supplementary Table 10). These results delineate the initial events along the cascade of functional consequences for these three Alzheimer's disease variants and provide important mechanistic insights into their development as potential therapeutic targets.

Splicing regulators associated with alternative splicing. Splicing of pre-mRNA is catalyzed by a large ribonucleoprotein complex called the spliceosome, which consists of five small nuclear RNAs and numerous splicing factors[33]. To identify brain splicing factors that regulate sQTL events in trans, we evaluated whether the lead sQTL SNPs that were identified in our study are enriched in RBP binding sites using publicly available cross-linking immunoprecipitation–sequencing (CLIP–seq) datasets from 76 RBPs in CLIPdb[34]. We found that binding targets of 18 RBPs are significantly enriched among lead sQTLs (Fig. 4a). The most enriched RBP is *PTBP1*, followed by *HNRNPC*, *CPSF7* and *ELAVL1* ($P < 0.05$, Fisher's exact test with Benjamini–Hochberg correction). Notably, the enrichment for neuronal *ELAVL1* RBP target sites is consistent with a recent report that, upon neuronal *ELAVL1* depletion, *BIN1* and *PICALM* transcripts were found to have lower exon inclusion for those sites in which *ELAVL* binding sites directly overlapped with SNPs associated with Alzheimer's disease[35].

On the other hand, we also observed significant enrichment for the lead sQTL SNPs within the binding sites for a number of heterogeneous nuclear ribonucleoproteins (hnRNP), including hnRNP C ($P < 0.009$). Furthermore, we find that the expression levels of hnRNP splicing factors are correlated with intronic excision levels of hundreds of genes, many of which are in Alzheimer's disease susceptibility loci including *BIN1*, *PICALM*, *APP* and *CLU* (Fig. 4b and Supplementary Figs. 7, 8). The hnRNP C factor has been linked to Alzheimer's disease in previous studies, including in a recent biochemical study that described the translational regulation of *APP* mRNA by hnRNP C[36]. This observation goes towards the mechanism of the sQTL: consistent with the assumption that altering the sequence of a binding site changes the likelihood that a splicing event occurs in vivo. In one example of an sQTL that affects intron usage, a SNP within an intron of *TBC1D7* is found within CLIP-defined binding sites for hnRNP C as well as other RBPs (Fig. 4c). Thus, incorporating RBP binding sites as a functional annotation enables us to improve our accuracy when selecting plausible variants that may disrupt binding of splicing factors that cause the alternative-splicing event. Further biochemical studies will be required to understand the full regulatory program that orchestrates the disease-related splicing changes.

Transcriptome-wide association studies prioritize Alzheimer's disease genes. To identify genes for which mRNA expression or alternative splicing is associated with Alzheimer's disease and mediated by genetic variation, we performed two transcriptome-wide association studies (TWAS)[37] using either the ROS/MAP expression data or its intronic excision levels as reference panels to reanalyze summary level data from the International Genomics of Alzheimer's Project (IGAP) GWAS[14]. A total of 4,746 genes and 15,013 differentially spliced introns could be analyzed, and we identified 21 genes at FDR < 0.05 for which imputed gene expression or intronic excision levels were significantly associated with

