scientific reports

OPEN

Deep post-GWAS analysis identifies potential risk genes and risk variants for Alzheimer's disease, providing new insights into its disease mechanisms

Zhen Wang[1,2], Quanwei Zhang[2], Jhih-Rong Lin[2], M. Reza Jabalameli[2], Joydeep Mitra[2], Nha Nguyen[2] & Zhengdong D. Zhang[2]✉

Alzheimer's disease (AD) is a genetically complex, multifactorial neurodegenerative disease. It affects more than 45 million people worldwide and currently remains untreatable. Although genome-wide association studies (GWAS) have identified many AD-associated common variants, only about 25 genes are currently known to affect the risk of developing AD, despite its highly polygenic nature. Moreover, the risk variants underlying GWAS AD-association signals remain unknown. Here, we describe a deep post-GWAS analysis of AD-associated variants, using an integrated computational framework for predicting both disease genes and their risk variants. We identified 342 putative AD risk genes in 203 risk regions spanning 502 AD-associated common variants. 246 AD risk genes have not been identified as AD risk genes by previous GWAS collected in GWAS catalogs, and 115 of 342 AD risk genes are outside the risk regions, likely under the regulation of transcriptional regulatory elements contained therein. Even more significantly, for 109 AD risk genes, we predicted 150 risk variants, of both coding and regulatory (in promoters or enhancers) types, and 85 (57%) of them are supported by functional annotation. In-depth functional analyses showed that AD risk genes were overrepresented in AD-related pathways or GO terms—e.g., the complement and coagulation cascade and phosphorylation and activation of immune response—and their expression was relatively enriched in microglia, endothelia, and pericytes of the human brain. We found nine AD risk genes—e.g., *IL1RAP*, *PMAIP1*, *LAMTOR4*—as predictors for the prognosis of AD survival and genes such as *ARL6IP5* with altered network connectivity between AD patients and normal individuals involved in AD progression. Our findings open new strategies for developing therapeutics targeting AD risk genes or risk variants to influence AD pathogenesis.

Alzheimer's disease (AD) is a progressive, chronic neurodegenerative disorder with a long prodromal phase[1]. With a complex genetic etiology and a high heritability, estimated ranging from 60 to 80%[2], AD is usually divided into two subgroups of diseases: the familial early-onset AD, caused by mutations in single genes including *APP*, *PSEN1*, and *PSEN2*, and the late-onset AD (LOAD), influenced by multiple common variants with low effect sizes[3]. Over the past decade, GWAS revealed a large number of AD-associated genetic loci (Supplementary Fig. S1 and Supplementary Table S1), including *SORL1*, *ABCA7*, *CLU*, *CR1*, *INPP5D*, *CD33*, *BIN1*, *PICALM*, *PTK2B*, and *APOE*, a locus that has been repeatedly validated across different studies[4]. Two recent meta-analyses of large cohorts of LOAD ($n = 455{,}258$ and 94,437) identified 29[5] and 25[6] risk loci, respectively. Interpretation of these results, however, remains elusive, because GWAS only detect statistical associations among a subset of all variants and ~86% of AD associated SNPs are non-coding (either intronic or intergenic, Supplementary Fig. S1). Studies have shown improvement on identifying potential risk AD genes by integrating GWAS and omics data. Most of those integration include rather limited or specific information (e.g. QTL or methylation data) alone to identify potential risk AD genes[7–9]. To better understand the biological mechanisms underlying AD etiology,

[1]College of Animal Sciences, Zhejiang University, Hangzhou, Zhejiang, China. [2]Department of Genetics, Albert Einstein College of Medicine, Bronx, NY, USA. ✉email: zhengdong.zhang@einsteinmed.org

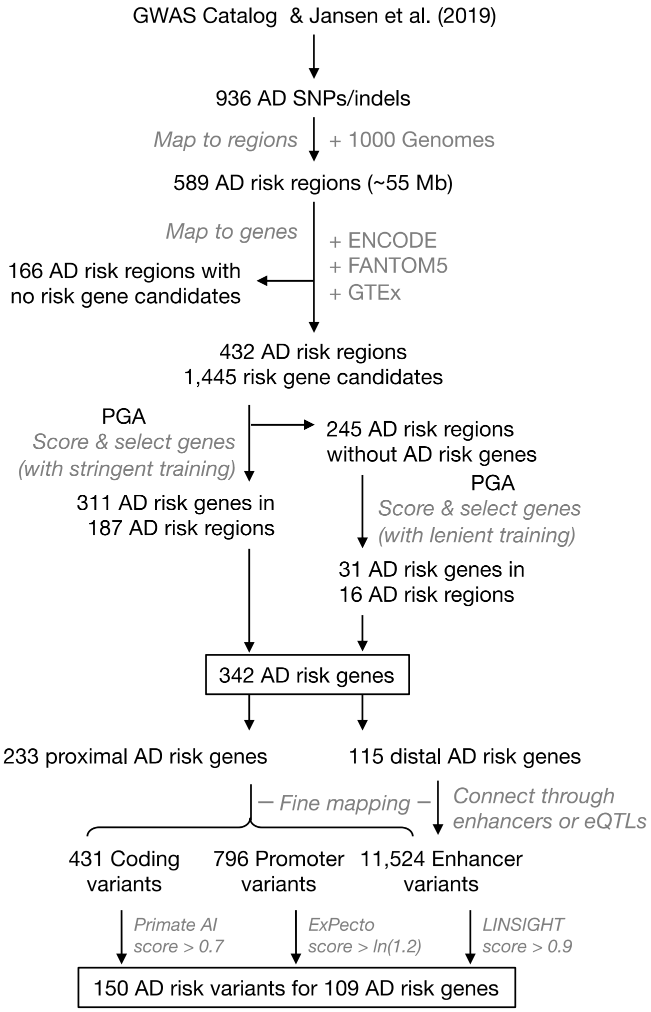

GWAS Catalog & Jansen et al. (2019)
936 AD SNPs/indels
Map to regions
+ 1000 Genomes
589 AD risk regions (~55 Mb)
Map to genes
+ ENCODE
+ FANTOM5
+ GTEx
166 AD risk regions with no risk gene candidates
432 AD risk regions
1,445 risk gene candidates
PGA
Score & select genes
(with stringent training)
245 AD risk regions without AD risk genes
311 AD risk genes in 187 AD risk regions
PGA
Score & select genes
(with lenient training)
31 AD risk genes in 16 AD risk regions
342 AD risk genes
233 proximal AD risk genes
115 distal AD risk genes
Fine mapping
Connect through enhancers or eQTLs
431 Coding variants
796 Promoter variants
11,524 Enhancer variants
Primate AI score > 0.7
ExPecto score > ln(1.2)
LINSIGHT score > 0.9
150 AD risk variants for 109 AD risk genes

Figure 1. Flowchart of AD risk genes and risk variants prediction. Briefly, using 936 AD-associated signals and the 1000 Genome Project data, we first systematically identified genomic AD-risk regions, mapped them to genes (enhancer from ENCODE and FANTOM5 or eQTLs from GTEx were used to identity distal risk gene) to identify AD risk gene candidates, integrated gene network, annotation data and training gene sets to score all gene candidates for AD risk, finally selected genes with a score overpassed the threshold as AD risk genes. Next, all the variants located in the AD risk regions were grouped into coding, promoter and enhancer variants, and corresponding functional annotation of variants (PrimateAI, ExPecto and LINSIGHT) were used to prioritize these variants. The variants for identified AD risk genes with functional annotation score overpassed the threshold will be predicated as potential risk variants (see details in method).

the functional impact of genetic association signals needs to be extensively investigated to identify disease genes and risk variants underlying the genetic signals detected by GWAS.

To this end, we sought to integrate genomic data from multiple sources—e.g., GWAS signals from the GWAS Catalog, disease genes databases (MalaCards[10], DISEASES[11], and DisGeNET v5.0[12]), functional annotation of genetic variants (LINSIGHT[13], ExPecto[14], and PrimateAI[15]), and the 1000 Genomes Project—to predict AD risk genes and risk variants. In this study, we aimed to first compile a list of high-confidence AD risk genes derived from association signals, then systematically uncover the characteristics of the identified AD risk genes, including the level and variation of their expression in different types of cells, and finally use a computational framework that we developed to identify putative risk variants connected to AD risk genes. Our results provide novel biological insights into the genetic architecture, expression profiles, functional pathways involved in the AD etiology, and ultimately a basis for future therapeutic development for the disease.

Results

AD risk regions and risk genes.

Using 936 GWAS AD SNPs and linkage disequilibrium, we identified 589 genomic risk regions, spanning ~ 55.0 Mb of the human genome (Fig. 1). Based on the genomic annotation of genes and regulatory information including enhancer and eQTL, we could connect 1,445 genes to 432 risk regions (Supplementary Fig. S2 and Supplementary Table S2). Among AD risk gene candidates, 967 are proximal genes, overlapping AD risk regions, and 506 are distal genes, linked to AD risk regions through long-range

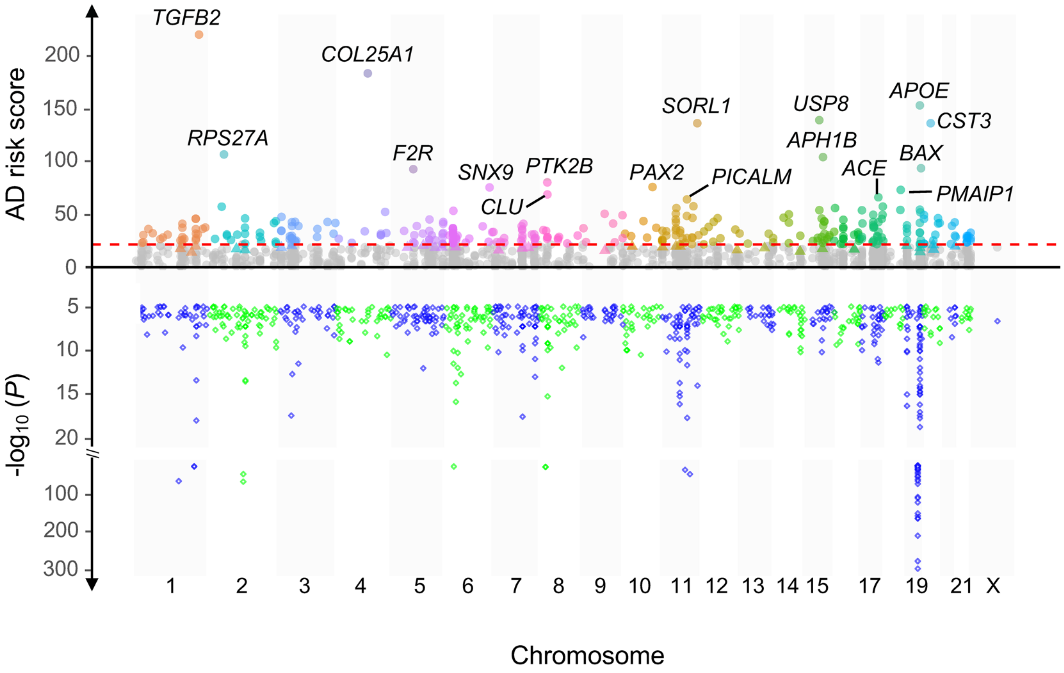


Figure 2. AD risk genes and AD-associated variants across chromosomes. The upper part of the figure shows the scores of AD risk gene candidates, which were calculated based on the stringent training gene set. Colored and gray dots represent AD risk gene candidates above and below the threshold (=21.4, the red dashed line), respectively. The colored small triangles below the threshold represent additional AD risk genes predicted with the lenient gene training set. Top AD risk genes are labelled with their gene symbols. The upper part of the figure is the Manhattan plot of the 936 AD-associated SNPs that we collected from the GWAS Catalog and Jansen et al.[5].

gene regulatory elements (e.g., enhancers) (Supplementary Table S2). 28 genes are both proximal for some risk regions and distal for other risk regions. 35 candidates were not scored due to their lack of GO annotation and/or exclusion from the gene functional linkage network. Among 1,410 scored candidates, 342 loci, distributed in 203 risk regions (Fig. 2 and Supplementary Table S3), surpassed the "high" threshold (see Method) and thus were considered as (putative) AD risk genes. They included 233 (68.1%) candidates proximal to AD risk regions, and additional 115 (33.6%) distal genes, which are likely to be regulated by regulatory elements in the risk regions (Fig. 1 and Supplementary Table S3). Comparing AD risk genes that we identified with ones reported by past GWAS, we found that 246 genes on our list are novel—they have not been identified as AD risk genes by AD GWAS so far, and many of on our list such as *CR1*[16], *ABCA7*[17], *TREM2*[18], *SORL1*[19] and *BIN1*[20] have been reported to play import roles in the pathologies of AD.

KEGG pathways and GO biological processes enriched with AD risk genes. Functional enrichment analysis of 342 AD risk genes using DAVID[21] showed that AD risk genes were overrepresented in 10 KEGG pathways (FDR < 0.05) and with 151 GO terms (adjusted $P < 0.01$ after the Bonferroni correction) (Fig. 3A,B and Supplementary Table S4), many of which are highly relevant to the AD pathology.

Clustered expression of AD risk genes in different human tissues. Based on their expression profiles in different human tissues, AD risk genes can be clustered into three groups (Fig. 3C and Supplementary Fig. S3). The first group of 32 genes was expressed almost exclusively in the central nervous system (CNS), especially the frontal and the prefrontal cortices. Many genes in this group, such as *BIN1*, *MAPT*, and *CNTNAP2*, have been implicated in the pathogenesis of AD[22–24]. The second group included 131 genes actively expressed in the immune cells such as B and T lymphocytes. 115 genes in the third group were expressed across a wide range of different tissues, including the CNS. Many genes in this group, such as *APOE*, *CR1*, and *EPHA1*, are known to be associated with AD. Human studies clearly indicate that ApoE isoforms differentially affect Aβ aggregation and clearance[4], and *CR1* may play a role in the clearance of Aβ[25].