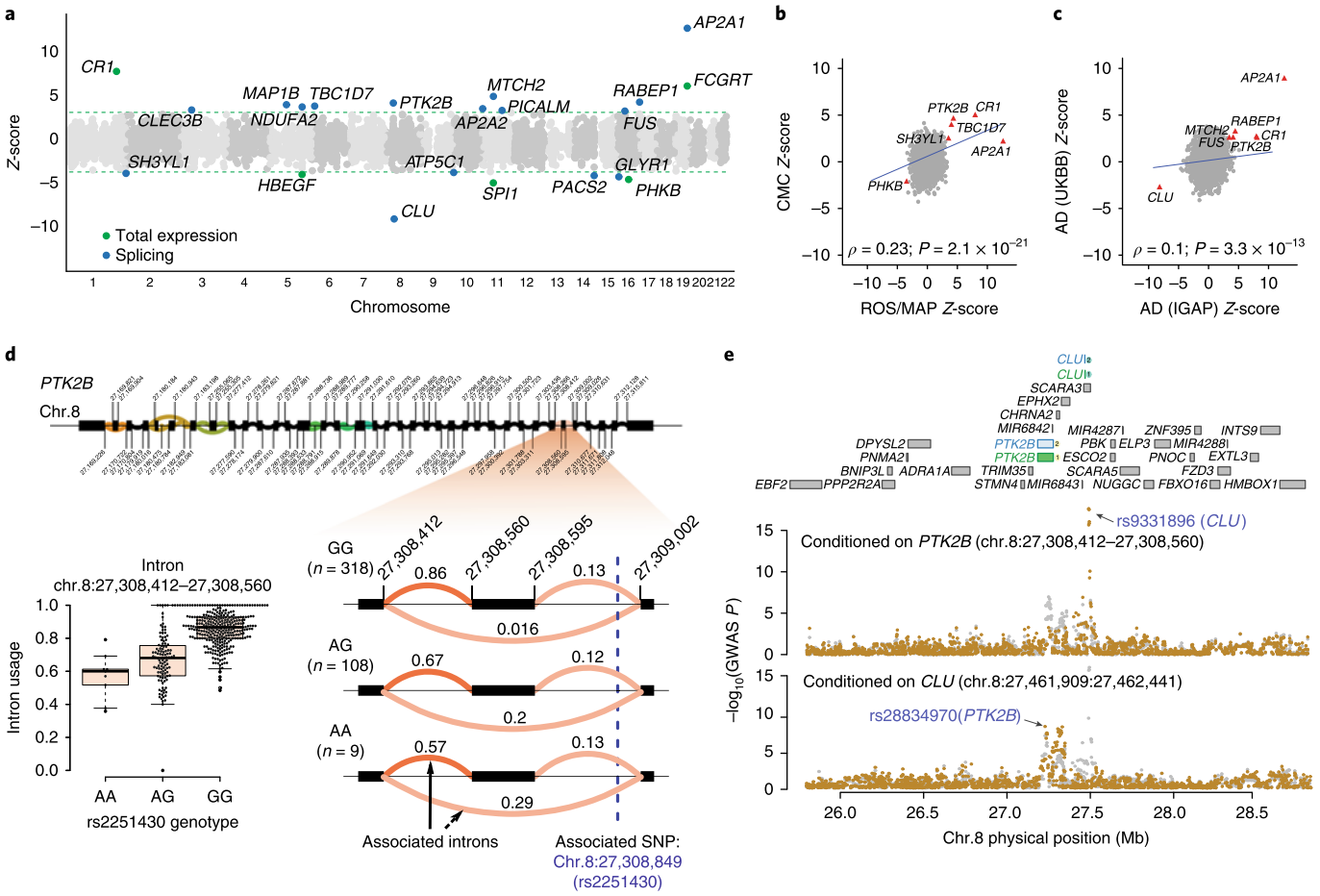

Fig. 5 | TWAS of Alzheimer's disease. a, Transcriptome-wide results using the IGAP GWAS summary statistics; each dot is one gene. The dotted green line denotes the threshold of significance (FDR < 0.05). Genes for which there is evidence of significant differential intron usage are highlighted in blue. In green, we highlight those genes for which the TWAS (total gene expression) results are significant. **b**, Replication of ROS/MAP TWAS in CMC DLPFC data. The red triangles denote genes for which the replication analysis is significant. **c**, Replication of IGAP Alzheimer's disease TWAS using the UKBB GWAS based on an independent set of subjects. **d**, *PTK2B* gene structure (top): clusters of differential splicing events are noted with the colored curves. The panel then zooms to highlight differential intronic usage for chr.8:27,308,412–27,308,560 stratified by rs2251430 genotypes (right). On the left, we show the same data as a box plot. **e**, Conditional analysis of IGAP GWAS results for two splicing effects for *PTK2B* and *CLU* in Alzheimer's disease GWAS data. As noted in the top aspect of the panel, these two Alzheimer's disease genes are located close to one another. The intronic excision events for *PTK2B* and *CLU* are present in both ROS/MAP (blue) and in CMC (green) dataset. When the Alzheimer's disease GWAS is conditioned on the *PTK2B* (chr.8:27,308,412–27,308,560) splicing effect, the *CLU* effect remained significant, demonstrating its independence from the *PTK2B* association. The reciprocal analysis conditioning on the *CLU* (chr.8:27,461,909–27,462,441) effect, the *PTK2B* association remained significant.

Alzheimer's disease status (Fig. 5a and Supplementary Table 11). Among these, there were genes in known Alzheimer's disease loci including *SPI1*, *CR1*, *PTK2B*, *CLU*, *MTCH2* and *PICALM*. These results help to pinpoint the likely gene that is the target of the known susceptibility variant in each locus, particularly at the *MTCH2* locus in which the functional consequences of the risk allele were unclear. However, the new genes associated with Alzheimer's disease are even more interesting, and eight of these associations are found in loci that showed only suggestive evidence of association in the IGAP study. These genes include *AP2A1*, *AP2A2*, *FUS*, *MAP1B*, *TBC1D7* and others that are now significant at a threshold adjusted for genome-wide testing. This analysis therefore helps to prioritize the long list of suggestive IGAP associations (Fig. 5a and Supplementary Figs. 9–17). Notably, both *AP2A1* and *MAP1B* were recently identified as hub proteins in Alzheimer's disease proteome networks and had lower protein expression levels in brains from patients with Alzheimer's disease compared to those from patients with amyotrophic lateral sclerosis, Parkinson's disease or from control subjects[38].

To replicate these results, we first assessed whether an expression imputation model built using the CMC dataset[28] yielded similar results when it is deployed in the IGAP GWAS data. We focused on the 21 significantly associated genes described here: four genes (*CR1*, *PTK2B*, *TBC1D7* and *SH3YL1*) replicated at FDR < 0.05 and two genes (*AP2A1* and *PHKB*) were suggestive at $P < 0.05$ with the expression and splicing inference from CMC (Fig. 5b). The directions of effect for all six associations were consistent in both datasets (Fig. 5b). Thus, we see replication of our results: they are not due to the unique properties of the ROS/MAP dataset. Second, we used a different Alzheimer's disease GWAS—the UK Biobank (UKBB) GWAS by proxy[39]—to replicate the IGAP TWAS results. We note that, despite analyzing data from 116,196 subjects, the UKBB GWAS is underpowered since the GWAS does not use Alzheimer's disease cases but, rather, subjects who have a first-degree relative with Alzheimer's disease as 'cases'. Nevertheless, we were able to replicate (at a nominal $P < 0.05$) seven of our IGAP TWAS associations in the UKBB TWAS (Fig. 5c). These two complementary replication efforts demonstrate the robustness of our results. Finally, we