High expression of AD risk genes in microglia, endothelia, and pericytes of human brain. The expression of AD risk genes was significantly enriched in microglia, endothelia and pericytes in the frontal and the visual cortices and cerebellum from human adults (Fig. 3D and Supplementary Table S5). The high-expression profile was evident for many genes (Supplementary Fig. S4), which were enriched for microglial markers (e.g., *HLA-DRA* and *TREM2*), endothelial markers (e.g., *CD34*), and pericyte markers (e.g., *NR1H3*).

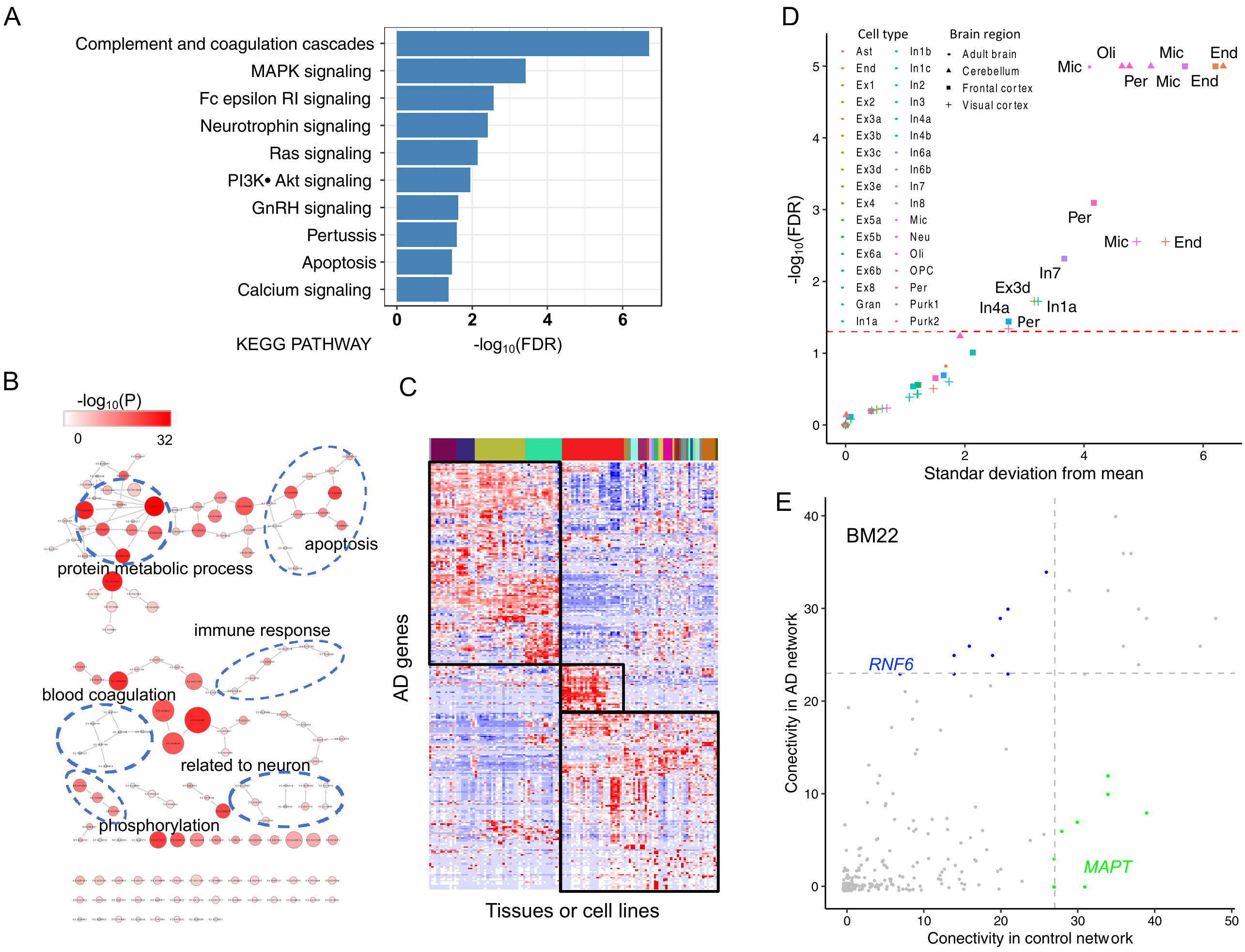


Figure 3. Functional annotation of AD risk genes. (**A**) KEGG pathways and (**B**) GO terms of biological processes enrichment of AD risk genes. In (**B**), each dot represents a significantly enriched GO term, whose $-\log_{10}$(*P*-value) and AD risk gene count are indicated by the color and size of the dot, respectively. (**C**) Expression of AD risk genes in different tissues. Three expression clusters are indicated by the black boxes. See Supplementary Fig. S3 for details. (**D**) Cell type enrichment analysis. We analyzed the expression of AD risk genes in different types of single cells from frontal and visual cortices, cerebellum, and adult brain. The red dashed line represents FDR = 0.05, and cell types whose transcriptomes were significantly enriched with the expression of AD risk genes are shown, with their brain regions indicated. Used cell type abbreviations: Ast, astrocytes; End, endothelial; Ex, excitatory; Gran, granule; In, inhibitory; Mic, microglia; Neu, neuron; Oil, oligodendrocytes; OPC, oligodendrocytes precursor; Per, pericytes; Purk, purkinje. (**E**) Network connectivity of AD risk genes in the co-expression network of both AD patients and normal controls in the brain region BM22. Each dot represents an AD risk gene. The gray dashed line marks the threshold for network hub genes either in the co-expression network of AD patients or normal controls. Blue and green dots represent network hub genes only in the co-expression network of AD patients and only in the co-expression network of normal controls, respectively. See Supplementary Fig. S6 for details and other brain regions.

Connectivity of AD risk genes in co-expression network. We carried out the gene co-expression network analysis across four brain regions to examine the gene regulation patterns among AD risk genes. Overall, we observed that connections among AD risk genes are less (BM10: *P* value = 0.50; BM22: *P* value = 5.8E-3; BM36: *P* value = 0.01; BM44: *P* value = 8.6E-7; all two-sided t-tests) in AD patients compared to normal controls (Supplementary Figs. S5 and S6). For each brain region, we considered top 20 genes with most interactions with other genes as network hubs. We found that hub genes with high connectivity in AD patients—e.g., *ARL6IP5* (BM10), *RNF6* (BM22), *TP53INP* (BM36), and *GGH* (BM44) —tended to have low connectivity in healthy individuals (Fig. 3E and Supplementary Fig. S5). On the other hand, many hub genes in normal people—e.g., *LMTK2* (BM10), *SPPL2A* (BM22), *MAPT* (BM44), and *USP8* (BM36)—usually had low connectivity in AD patients.

We also analyzed co-expression of AD risk genes at the proteomic level and observed similar patterns that the AD risk genes were less connected among AD patients than normal controls in the ACG region (*P* value = 2.2E-16, two-sided t-test), while opposite pattern in the FC region (*P* value = 1.6E-3, two-sided t-test, Supplementary

Fig. S7). For example, *ARL6IP5* and *GGH* were network hubs in AD patients but less connected in controls in the FC region (Supplementary Fig. S8). In the ACG region, we observed network hubs such as *PTK2B*, *SPARC* , and *RAD50* showing large alteration between AD patients and controls (Supplementary Fig. S8). SPARC is a matricellular protein which can facilitate the migration of immune cells (e.g., blood-derived dendritic cells). Although its role in AD-related neuroinflammation is still not clear, a study has shown that there are significant alterations in its expression and it collocates to Aβ protein deposits in AD brain tissues[26].

Expression of AD risk genes in human brain and its connection to disease survival. Using data from three studies of differential gene expression between AD cases and controls in different brain regions[27–29], we found 171 (50%) AD risk genes were differentially expressed in at least one brain region, including 102 up-regulated genes, 64 down-regulated genes, and 5 genes showing both up- and down-regulation in different brain regions (Supplementary Table S6). Differential expression of AD risk genes was either widespread, occurring in multiple brain regions, or limited to a specific brain region. *TGFB2,* the highest ranked risk gene, was up-regulated in frontal cortex (FC), central nervous system (CNS), temporal cortex (TCX), superior temporal gyrus (STG) and parahippocampal gyrus (PHG), while *PTK2B* was down-regulated in brain cerebellum (CBE), TCX, and PHG. *COL25A1*, the second highest ranked risk gene, and *PMAIP1* were separately down- and up-regulated only in the TCX region. Differential expression of some AD risk genes was discordant in different brain regions. For example, *ApoE* and *CST3* in AD patients were up-regulated in TCX region but down-regulated in cerebellum.

Since AD is mainly a late-onset neurodegenerative disorder, we examined how AD risk genes are expressed specifically among adults. Using a binarization procedure[30], we analyzed their spatiotemporal expression patterns using RNA-seq data from BrianSpan. Although no strong pattern was found (Supplementary Fig. S9), the proportion of AD risk genes with dramatically suppressed expression was increased at the age of 40 compared to early ages. The proportion of AD risk genes that tend to be transcriptionally actively was relatively higher at the early ages (of 23 and 30) in comparison to the old ages (of 36 and 40). We next examined the spatiotemporal expression pattern of AD risk genes during the development of the frontal cortex across an extended range of ages (from 18.05 to 78.23). We did not observe any distinct expression pattern across this range of ages.

As a chronic neurodegenerative disease, AD starts slowly and gradually worsens overtime. We hypothesized that genes whose expression correlates with AD progression may mark AD severity and thus can be used to predict AD prognosis. To test this hypothesis, we assessed the impact of AD genes on survival using the Kaplan–Meier analysis. Based on expression levels, nine genes—*NRG3*, *IL1RAP*, *PMAIP1*, *STRADA*, *SGK3*, *LAMTOR4*, *MAPK12*, *PHB*, and *GRB2*—separated AD patients into low- and high-risk groups with different disease survival ($P<0.05$). Their expression also trends differently with age between healthy individuals and AD patients in at least one brain region ($P<0.05$) (Fig. 4 and Supplementary Fig. S10).

Predicted AD risk variants. Using the computational framework that we developed for this project, we predicted 150 unique potential risk variants (51% of them located in risk regions with p-value < 5E-8) for 109 AD risk genes (Supplementary Table S7 and Fig. S11A). To evaluate this prediction, we analyzed their effect on sequence motifs of transcription factors binding sites (TFBS) and compared these to the eQTL data from GTEx. Motif analysis revealed that 69 predicted risk variants (46%, Supplementary Table S7) cause either gain or loss of TFBS motifs, likely affecting TF binding. Among them, 32 (21%, Supplementary Table S7) have also been identified as eQTLs. Together, 85 (57%) of the predicted risk variants can be functionally annotated (Supplementary Fig. S11B). Three modules were developed in our computational framework to predict risk variants in different functional genomic regions:

Coding variants. We predicted 54 risk coding variants. For example, rs7412 and rs4147934 are two missense coding SNPs, each in high LD ($r^2>0.5$) with one of the AD GWAS lead SNPs in its corresponding risk region, were predicted as risk variants (Fig. 5 A and B, Supplementary Table S7). rs7412, in *APOE* with a PrimateAI score = 0.80, is a well-known variant reported to be associated with AD. rs4147934, in *ABCA7* with a PrimateAI score = 0.78, has been proposed as a functional candidate variant accounting for the GWAS signal at *ABCA7* locus in Caucasians[31]. Although AD risk from rs4147934 is probably population-specific since its association signal was not replicated in the African American cohort[32], our analysis provides additional evidence in support of its causal role in AD and thus its impact in non-European ancestry populations merits further investigation.

Non-coding variants in promoters. We also predicted 33 risk promotor variants in brain tissues or cells (Fig. 5C–E), including rs76516995 (ExPecto score = 0.194) for *BIN1* in astrocytes, rs4292 (ExPecto score = 0.390) for *ACE* in neural cells, and rs12691088 (ExPecto score = 0.224) for *APOC1* in astrocytes. rs76516995 and rs4292 have also been identified as eQTLs by GTEx, while rs12691088 has been shown to be associated with AD-related phenotypes in multiple brain regions[33].

Non-coding variants in enhancers. We predicted 64 risk enhancer variants. Two of them, rs2271920 (LINSIGHT score = 0.951) and rs117423666 (LINSIGHT score = 0.966) Fig. 5F and Supplementary Table S7), are risk variants for risk gene *PTK2B*. rs2271920 is an AD GWAS lead SNP itself[34]. As a GTEx eQTL, it changes the expression of *PTK2B*, likely by altering the binding sites motif of BCL6 and ZNF467. For the *SNX1* locus, we also predicted two risk enhancer variants: rs146600064 (LINSIGHT score = 0.914) and rs60226406 (LINSIGHT score = 0.977) (Fig. 5G and Supplementary Table S7). rs146600064 is in total LD ($r^2=1$) with the lead AD SNP rs74615166[3]. In another risk region indexed by the AD lead SNP rs7207400[34], several predicted risk variants—rs2049515, rs4341787, rs549929529, and rs242557—are connected to multiple AD-risk genes—*NMT1,*