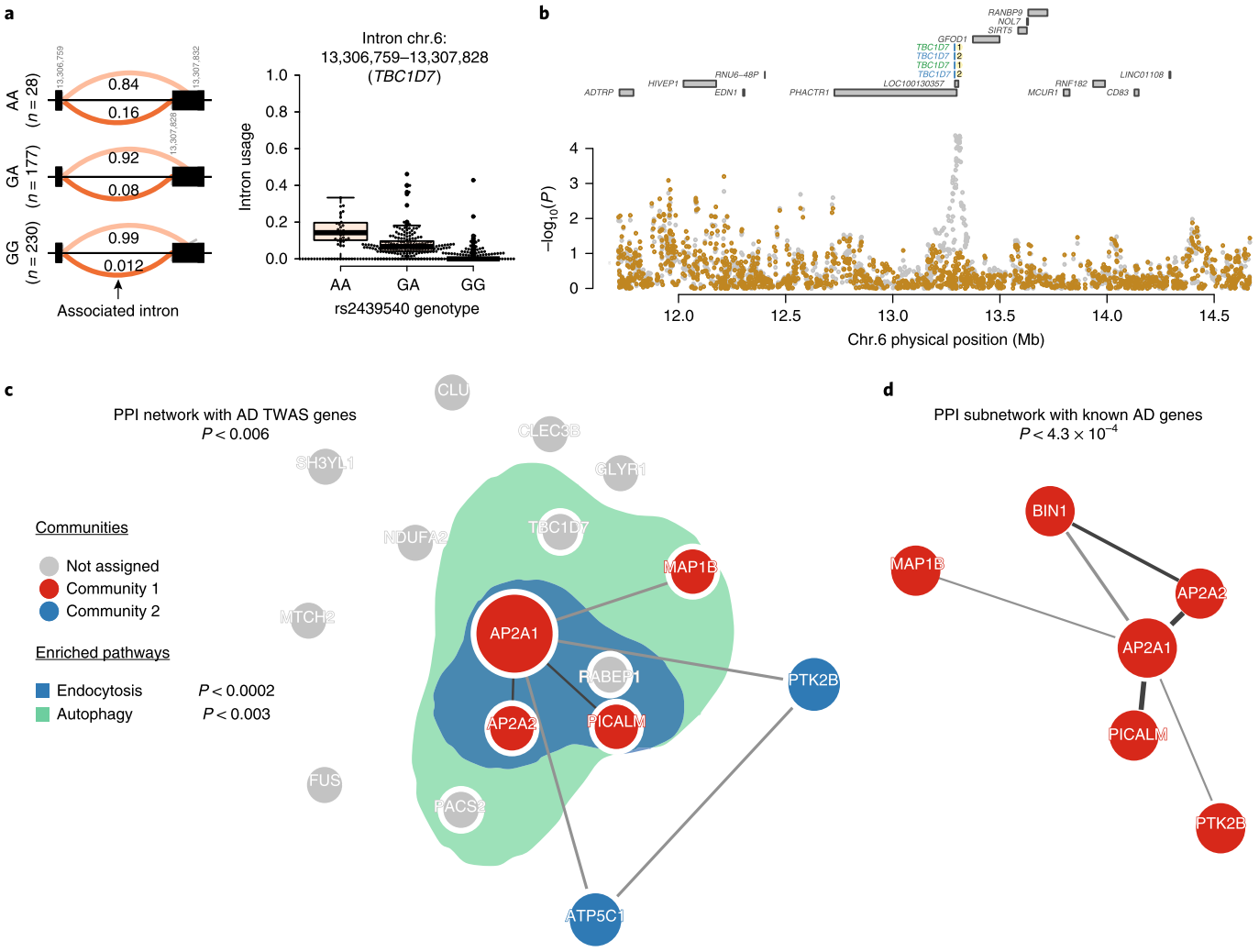

Fig. 6 | TWAS prioritizes Alzheimer's disease genes in endocytosis- and autophagy-related pathways. a, Differential intronic usage for chr.6:13,306,759–13,307,828 (*TBC1D7*) stratified by rs2439540 genotypes (left). Box plot of the same data (right). **b**, Regional plot showing the IGAP *P* values in *TBC1D7* locus. Two intronic excision events at *TBC1D7* are present in both the ROS/MAP (blue) and CMC (green) datasets. The Alzheimer's disease GWAS effect is mostly explained by intronic usage of chr.6:13,306,759–13,307,828. The Alzheimer's disease GWAS at *TBC1D7* is suggestive in the original IGAP study ($P<10^{-5}$). **c**, The product of three of the novel Alzheimer's disease genes (*AP2A2*, *AP2A1* and *MAP1B*) are members of the same PPI network ($P<0.006$). The genes in this network and others not in the network (that is, *TBC1D7*, *PACS2* and *RABEP1*) are significantly enriched in genes annotated as being involved in endocytosis (blue; $P<0.0002$) and autophagy-related pathways (green; $P<0.003$). **d**, The novel Alzheimer's disease genes (*AP2A2*, *AP2A1* and *MAP1B*) form a significant PPI subnetwork ($P<4.3\times10^{-4}$) with known Alzheimer's disease genes (that is, *PICALM*, *BIN1* and *PTK2B*).

performed a TWAS using the summary statistics of a meta-analysis of IGAP and UKBB GWAS, and identified three additional genes (*ABCA7*, *RHBDF1* and *VPS53*) that meet a genome-wide significant threshold in the meta-analysis; *ABCA7* is one of the well-validated Alzheimer's disease loci (Supplementary Table 12).

Most of the TWAS associations are the result of differential intron usage, suggesting the importance of pre-mRNA splicing in Alzheimer's disease (Fig. 5a). As an example, a TWAS association with intron usage at *PTK2B*, a known Alzheimer's disease susceptibility locus, is shown in Fig. 5d. We often observed multiple TWAS-associated genes in the same locus, likely owing to co-expression of genes in close physical proximity or allelic heterogeneity within the susceptibility locus[40]. To account for multiple associations in the same locus, we applied conditional and joint association methods that rely on summary statistics[40,41] to identify genes that had significant TWAS associations when analyzed jointly (Figs. 5e, 6b). A region with multiple TWAS associations includes the *PTK2B–CLU* locus, which shows independent co-localized associations for both GWAS[14] and splicing effects (Fig. 5e).

Refining known associations is important to translate results into functional studies, but the newly validated Alzheimer's disease genes (Fig. 6a,b) also offer new insights into disease: we used GeNets[42] to evaluate the connectivity of these new Alzheimer's disease genes with the network of known susceptibility genes that are interconnected by protein–protein interaction (PPI)[43]. These new and known genes associated with Alzheimer's disease susceptibility are directly connected (that is, they form shared 'communities'; $P<0.006$) (Fig. 6c). Furthermore, this joint network is enriched for endocytosis pathways ($P<0.0002$), highlighting the existing suggestion that endocytosis pathways are preferentially targeted in Alzheimer's disease. The enrichment for an autophagy–lysosomal-related pathway ($P<0.003$) is more interesting (Fig. 6c). The genes in the autophagy–lysosomal-related pathway (*AP2A2*, *AP2A1* and *MAP1B*) form a statistically significant ($P<4.3\times10^{-4}$) PPI subnetwork with known Alzheimer's disease genes (*PTK2B*, *PICALM* and *BIN1*) (Fig. 6d). This is consistent with prior work that has implicated protein-degradation pathways in Alzheimer's disease[44].

Overall, these PPI analyses suggest that our new TWAS-derived genes are not a random set of genes but are part of an Alzheimer's disease network.

Discussion

In this study, we directly examined alternative splicing events in a large dataset of aging brains, which led to the observations that specific alternative splicing events are reproducibly associated with Alzheimer's disease and that certain validated genetic associations affect splicing of a nearby gene as the proximal functional consequence of the susceptibility allele. Our replication efforts demonstrate that the observed Alzheimer's disease-related perturbations in splicing are not simply owing to spliceosomal failure: specific genes are reproducibly affected in a specific manner. Furthermore, results from our in vitro model of tau overexpression in iPSC-derived neurons suggest that perturbation of *MAPT* may be sufficient to cause at least a subset of the disease-related splicing changes that are observed in the human cortex at autopsy.

We used the TWAS approach, which leverages our splicing map and common genetic variants to test the hypothesis that the effect of such variants in Alzheimer's disease is mediated by altering splicing levels. These analyses confirmed many of the known Alzheimer's disease genes (that is, *CLU* and *PTK2B*), which supports the role of regulation of splicing levels as key mechanisms in certain loci, but also found several new loci: *TBC1D7*, *AP2A1*, *AP2A2* and *MAP1B* (Figs. 5a,6b and Supplementary Figs. 7, 9, 13). These new genes reinforce the association of the Clathrin–AP2 adaptor complex with Alzheimer's disease susceptibility[45]. Both *AP2A2* and *AP2A1* encode components of the AP2 adaptor complex that serves as a cargo receptor, selectively sorting membrane proteins involved in receptor-mediated endocytosis[46]. The AP2 complex and PICALM interact with APP, directing it towards degradation and autophagy[46].

Our study also offers insights for several well-known Alzheimer's disease loci in which the gene was known but the functional mechanism remained unclear. Similar to our work in *CD33*[16], the careful analysis of these cortical data highlights a specific splicing mechanism for the Alzheimer's disease risk alleles at *CLU*, *PICALM* and *PTK2B*. All three are complex proteins with a large number of exons, and our results prioritize specific domains in these proteins that contain the functional domain that influences Alzheimer's disease risk. Furthermore, our analyses of RBPs involved in splicing regulation of Alzheimer's disease susceptibility genes including *PICALM*, and RNA-binding site analysis of *HNRNPC* (Fig. 4c and Supplementary Fig. 6) and *ELAVL* helps to prioritize the variant that may be driving the genetic association and to elaborate the series of events upstream of the susceptibility variant that enable its expression. Thus, the catalog of splicing variants made available with this study provides a starting point for further focused molecular and biochemical experimental validation to fully elucidate the role of these splicing variants in the etiology of Alzheimer's disease.