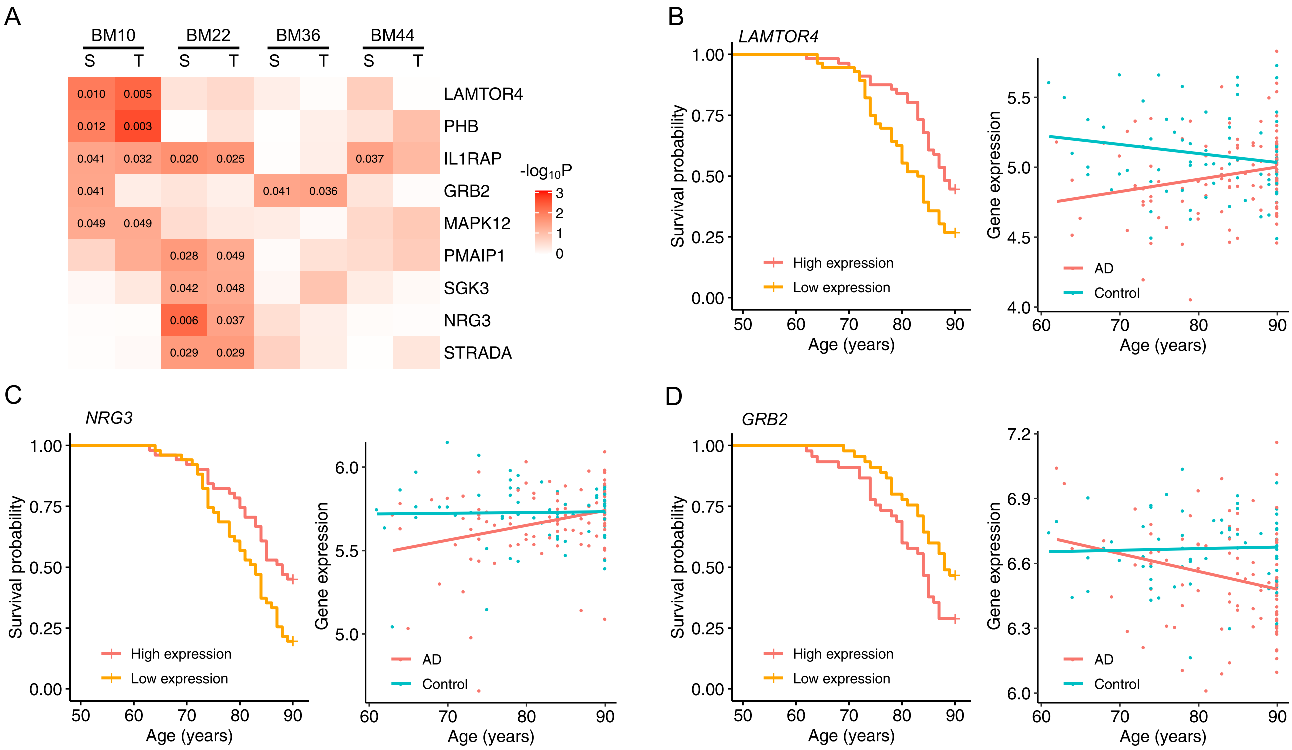


Figure 4. AD survival curves and gene expression trajectories with age. We analyzed the expression of each AD risk gene to see how it may affect the survival of AD patients with low and high expression levels and how it trends with age in AD patients and normal controls. (**A**) The heatmap of *P*-values of Kaplan–Meier survival analyses and gene expression trend tests across four brain regions (see Methods). Nine AD risk genes with at least one *P*-value less than 0.05 (labeled) are included in the heatmap. This figure includes AD survival curves and gene expression trajectories with age for three genes with $P < 0.05$: (**B**) *LAMTOR4* in the brain region BM10, (**C**) *NRG3* in the brain region BM22, and (**D**) *GRB2* in the brain region BM36. See Supplementary Fig. S10 for plots of other significant results.

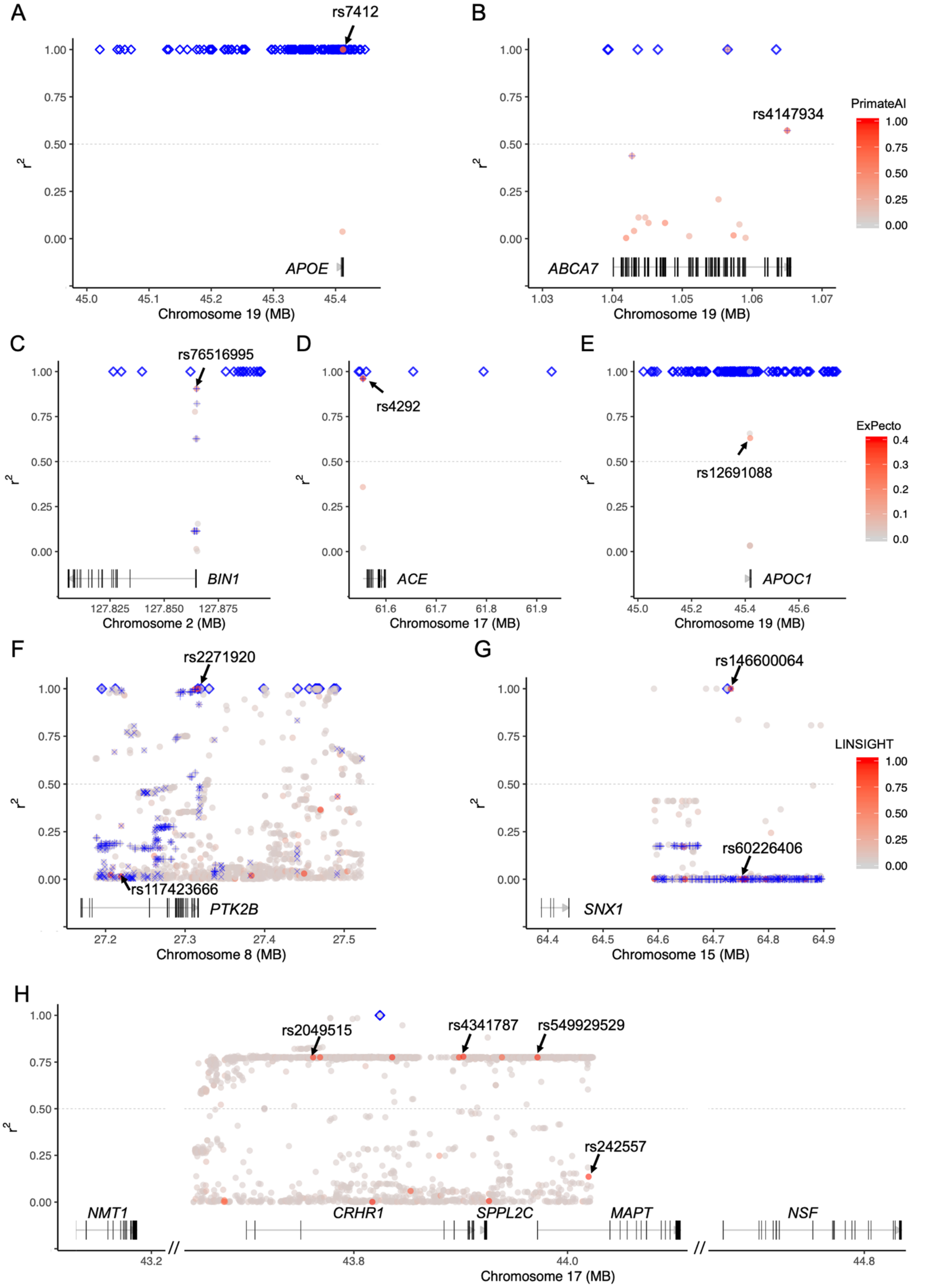
CRHR1, SPPL2C, MAPT, and *NSF* (Fig. 5H and Supplementary Table S7). Although a disease risk gene could be influenced by multiple risk variants, we examined additional data—e.g., TFBS motifs, eQTL, and ClinVar—to further prioritize them. rs549929529 is most likely a risk variant for *MAPT* as it is reported in ClinVar to be associated with 'MAPT-Related_Spectrum_Disorders'. Compared with other SNPs, rs4341787 is in the highest LD ($r^2 = 0.779$) with the AD lead SNP rs7207400 and is an eQTL identified by GTEx for *NMT1*. Also, in high LD ($r^2 = 0.775$) with rs7207400, rs2049515 could be a risk variant for multiple genes as it is an eQTL for *CRHR1*, *SPPL2C*, and *NSF* (Fig. 5H and Supplementary Table S7). Moreover, rs1522388, a predicted risk variant (LINSIGHT score = 0.979) for *FLNB*, was also identified as a reporter assay QTL in HepG2 cell line[35], which experimentally demonstrated its functional impact as a regulatory variant.