This study has several limitations. We only characterize splicing events in one region (the DLPFC) of the aging brain. The DLPFC is a region that is affected by amyloid-β pathology relatively early as it spreads throughout the neocortex[47]. The accumulation of tau pathology progresses in a stereotypic manner captured by the Braak stages[48], and the DLPFC displays accumulation of neurofibrillary tangles containing tau typically when individuals begin to be symptomatic. Thus, both pathological amyloid-β and tau accumulate in the DLPFC in Alzheimer's disease, and we use quantitative measures of these pathologies to enhance our power in discovering the molecular features that are associated with these pathologies. Some of these splicing changes may contribute to the accumulation of pathological amyloid-β or tau whereas others may be a reaction to the presence of pathological amyloid-β or tau accumulation or may be the result of indirect effects of the pathology in other brain regions. Currently, we cannot differentiate these three sources of variation in our results. Expression datasets from multiple brain regions exist in the Genotype-Tissue Expression (GTEx) project[49], but the sample size is too small ($n = 88$–136) to build a robust transcriptome model for TWAS. The MSBB has RNA-seq data across three brain regions but owing to the lack of availability for individual level genotypes, we are unable to build reference models from those data. Another limitation of this study is the small sample size of the in vitro experiment; thus, these intriguing results will require testing in a much larger number of iPSC lines to confirm that this effect of *MAPT* overexpression is generalizable. We note that these *MAPT*-overexpressing iPSC-derived neurons are functioning normally at the time when they were sampled; thus, these data suggest that at least some of the disease-associated splicing changes that we report may occur very early in the series of molecular events that are caused by perturbation in *MAPT* expression.

This transcriptome-wide reference map of RNA splicing in the aging cortex is a new resource that highlights strong effects of neuropathology and genetic variation on splicing. It will be useful in annotating the results of genetic and epigenomic studies of neurologic and psychiatric diseases; however, it has an immediate influence on the identification of the functional consequences of several Alzheimer's disease susceptibility alleles, helps to expand the list of loci involved in Alzheimer's disease, and implicates the protein-degradation machinery in the pathology of Alzheimer's disease.

URLs. ROS/MAP sQTL browser, https://rajlab.shinyapps.io/sQTLviz_ROSMAP/; LeafCutter, https://github.com/davida-knowles/leafcutter; xQTL Browser, http://mostafavilab.stat.ubc.ca/xQTLServe; FUSION, http://gusevlab.org/projects/fusion/; MISO, http://genes.mit.edu/burgelab/miso/; SpliceAid-F, http://srv00.recas.ba.infn.it/SpliceAidF/; Roadmap Epigenomics Project, http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html; GREGOR, http://genome.sph.umich.edu/wiki/GREGOR; GARFIELD, http://www.ebi.ac.uk/birney-srv/GARFIELD; GeNets, https://apps.broadinstitute.org/genets; Michigan Imputation Server, https://imputationserver.sph.umich.edu/index.html; The RUSH Alzheimer's Disease Research Center Research Resource Sharing Hub, https://www.radc.rush.edu; AMP-AD Synapse Portal, https://www.synapse.org/#!Synapse:syn2580853/wiki/409844; CommonMind Consortium Knowledge Portal, https://www.synapse.org/#!Synapse:syn2759792/wiki/69613; IGAP GWAS summary statistics, http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php; UK Biobank summary statistics, http://gwas-browser.nygenome.org/downloads/gwas-browser/.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-018-0238-1.

Received: 23 August 2017; Accepted: 16 August 2018;
Published online: 8 October 2018

References

1. Wang, E. T. et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
2. Kornblihtt, A. R. et al. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat. Rev. Mol. Cell Biol.* **14**, 153–165 (2013).
3. Barbosa-Morais, N. L. et al. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587–1593 (2012).
4. Wang, G. S. & Cooper, T. A. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.* **8**, 749–761 (2007).
5. Dredge, B. K., Polydorides, A. D. & Darnell, R. B. The splice of life: alternative splicing and neurological disease. *Nat. Rev. Neurosci.* **2**, 43–50 (2001).
6. Parikshak, N. N. et al. Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* **540**, 423–427 (2016).

7. Arai, T. et al. TDP-43 is a component of ubiquitin-positive tau-negative inclusions in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Biochem. Biophys. Res. Commun.* **351**, 602–611 (2006).
8. Trabzuni, D. et al. MAPT expression and splicing is differentially regulated by brain region: relation to genotype and implication for tauopathies. *Hum. Mol. Genet.* **21**, 4094–4103 (2012).
9. Rockenstein, E. M. et al. Levels and alternative splicing of amyloid beta protein precursor (APP) transcripts in brains of APP transgenic mice and humans with Alzheimer's disease. *J. Biol. Chem.* **270**, 28257–28267 (1995).
10. Buee, L., Bussiere, T., Buee-Scherrer, V., Delacourte, A. & Hof, P. R. Tau protein isoforms, phosphorylation and role in neurodegenerative disorders. *Brain Res. Rev.* **33**, 95–130 (2000).
11. 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).
12. Bai, B. et al. U1 small nuclear ribonucleoprotein complex and RNA splicing alterations in Alzheimer's disease. *Proc. Natl Acad. Sci. USA* **110**, 16562–16567 (2013).
13. Vaquero-Garcia, J. et al. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife* **5**, e11752 (2016).
14. Lambert, J. C. et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
15. Raj, T. et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science* **344**, 519–523 (2014).
16. Raj, T. et al. CD33: increased inclusion of exon 2 implicates the Ig V-set domain in Alzheimer's disease susceptibility. *Hum. Mol. Genet.* **23**, 2729–2736 (2014).
17. Bennett, D. A., Schneider, J. A., Arvanitakis, Z. & Wilson, R. S. Overview and findings from the Religious Orders Study. *Curr. Alzheimer. Res.* **9**, 628–645 (2012).
18. Bennett, D. A. et al. Selected findings from the Religious Orders Study and Rush Memory and Aging Project. *J. Alzheimers. Dis.* **33**, S397–S403 (2013).
19. 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).
20. Li, Y. I. et al. Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).
21. Li, Y. I. et al. RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).
22. Tollervey, J. R. et al. Analysis of alternative splicing associated with aging and neurodegeneration in the human brain. *Genome Res.* **21**, 1572–1582 (2011).
23. Mitchelmore, C. et al. NDRG2: a novel Alzheimer's disease associated protein. *Neurobiol. Dis.* **16**, 48–58 (2004).
24. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, e17 (2005).
25. Wang, M. et al. Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer's disease. *Genome Med.* **8**, 104 (2016).
26. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
27. Bernstein, B. E. et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
28. Fromer, M. et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19**, 1442–1453 (2016).
29. 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).
30. Nicolae, D. L. et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, e1000888 (2010).
31. Chen, L. et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* **167**, 1398–1414 (2016).
32. Malik, M. et al. CD33 Alzheimer's risk-altering polymorphism, CD33 expression, and exon 2 splicing. *J. Neurosci.* **33**, 13320–13325 (2013).
33. Sibley, C. R., Blazquez, L. & Ule, J. Lessons from non-canonical splicing. *Nat. Rev. Genet.* **17**, 407–421 (2016).
34. Yang, Y. C. et al. CLIPdb: a CLIP-seq database for protein–RNA interactions. *BMC Genomics* **16**, 51 (2015).
35. Scheckel, C. et al. Regulatory consequences of neuronal ELAV-like protein binding to coding and non-coding RNAs in human brain. *eLife* **5**, e10421 (2016).
36. Borreca, A., Gironi, K., Amadoro, G. & Ammassari-Teule, M. Opposite dysregulation of fragile-X mental retardation protein and heteronuclear ribonucleoprotein C protein associates with enhanced APP translation in Alzheimer disease. *Mol. Neurobiol.* **53**, 3227–3234 (2016).
37. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
38. Seyfried, N. T. et al. A multi-network approach identifies protein-specific co-expression in asymptomatic and symptomatic Alzheimer's disease. *Cell Syst.* **4**, 60–72 (2017).
39. Liu, J. Z., Erlich, Y. & Pickrell, J. K. Case–control association mapping by proxy using family history of disease. *Nat. Genet.* **49**, 325–331 (2017).
40. Gusev, A. et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat. Genet.* **50**, 538–548 (2018).
41. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
42. Li, T. et al. GeNets: a unified web platform for network-based genomic analyses. *Nat. Methods* **15**, 543–546 (2018).
43. Raj, T. et al. Alzheimer disease susceptibility loci: evidence for a protein network under natural selection. *Am. J. Hum. Genet.* **90**, 720–726 (2012).
44. Nixon, R. A. New perspectives on lysosomes in ageing and neurodegenerative disease. *Ageing Res. Rev.* **32**, 1 (2016).
45. Emmett, M. J. et al. Histone deacetylase 3 prepares brown adipose tissue for acute thermogenic challenge. *Nature* **546**, 544–548 (2017).
46. Tian, Y., Chang, J. C., Fan, E. Y., Flajolet, M. & Greengard, P. Adaptor complex AP2/PICALM, through interaction with LC3, targets Alzheimer's APP-CTF for terminal degradation via autophagy. *Proc. Natl Acad. Sci. USA* **110**, 17071–17076 (2013).
47. Ingelsson, M. et al. Early Aβ accumulation and progressive synaptic loss, gliosis, and tangle formation in AD brain. *Neurology* **62**, 925–931 (2004).
48. Guillozet, A. L., Weintraub, S., Mash, D. C. & Mesulam, M. M. Neurofibrillary tangles, amyloid, and memory in aging and mild cognitive impairment. *Arch. Neurol.* **60**, 729–736 (2003).
49. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
50. Fairfax, B. P. et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).

Acknowledgements

We thank the participants of ROS and MAP for their essential contributions and gift to these projects; A. Gusev for helpful discussions and for sharing the source code and scripts for TWAS; the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses. This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai. T.R. is supported by grants from the NIH National Institute on Aging (R01AG054005) and the Alzheimer's Association. P.L.D. is supported by NIH R01AG036836. D.A.B. is supported by NIH P30AG10161, R01AG015819, R01AG017917. P.L.D. and D.A.B. are supported by NIH U01AG046152. B.Z. is supported by NIH R01AG046170, RF1AG054014, RF1AG057440 and R01AG057907. We thank the patients and families who donated material for CommonMind Consortium data. The CommonMind Consortium data are available in CMC Knowledge Portal: https://www.synapse.org/#!Synapse:syn4923029. Data were generated as part of the CMC supported by funding from Takeda Pharmaceuticals Company Limited, F. Hoffman-La Roche Ltd and NIH grants R01MH085542, R01MH093725, P50MH066392, P50MH080405, R01MH097276, RO1-MH-075916, P50M096891, P50MH084053S1, R37MH057881 and R37MH057881S1, HHSN271201300031C, AG02219, AG05138 and MH06692. Brain tissue for the study was obtained from the following brain bank collections: the Mount Sinai NIH Brain and Tissue Repository, the University of Pennsylvania Alzheimer's Disease Core Center, the University of Pittsburgh NeuroBioBank and Brain and Tissue Repositories and the NIMH Human Brain Collection Core. CMC Leadership: P.S., J. Buxbaum (Icahn School of Medicine at Mount Sinai), B. Devlin, D. Lewis (University of Pittsburgh), R. Gur, C.-G. Hahn (University of Pennsylvania), K. Hirai, H. Toyoshiba (Takeda Pharmaceuticals Company Limited), E. Domenici, L. Essioux (F. Hoffman-La Roche Ltd), L. Mangravite, M. Peters (Sage Bionetworks), T. Lehner and B. Lipska (NIMH).

Author contributions

T.R. and P.L.D. conceived the project and planned the experiments. T.R. and Y.I.L. analyzed and interpreted the data with support from G.W., S.R., J.H., Y.W., I.G., B.N. and S.M. P.L.D., D.A.B., M.W., P.S., E.E.S., V.H. and B.Z. contributed samples and/or data. T.Y.P. performed the tau overexpression in iPSC neurons. T.R. and P.L.D. prepared the first draft of the manuscript. All authors contributed to the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0238-1.

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

Correspondence and requests for materials should be addressed to T.R. or P.L.D.

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

Methods

Study cohorts. *ROS.* From January 1994 to June 2010, 1,148 individuals agreed to annual detailed clinical evaluation and brain donation at the time of death. Of these, 1,139 have completed their baseline clinical evaluation: 68.9% were women; 88.0% were white, non-Hispanic; their mean age was 75.6 years; and mean education was 18.1 years. There were 287 cases of incident dementia and 273 cases of incident Alzheimer's disease with or without a coexisting condition. Details of the clinical and pathologic methods have been previously reported[17].

MAP. From October 1997 to June 2010, 1,403 individuals agreed to annual detailed clinical evaluation and donation of the brain, spinal cord, nerve and muscle at the time of death. Of these, 1,372 completed their baseline clinical evaluation: 72.7% were women; 86.9% were white, non-Hispanic; their mean age was 80.0 years; and mean education was 14.3 years with 34.0% with 12 or fewer years of education. There were 250 cases of incident dementia and 238 cases of incident Alzheimer's disease with or without a coexisting condition. Details of the clinical and pathologic methods have been previously reported[51]. To avoid population stratification artifacts in the genetic analyses, the study was limited to non-Hispanic whites.