To date, the functional impact of the aforementioned variants is still poorly understood in AD etiology, but our findings provide promising risk variant candidates for further experimental validation, which will in turn identify potential drug targets for the development of AD treatment. To explore the molecular function of the predicted risk variants, we examined eQTLs (FDR < 0.05) identified in three brain regions (temporal cortex, dorsolateral prefrontal cortex, and cerebellum) in an AD cohort (Synapse: syn17015233[36]). We found 16 (11%, Supplementary Table S8) risk variants were AD-related eQTLs. Several risk variants—e.g., rs2236393 for *CDH3* and rs12752439 for *HSPG2*—were not identified as eQTLs by GTEx but instead are AD-related eQTLs. This analysis provides direct evidence for their involvement in AD pathogenesis.

Discussions

GWAS have uncovered thousands of genetic variants that influence risk for complex human diseases. However, there is still a large gap between the statistical associations linking locus and trait and the functional impact of risk variants underlying disease risk. Multiple factors have made it difficult to bridge this gap. First, the association of a locus with disease does not reveal the underlying causal variant as many co-inherited variants in strong LD with one another at the locus often have statistically indistinguishable disease associations. Second, genes affected by the causal variants are usually unknown. Consequently, functional studies aimed at determining the causal genetic variants and the biological mechanisms underlying the observed disease association have lagged. Thus, it is important to not only identify disease risk genes but also explore their risk variants. Responding to this need, we proposed a post-GWAS computational framework that could predict the risk variants for a specific disease.

Using this framework for AD, we predicted 342 AD risk genes and, for 109 of them, 150 risk variants. The replication of finding many well-known AD risk genes using our framework shows a reliability of our approach.



Causal variants			AD risk genes (score)				
SNP ID	LINSIGHT score	Motif	*NMT1* (23)	*CRHR1* (39.4)	*SPPL2C* (49.1)	*MAPT* (42.4)	*NSF* (37.1)
rs2049515	0.903492	2 gains	Non-eQTL	eQTL	eQTL	eQTL	eQTL
rs4341787	0.935945	-	eQTL	eQTL	eQTL	eQTL	eQTL
rs549929529	0.940552	-	-	-	-	eQTL	-
rs242557	0.959584	1 gain	eQTL	eQTL	Non-eQTL	eQTL	Non-eQTL

Figure 5. Predicted risk variants. Plotted are examples of coding, promoter, and enhancer risk variants identified using PrimateAI (**A**, **B**), ExPecto (**C**–**E**), and LINSIGHT (**F**–**H**), respectively, with their affected AD risk genes. Each dot represents a variant, whose annotation score is indicated by the color of the dot. Predicted risk variants are marked by arrows with their SNP IDs. Blue diamonds represent AD-associated variants identified by GWAS. Plus and cross signs represent eQTLs and variants located in enhancers. The gray dashed line represents LD $r^2 = 0.5$.

For example, the predicted AD risk genes with the highest risk scores are well-known AD risk genes, such as *ApoE* and *SORL1* have already been shown to functionally underlie the disease pathology[37] (Fig. 2). Moreover, we also provide AD risk genes with highest risk scores as candidates, such as *TGFB2*, *CST3* and *USP8* genes, for researchers to further study their functions in AD. The role of the *TGFB2* gene in AD is still poorly understood, although it was found to autocrinally induce the apoptosis of primary cortical neurons[38] and up-regulated in multiple brain regions among AD patients. TGFβ signaling and *COL25A1* play important roles in the pathogenesis of AD[39,40], while *CST3* may offer neuroprotection against AD[41]. *USP8* gene depletion leads to decreased levels and activity of *BACE1*, the rate-limiting enzyme in the production of amyloid-β[42].

Several recent studies explored the connection between diseases and genetic variants in coding sequences[15], non-coding regions[13], and promoters[14] in human genome. Our framework for post-GWAS analysis integrates these data to predict disease causal variants. Over 90% of disease-associated variants found in GWAS are located in non-coding regions, and aggregate analysis of them has shown that they are significantly enriched in transcription-factor-occupied regions and DNase I hypersensitive sites, most of which overlap enhancer regions[43], suggesting that many disease loci identified by GWASs influence transcriptional output of one or more target genes through enhancer variants. However, it has been difficult, computationally or experimentally, to identify risk variants in enhancers for any disease. In our study, AD risk variants located in promoters or enhancers likely modulate the disease risk by influencing the expression of 85 AD risk genes, among which 50 (58.8%) are differentially expressed in at least one brain region, including 27 up-regulated genes, 20 down-regulated genes, and 3 genes showing both up- and down-regulation in different brain regions (Supplementary Tables S6 and S7). This indicates that these genes are likely to modulate the disease risk through altering gene expression in brain regions. For AD, we predicted 62 risk enhancer variants. For over 66% of them, there is additional support for their functional impact on enhancer: 16 are known eQTLs and 29 can alter transcriptional factor binding through gain or loss of TFBS motifs, implying a strong reliability of these variants should be the risk variants.

AD risk genes enriched in KEGG pathways and GO biological processes highly relevant to the disease pathology. The most significantly enriched KEGG pathway is the complement and coagulation cascades. Increasing evidence suggests that deregulation of the complement cascade is a contributing factor leading to chronic inflammation and neurodegeneration observed in AD[44,45]. The complement system plays an important role in the innate and adaptive immune responses, restricts amyloid plaque formation, and helps clearance of plaque components associated with AD[46]. AD risk genes are also enriched in MAPK signaling pathway, which contributes to the AD pathogenesis through multiple mechanisms, including the regulation of neuronal apoptosis and phosphorylation of APP and tau[47]. Additional enriched pathways, such as GnRH signaling, PI3K-Akt signaling, neurotrophin signaling, and calcium signaling, all have been shown to likely play a role in AD[48–51], whereas Ras signaling may play an important role in aging[52]. Many enriched GO terms—e.g., “nervous system development”, "neurogenesis", “neuron differentiation”, “neuron development”—are related to the development and differentiation of neurons. Some are directly related to AD: e.g., “phosphorylation ($P<0.01$)”, “activation of immune response ($P<0.01$)”, and “learning memory” are part of the currently predominant hypothesis of the pathogenesis of AD[53,54]. In additional, several blood-related GO terms are enriched: “blood vessel development”, “vasculature development”, “negative regulation of blood coagulation”, and “hemopoiesis”. Increasing evidences has shown that the hematopoietic system may contribute to the initiation and/or progression of AD[55].

Clustered expression of our AD risk genes in the second group actively expressed in the immune cells such as B and T lymphocytes, implying immune related genes might highly correlate with AD. Recent studies showed that inflammation contributes to the pathogenesis of AD[56]. It is important to note that many immune related genes, such as *TREM2*, *INPP5D*, *CD34,* and *CD55*, were included in this group. TREM2 is a cell surface receptor of the immunoglobulin superfamily protein expressed in microglia in the CNS. As a potential key molecule in AD pathogenesis, it might protect against neurodegeneration by promoting phagocytosis to clear apoptotic neurons[57] and a broad array of microglial functions in response to Aβ deposition[58]. INPP5D plays a significant role in inflammatory responses and has been implicated in the pathogenesis of late-onset AD through the regulation of microglial and myeloid cell function[3]. This suggests that immune processes may directly contribute to the pathology and progression of AD, rather than being the consequence of the neurodegeneration. Cholesterol metabolism-associated genes in this group, such as *SORL1* and *ABCA7*, have been linked to AD in previous studies. Studies showed that the suppression of *SORL1* expression contribute to the overexpression of Aβ and an increased risk of AD[59]. *ABCA7* is a genetic risk factor for late-onset AD and may participate in the regulation of Aβ homoeostasis in the brain[60]. Moreover, this group also includes tauopathy-associated AD risk genes, such as, *PTK2B* and *PICALM*. Previous studies in *Drosophila* indicated that *PTK2B* acts as an early marker and in vivo modulator of tau toxicity[61]. Cell-based and in vivo data showed that perturbations of *PICALM* levels might be a key for the regulation of autophagy and tau levels and therefore essential for modulating tau toxicity[62].

AD risk genes are expressed in many different types of brain cells. We found that they are over-expressed in microglia, endothelia, and pericytes from three different regions of adult brains, consistent with previous reports[63,64] showing that these cell types are likely to be associated with AD pathology. Previous studies showed that microglia are the primary cells contributing to the initiation of the immune response to AD pathology, and the aberrant microglial activation is a causal factor for the development of AD[65]. Recent studies also suggests endothelial dysfunction may be involved in the pathogenesis of AD[66]. Pericytes, cells in the blood–brain barrier, degenerate in AD and are reported to control multiple steps of AD-alike neurodegeneration cascade in mice overexpressing Aβ-precursor protein[67]. Moreover, oligodendrocytes are only significantly enriched with overexpressed AD risk genes in cerebellum. The major function of oligodendrocytes is the formation of myelin, whose breakdown is associated with AD[68]. We also observed over-expression of AD risk genes in specific cell types or brain regions or both. For example, only oligodendrocytes from cerebellum showed active expression of AD risk genes. When frontal cortex with visual cortex were compared, only In4a and In7 cells from the former and In1a cells from the later showed enriched expression for AD risk genes. These findings were replicated using a single-cell transcriptome data set[63] (Fig. 3D). A previous expression weighted cell-type enrichment analysis of a set of 178 AD risk genes using the same data set failed to identify significantly enriched cell types[64], likely due to the incomplete list AD risk genes used in that analysis.

AD is a progressive neurodegenerative disease that involves alteration of gene expression at the whole transcriptome level. The perturbation in the sub-networks of co-expression involving AD risk genes can partially reflect AD progression. Finding the altered network hub genes involved in AD progression may help identify AD biomarkers.

We found hub genes, such as *RNF6*, *TP53INP*, and *GGH* genes, with high connectivity in AD patients but tended to have low connectivity in healthy individuals. Rnf6, a ring-finger-dependent ubiquitin ligase, functions for proteasomal degradation in axonal growth cones of primary hippocampal neurons in mice by regulating the levels of *Limk1*, which play a crucial role in neurodevelopment and synaptic plasticity[69]. *TP53INP1*, a major regulator of p53 in response to oxidative stress[70], is a tumor suppressor associated with malignant tumor metastasis in breast, liver, pancreas, and stomach and plays a critical role in cancer progression. Interestingly, previous studies showed inverse correlation between cancer and AD[71]. It has been reported that tripeptide GGH might be used for Cu chelation therapy for AD treatment as Cu ion level was reported to be elevated in AD brains and accumulation of amyloid plaques leading to metal homeostasis dysregulation[72]. Also, we found many hub genes, such as *LMTK2*, *MAPT*, *USP8*, and *SPPL2A* genes, in normal people turn into have low connectivity in AD patients. *LMTK2* may contribute to the neurodegenerative process by disrupting axonal transport, tau hyperphosphorylation and enhancing apoptosis[73]. Its expression is decreased in a tau mouse model of AD[74]. As one of the deubiquitinases, which play a critical role in regulating synaptic function and whose dysfunction results in several neurological disorders, *USP8* has been shown to be associated with AD[42], Parkinson's disease, and Lewy body disease. *MAPT* encodes tau protein, whose hyper-phosphorylation and subsequent intracellular neurofibrillary entanglement is one of definitive neuropathological hallmarks of AD. SPPL2a is an intramembrane protease of lysosomes/late endosomes and plays a critical role in regulation of intramembrane proteolysis in B cells and the regulation of innate and adaptive immunity[75].

As expected, many predicted AD risk genes were also identified by survival analysis as predictors for AD prognosis. They include *IL1RAP*, *PMAIP1*, *LAMTOR4*, and *GRB2*. IL1RAP, a key immune signaling factor, impacts amyloid accumulation by modulating the activity of microglia and is crucial in clearing brain amyloid and limiting plaque growth[76]. PMAIP1 is an essential mediator of p53-dependent apoptosis, an important biological process in neurodegenerative disorders[77]. LAMTOR4 is a component of the Rag-Regulator complex and an essential regulator of lysosomes in microglia. Its absence has been shown to result in diminished number of microglia in Zebrafish[78]. GRB2, a cytoplasmic protein, are involved in protecting the cytoskeletal architecture in AD-like conditions[79] and interacts with the C-terminal fragment or tyrosine-phosphorylated APP. This interaction intensifies significantly in neuronal cells and AD brains[80]. Finally, it is worth noting that the predictor genes for AD prognosis are region-specific, as none was identified in BM44 and only one (*GRB2*) in BM33. Because brain samples can only be obtained after death, with the assumption that postmortem data are indicative of the long-term gene expression state before death, we performed the survival analysis to detect predictors for AD prognosis and to understand the expression patterns of AD risk genes. Although this assumption may not be true for every gene and thus could lead to bias in our results, the survival analysis can nevertheless be helpful in detecting potential predictors for AD prognosis, which can be experimentally validated and examined.

Limitations. In this study, we developed an integrated computational framework aimed to predict both disease genes and corresponding risk variants using multiple omics data sets. Although our results showed that the predicted risk variants are likely to be functional, they need be experimentally validated and studied using transgenic AD mouse models[81] and the luciferase reporter assay technology. Moreover, this study did not consider the impact of other critical covariates such as sex and APOE on the results.