See Supplementary Note for the details of CMC and MSBB datasets.

Data acquisition, quality control and normalization. *Genotyping.* DNA from ROS and MAP subjects was extracted from whole blood, lymphocytes or frozen post-mortem brain tissue and genotyped on the Affymetrix GeneChip 6.0 platform at the Broad Institute's Center for Genotyping. Only self-declared non-Hispanic Caucasians were genotyped to minimize population heterogeneity. PLINK software[52] was used to implement our quality-control pipeline. We applied standard quality-control measures for subjects (genotype success rate >95%, genotype-derived gender concordant with reported gender, excess inter/intraheterozygosity) and for SNPs (Hardy-Weinberg equilibrium $P > 0.001$; minor allele frequency> 0.01, genotype call rate >0.95; misshap test $>1 \times 10^{-9}$) to these data. Subsequently, EIGENSTRAT[53] was used to identify and remove population outliers using default parameters. Imputation was performed using Michigan Imputation Server with Minimac3[54] using Haplotype Reference Consortium (HRC version r1.1, 2016)[55] panel, which consists of 64,940 haplotypes of predominantly European ancestry. Imputation filtering of $r^2 > 0.3$ was used for quality control. After quality control, 450 individuals and 8,383,662 genotyped or imputed markers were used for sQTL analysis.

RNA-seq data. RNA was sequenced from the gray matter of the DLPFC of 542 samples, corresponding to 540 unique brains. These samples were extracted using Qiagen's miRNeasy mini kit and the RNase-free DNase Set. RNA was quantified using Nanodrop. The quality of RNA was evaluated by the Agilent Bioanalyzer. All samples were chosen to pass two initial quality filters: RNA integrity (RIN) score > 5 and quantity threshold of 5 µg (and were selected from a larger set of 724 samples). RNA-seq library preparation was performed using the strand-specific dUTP method with poly-A selection. Sequencing was performed on the Illumina HiSeq with 101-bp paired-end reads and achieved coverage of 150 million reads of the first 12 samples. These 12 samples served as a deep-coverage reference and included two males and two females of non-impaired, mild cognitive impaired and Alzheimer's cases. The remaining samples were sequenced with target coverage of 50 million reads; the mean coverage for the samples to pass quality control is 95 million reads (median 90 million reads). The libraries were constructed and pooled according to the RIN scores such that similar RIN scores would be pooled together. Varying RIN scores result in a larger spread of insert sizes during library construction and leads to an uneven coverage distribution throughout the pool.

The RNA-seq data were processed by a parallelized pipeline. This pipeline includes trimming the beginning and ending bases from each read, identifying and trimming adapter sequences from reads, detecting and removing rRNA reads, and aligning reads to the reference genome. Specifically, RNA-seq reads in FASTQ format were inspected using FASTQC program. Barcode and adapter contamination, low-quality regions (8 bp at the start and 7 bp at the end of each FASTQ reads) were trimmed using FASTX-toolkit. To remove rRNA contamination, we aligned trimmed reads to the rRNA reference (rRNA genes were downloaded from the UCSC genome browser selecting the RepeatMask table) by BWA, and we then extracted only paired unmapped reads for transcriptome alignment. STAR (v.2.5)[56] was used to align reads to the transcriptome reference, and RSEM (v.1.3.0)[57] was used to estimate expression levels for all transcripts. To quantify the contribution of experimental and other confounding factors to the overall expression profiles, we used the COMBAT algorithm[58] to account for the effect of batch and linear regression to remove the effects of RIN, postmortem interval, sequencing depth, study index (ROS sample or MAP sample), genotyping principal components, age at death and sex. Finally, only highly expressed genes were kept (mean expression > 2 $\log_2$ fragments per kilobase million), resulting in 13,484 expressed genes for eQTL analysis. The details for cis-eQTL analysis have previously been described[29].

Intron usage mapping and quantification. We used LeafCutter[20] to obtain clusters of variably spliced introns. LeafCutter enables the identification of splicing events without relying on existing annotations, which are typically incomplete, especially in the setting of large genes or individual- and/or population-specific isoforms. LeafCutter defines 'clusters' of introns that represent alternative splicing choices. To do this, it first groups together introns that overlap (defined by spliced reads). For each of these groups, LeafCutter constructs a graph in which nodes are introns and edges represent overlapping introns. The connected components of this graph define the intron clusters. Singleton nodes (introns) are discarded. For each intron cluster, it iteratively (1) removed introns that were supported by fewer than 100 reads or fewer than 5% of the total number of intronic read counts for the entire cluster, and (2) reclustered introns according to the procedure above. The intron usage ratio for each cluster was next computed and standardized (across individuals) and quantile normalized (across sample) as previously described[20]. LeafCutter was carefully benchmarked against other methods[20], and was able to, for example, identify as many or more differentially spliced events when compared to other methods.

Association of intron usage with Alzheimer's disease and neuropathology traits. The association analysis with neuropathology traits and intron usage was performed using a linear model, adjusting for experimental batch, RIN, sex, age at death and postmortem interval. To test for association with Alzheimer's disease, we limited the comparison to those participants with clinical and pathological diagnosis of Alzheimer's disease and those who have neither diagnosis (Supplementary Table 1). We used LeafCutter[20] to identify intron clusters with at least one differentially excised intron by jointly modeling intron clusters using a Dirichlet-multinomial generalized linear model[20]. To account for neuronal loss and cell type proportion in each brain sample, we used gene expression level of cell type-specific genes as an additional covariate. However, these measures did not affect our association analysis. We report differentially spliced introns at Bonferroni-corrected $P < 0.05$ to correct for multiple hypothesis testing.

We used variancePartition[59] to estimate the proportion of variance explained by differently excised introns association with Alzheimer's disease, burden of amyloid, burden of tangles and neuritic plaques.

sQTL mapping. We used LeafCutter[20] to obtain the proportion of intron-defining reads to the total number of reads from the intron cluster it belongs to. This intron ratio describes how often an intron is used relative to other introns in the same cluster. We used WASP[60] to remove read-mapping biases caused by allele-specific reads. This is particularly important when a variant is covered by reads that also span intron junctions as it can lead to a spurious association between the variant and intron excision level estimates. We standardized the intron ratio values across individuals for each intron and quantile normalized across introns[61] and used this as our phenotype matrix. We used linear regression (as implemented in fastQTL)[26] to test for associations between SNP dosages (MAF ≥ 0.01) within 100 kb of intron clusters and the rows of our phenotype matrix that correspond to the intron ratio within each cluster. As covariate, we used the first three principal components of the genotype matrix to account for the effect of ancestry plus the first 15 principal components of the phenotype matrix (PSI) to regress out the effect of known and hidden factors. The principal components regress out the technical and biological covariates such as experimental batch, RIN, sex, age at death, and postmortem interval. To estimate the number of sQTLs at any given FDR, we ran an adaptive permutation scheme[26], which maintains a reasonable computational load by tailoring the number of permutations to the significance of the association. We computed the empirical gene-level P value for the most significant QTL for each gene. Finally, we applied Benjamini–Hochberg corrections to the permutation P values to extract all significant sQTL pairs with an FDR < 0.05.

TWAS. We used RNA-seq data and genotypes from ROS/MAP to impute the cis genetic component of expression/intron usage[37,40] into a large-scale late-onset Alzheimer's disease GWAS of 74,046 individuals from the IGAP[14]. The complete TWAS pipeline is implemented in the FUSION suite of tools[40]. The details steps implemented in FUSION are as follows. First, we estimated the heritability of gene expression or intron usage unit and stopped if not significant. We estimated using a robust version of GCTA-GREML[62], which generates heritability estimates per feature as well as the likelihood ratio test P value. Only features that have a heritability of Bonferroni-corrected $P < 0.05$ were retained for TWAS analysis. Second, the expression or intron usage weights were computed by modeling all cis-SNPs (±1 Mb from the transcription start site) using best linear unbiased prediction, or modeling SNPs and effect sizes with Bayesian sparse linear mixed model, least absolute shrinkage and selection operator, Elastic Net and top SNPs[37,40]. A cross-validation for each of the desired models were then performed. Third, a final estimate of weights for each of the desired models was performed and the results were stored. The imputed unit is treated as a linear model of genotypes with weights based on the correlation between SNPs and expression in the training data while accounting for linkage disequilibrium (LD) among SNPs. To account for multiple hypotheses, we applied an FDR < 0.05 within each expression and splicing reference panel that was used.

We used the same TWAS pipeline to process the CMC datasets (see Supplementary Note).

Joint and conditional analysis. Joint and conditional analysis of TWAS results was performed using the summary statistic-based method described previously[41], which we applied to genes instead of SNPs. We used TWAS statistics from the main results and a correlation matrix to evaluate the joint or conditional model. The correlation matrix was estimated by predicting the cis-genetic component of expression for each TWAS gene and computing Pearson correlations across all pairs of genes. We used FUSION tool to perform the joint and conditional analysis and to generate conditional outputs and plots.

Gene expression, DNA methylation, Histone modification QTL mapping. The details of ROS/MAP gene expression, DNA methylation and histone modification data are described in the Supplementary Note. The quantitative trait locus (xQTL) analysis on a multi-omic dataset has been described previously[29]. The xQTL results and analysis scripts can be accessed through online portal, xQTL Serve (see URLs).

QTL sharing. We used the Storey's π_1 statistics[63] also described previously[64], QTL sharing was estimated as the proportion of true associations π_1 among the top SNP in each QTL in the second QTL.

Enrichment of sQTLs within epigenomic marks and splicing factor binding sites. We selected a set of 71 human curated RBP splicing regulatory proteins from the SpliceAid-F database[65] to analyze the relationship between gene expression levels of RBP and intron usage patterns across all samples. To test for enrichment of sQTLs in RBP binding sites, we downloaded human CLIP data in BED format from CLIPdb[34]. We used GREGOR[66] (Genomic Regulatory Elements and Gwas Overlap algoRithm) to evaluate global enrichment of trait-associated variants in splicing factor binding sites. GREGOR[66] evaluates the significance of the observed overlap (of sQTL and splicing factor binding sites) by estimating the probability of the observed overlap of the lead sQTL relative to expectation using a set of matched control variants (random control SNPs are selected across the genome that match the index SNP for a number of variants in LD, minor allele frequency and distance to nearest intron). We used Fisher's exact test in combination with Benjamini–Hochberg FDR correction for multiple testing.

Enrichment of sQTLs in chromatin states. We downloaded chromatin states from the Roadmap Epigenomics Project. The 15 chromatin states were generated from five chromatin marks in DLPFC of a cognitively non-impaired MAP subject with minimal pathology as part of the Roadmap Epigenomics Consortium[27]. A ChromHMM model applicable to brain epigenome was learned by virtually concatenating consolidated data corresponding to the core set of five chromatin marks assayed (H3K4me3, H3K4me1, H3K36me3, H3K27me3 and H3K9me3). BED files were downloaded from Washington University in St. Louis Roadmap Epigenome Browser. To test for enrichment for sQTLs among the 15 chromatin states, we used GREGOR[66] to evaluate global enrichment of trait-associated variants in splicing factor binding sites.

GWAS enrichment analyses. We used GARFIELD (see URLs) to test for enrichment of IGAP Alzheimer's disease GWAS SNPs among sQTLs and other publicly available QTL datasets. GARFIELD performs greedy pruning of GWAS SNPs (LD $r^2 > 0.1$) and then annotates them based on functional information overlap. It quantifies fold enrichment at GWAS $P < 10^{-5}$ significant cut-off and assesses them by permutation testing, while matching for minor allele frequency, distance to nearest transcription start site and a number of LD proxies ($r^2 > 0.8$).

Q–*Q* plots show quantiles of one dataset against quantiles of a second dataset and are commonly used in GWAS to show a departure from an expected *P*-value distribution. We generated *Q*–*Q* plots for LD-pruned GWAS SNPs (PLINK with the settings '– indep- pairwise 100 5 0.8'). We compared the sQTLs overlapping with LD-pruned GWAS SNPs and compared the distribution to a random set of SNPs with similar MAF.

GWAS datasets. We performed the TWAS using GWAS summary statistics from: (1) Alzheimer's disease GWAS from the IGAP (stage 1 data)[14]; (2) Alzheimer's disease genome-wide association study by proxy (GWAx) in 116,196 individuals from the UKBB[39].

PPI network and pathway analysis. We constructed a PPI network using GeNets[42] to determine whether the Alzheimer's disease TWAS genes significantly interact with each other and with known Alzheimer's disease-associated proteins. GeNets create networks of connected proteins using evidence of physical interaction from the InWeb database, which contains 420,000 high-confidence pairwise interactions involving 12,793 proteins[67]. Community structures of the underlying genes are displayed in GeNets. These 'communities' are also called modules or clusters. This feature highlights genes that are more connected to one another than they are to other genes in other modules. To assess the statistical significance of PPI networks, GeNets applies a within-degree node-label permutation strategy to build random networks that mimic the structure of the original network and evaluates network connectivity parameters on these random networks to generate empirical distributions for comparison to the original network. In addition to PPI network analysis, GeNets allows for gene set enrichment analysis on genes within the PPI network. We used Molecular Signatures Database Curated Gene Sets, curated from various sources such as online pathway databases, the biomedical literature and knowledge of domain experts and Canonical Pathways, curated from pathway databases such the Kyoto Encyclopedia of Genes and Genomes, BioCarta, and Reactome to test for gene set enrichment within the PPI network. Then a hypergeometric testing is applied to get *P* value for gene set enrichment. We used Bonferroni-corrected $P < 0.05$ to correct for multiple hypothesis testing.

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

Data availability

The ROS/MAP sQTL visualization (Shiny App) browser is available at https://rajlab.shinyapps.io/sQTLviz_ROSMAP/. The ROS/MAP data are available at the RADC Research Resource Sharing Hub at http://www.radc.rush.edu/. The ROS/MAP and MSBB mapped RNA-seq data that support the findings of this study are available from the AMP-AD Knowledge Portal (https://www.synapse.org/#!Synapse:syn2580853) upon authentication by the Consortium. The CMC data are available from the CMC Knowledge Portal (https://www.synapse.org/#!Synapse:syn4923029).

References

51. Bennett, D. A. et al. Overview and findings from the Rush Memory and Aging Project. *Curr. Alzheimer. Res.* **9**, 646–663 (2012).
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. Patterson, N., Price, A. L. & Reich, D. Population structure and Eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
54. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
55. The Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
56. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
57. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
58. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
59. Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, 483 (2016).
60. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).
61. Degner, J. F. et al. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
62. 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).
63. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
64. Nica, A. C. et al. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.* **7**, e1002003 (2011).
65. Giulietti, M. et al. SpliceAid-F: a database of human splicing factors and their RNA-binding sites. *Nucleic Acids Res.* **41**, D125–D131 (2013).
66. Schmidt, E. M. et al. GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics* **31**, 2601–2606 (2015).
67. Lage, K. et al. A human phenome–interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* **25**, 309–316 (2007).

nature research

Corresponding author(s): Towfique Raj and Philip De Jager

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

No sample-size calculation was performed before the study. The number of samples (n=450) was determined by the availability of high-quality brain specimens of ROS/MAP study participants. The goal of this study is to identify the impact of common (minor allele frequency > 5%) genetic variants on local splicing events. Assuming an additive model of effects on gene expression levels, we estimate that the sample size of 450 controls provides excellent power to detect modest genetic effects at genome-wide level of significance 5 x 10-8 (91% power to detect effects explaining 10% of expression variability and 23% power to detect effects explaining 5% of variability). Sample size of hundreds (usually 200-400) is typical for a QTL studies. Therefore, we believe that the sample size of 450 is sufficient to detect significant associations. Finally, we also note that this is one of the largest brain transcriptome data available.

2. Data exclusions

Describe any data exclusions.

In total, 92 out of 542 total samples were excluded due to insufficient RNA-seq quality or they did not have available genotyping data. All quality measures are given in Online Methods and Supplementary Notes. The QC process was independent of the samples' clinical or pathological characteristics.

3. Replication

Describe whether the experimental findings were reliably reproduced.

We have made every effort to replicate/validate our findings in two independent RNA-sequencing datasets from the Commonmind Consortium (CMC) (Fromer et al. 2016 Nature Neuroscience) and from Mount Sinai Brain Bank (MSBB) (https://www.synapse.org/#!Synapse:syn2580853). This replication effort is rather unusual for a genetical genomics study. The following validation/replication analyses were carried out:

-MSBB RNA-seq dataset (n=301) was used to replicate aberrant splicing alterations in AD brains. We used MSBB to perform independent replication of AD-specific alternative splicing events because the RNA-seq dataset is from an aging Alzheimer's disease cohort.
-To further validate and explore the mechanism of aberrant splicing alterations in AD brains, we analyzed RNA-Seq data derived from control iPSC- derived neurons (iN) and iN overexpressing Tau.
-CommonMind Consortium RNA-seq data (n=537) from the dorsolateral prefrontal cortex (DLFPC) was used to replicate the splicing QTLs discovered in ROS/MAP.
-To replicate the transcriptome-wide association study results (built using ROS/MAP data as a reference panel), we assessed whether using the expression imputation model built using the CMC dataset that was deployed in AD GWAS yields significant results.
-The UK BioBank (UKBB) AD GWAS by proxy data was used to replicate the International Genomics of Alzheimer's Project (IGAP) AD TWAS results.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

No allocation into groups was performed. Participants of the ROS and MAP studies were not cognitively impaired at enrollment. After death, pathological assessment was performed to measure tau, tangles, and amyloid-beta burden. Subsequently, brain samples were sent to RNA-sequencing in an arbitrary order.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Individuals generating the brain-derived RNA- sequence data were blinded to the outcome measures.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

Calling local splicing events or intronic excision: LeafCutter, http://davidaknowles.github.io/leafcutter/index.html
RNA-seq aligner: STAR v2.5, https://github.com/alexdobin/STAR
Quantifying gene expression: RSEM v1.3.0, https://github.com/deweylab/RSEM
Mapping QTL: fastQTL, http://fastqtl.sourceforge.net/
Allele-specific read mapping: WASP, https://github.com/bmvdgeijn/WASP
Removing batch effects from RNA-seq data: COMBAT, https://www.bu.edu/jlab/wp-assets/ComBat/Abstract.html
Genotype, genetic association and QC: PLINK, https://www.cog-genomics.org/plink2
Population stratification: EIGENSTRAT, https://github.com/DReichLab/EIG/tree/master/EIGENSTRAT
Imputing GWAS data: Michigan Imputation Server, https://imputationserver.sph.umich.edu/index.html
Transcriptome-wide associaiton study,Performing the expression imputation using several learning models, joint/conditional analysis: FUSION, http://gusevlab.org/projects/fusion/
Heritability estimates, GCTA- GREML, http://cnsgenomics.com/software/gcta/#Overview
Quantitates the expression level of alternatively spliced genes: MISO, http://genes.mit.edu/burgelab/miso/;
Evaluate enrichment of trait-associated variants in annotated features: GREGOR, http://genome.sph.umich.edu/wiki/GREGOR;
GWAS analysis of regulatory or functional information enrichment with LD correction: GARFIELD, http://www.ebi.ac.uk/birney-srv/GARFIELD;
Figures were generated using FUSION tools, R and ggplot2 2.2.1.

All software is freely available.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No unique materials were used. Additional human samples are available from the individuals in our study through the RUSH University Alzheimer Disease Center.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

N/A.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

N/A.

b. Describe the method of cell line authentication used.

N/A.

c. Report whether the cell lines were tested for mycoplasma contamination.

N/A.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

N/A.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

N/A.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

The mean age of death of the 450 subjects with RNA-seq profiles was 88.54 years. 281 of the 450 subjects were female, and 268 subjects were diagnosed with AD at pathological assessment. All analyses were adjusted for age and gender and other technical covariates. The details of covariates are in Supplementary Table 1. The ROS and MAP studies are described in detail at the RADC Research Resource Sharing Hub: http://www.radc.rush.edu/