Overall, in this study we prioritized new potential AD risk genes and risk variants from AD GWAS by integrating comprehensive omics data. Moreover, we provided comprehensive functional annotation to those putative risk genes and variants. Our findings will facilitate translating AD genetics to potential therapeutic target genes/variants and can be used to devise new strategies for AD therapeutics development.

Materials and methods

In this study, we developed a computational pipeline (Fig. 1) that integrates multiple data sets to predict AD risk genes as well as their risk variants, and systematically investigated their characteristics and functions. We first used the PGA software tool[82] to identify AD risk genes on the whole-genome level, and then examined their expression patterns across different human cells and tissues and during brain development, their co-expression network, and survival curves at the transcriptomic level. Next, using an analytical framework that we developed, we predicted potential risk variants for identified AD risk genes.

Data sets and their sources.

AD GWAS SNPs. From 58 AD GWAS collected in the GWAS Catalog[83] (as of December 2018) and the latest study[5] (meta-analysis of AD-proxy samples and LOAD samples, a valuable resource for our study to identify potential AD risk regions and is beneficial for our study to identify potential AD risk genes and variants), we collected 936 AD-associated SNPs ($P<1E{-}5$, Fig. 2 and Supplementary Table S1) with a lenient P value (suggestive evidence of association[84]) to contain more potential AD risk genes candidates being analyzed.

AD genes for PGA training. 338 AD genes (Supplementary Table S9) were collected from three disease-gene databases: MalaCards[10], DISEASES[11], and DisGeNET v5.0[12]. They were separated into two groups: a stringent set of 98 AD genes supported by at least two of these three disease-gene databases and a lenient set of 240 AD genes supported by only one of the three databases for AD.

Differentially expressed genes (DEGs). We assembled a list of 10,314 genes that have been shown in three studies[27–29] to be differentially expressed in nine brain regions between AD patients and normal controls (Supplementary Table S10).

Variant and genome annotation. We used multiple annotation metrics for various data analyses in this study, including scores from LINSIGHT[13], ExPecto[14], and PrimateAI[15], enhancer-gene connection from HEDD[85], eQTL from GTEx v.6.p[86], and Synapse (syn17015233[36]).

Identification of genomic risk regions and putative candidate genes. Using PGA[82], able to integrate different types of data to uncover plausible risk genes implicated by GWAS signals that might be missed by other methods, we first systematically identified genomic AD-risk regions. Each AD risk region is an LD block that is seeded with a GWAS-identified AD variant and defined by 1000 Genomes Project variants that are in substantial LD ($r^2 > 0.5$[30]) with the AD variant and less than 400 kb away from it. Overlapping or close (< 250 kb) LD blocks were merged. AD risk gene candidates consisted of proximal genes that overlap (after extending their ranges by 20 kb on each end[87]) these genomic AD risk regions and distal genes that are more than 20 kb away from AD risk regions but are linked to them by regulatory elements within them. We then integrated both gene network and annotation data with GWAS signals to score all candidates for AD risk. Gene regulatory information about enhancers and eQTLs in 44 human tissues (including non-brain tissues, which could provide complementary gene regulatory information) and their target genes from HEDD[85] and GTEx (v6p)[86] were used to identify distal AD risk gene candidates. Risk gene candidates were scored by:

$$S_g = \alpha S_f^{(n)} + (1 - \alpha) S_f^{(a)} (0 < \alpha < 1)$$

in which S_g is gene score, $S_f^{(n)}$ and $S_f^{(a)}$ are the network and annotation-based scores, respectively, and α is a coefficient controlling the relative weights of these two scores on the final gene score (see more details in "scoring risk gene candidates" section in[82]) with aforementioned training gene sets, and high-scoring candidates with scores greater than the threshold (achieves a prediction precision ≥ 0.8) estimated with either the stringent (= 21.4) or the lenient (= 13.1) gene set were taken as (putative) AD risk genes for downstream analysis.

Characterization of AD risk genes. *Functional enrichment analysis.* We used DAVID[21] to identify KEGG[88] pathways and GO terms[89] enriched with AD risk genes. *P*-values were adjusted for multiple tests using the Bonferroni or FDR methods.

Expression weighted cell-type enrichment (EWCE) analysis. We used EWCE (v1.2.0)[90] to identify brain cell types, which are more likely to be affected by AD risk genes. Single-cell RNA-seq data from 33 types of cells (Supplementary Table S5) from human adult brains (20–51 years old; a total of 35,289 cells; from the visual cortex, frontal cortex, and cerebellum)[91] were used in this analysis. The average expression across all samples for each cell type was used to determine enriched expression by EWCE.

Tissue gene expression analysis. To investigate gene-tissue expression specificity, we examined the expression profiles of AD risk genes across different tissues from the Gene Enrichment Profiler (http://xavierlab2.mgh.harvard.edu/EnrichmentProfiler/help.html)[92], which catalogs normalized expression values of ~ 12,000 genes across 126 primary human tissues. We grouped the putative risk genes into different clusters according to their different expression profile across tissues using the Euclidean distance and the Ward's clustering method[93]. The heatmap of the gene expression was plotted using R 'gplots' package with the 'heatmap.2' function.

Analysis of expression of AD risk genes during human brain development. Temporal gene expression patterns in different brain tissues and their regulation across the lifespan reveal molecular mechanisms involved in the formation, maturation, and degeneration of the human brain. Here, we examined the spatiotemporal expression patterns of AD risk genes in adult human brain. We first compiled gene expression profiles across ten regions of adult human brains from individuals of 19 to 40 years old, using RNA-seq data from BrainSpan[94] (http://www.brainspan.org/, as of April 2018). We then explored the temporal dynamics of gene expression in a more specific human brain region (i.e., prefrontal cortex) during brain development from fetuses to older adults, using spotted oligonucleotide microarrays yielding data from Brain Cloud[95] (http://braincloud.jhmi.edu/, as of May 2018).

Gene co-expression networks. We built gene co-expression networks using transcriptomic and proteomic data. From Synapse (syn7391833[96]), we collected expression profiles of ~ 23,200 genes (included 315 AD risk genes) in four human brain regions from normal controls and AD patients with the definitive disease status: frontal pole (Brodmann area 10, BM10) with 111 cases and 76 controls, superior temporal gyrus (BM22) with 102 cases and 65 controls, parahippocampal gyrus (BM36) with 90 cases and 68 controls, and inferior frontal gyrus (BM44) with 90 cases and 64 controls. From Synapse (syn10239444[97]), we also collected proteomic profiles of ~ 10,000 gene products in two human brain regions: frontal cortex and anterior cingulate gyrus (ACG), both with 10 cases and 10 controls. We first removed proteins with less than three samples in AD patients or normal controls, and then imputed missing values (0 or '#N/A') by replacing them with averaged values from AD and control samples, separately. We constructed gene co-expression networks using the Pearson's correlation ($r > 0.7$) among AD patients and normal controls for each brain region. The co-expression networks were drawn using the 'igraph' (v1.2.4.1) R package.

Survival analysis. Among 342 AD risk genes, only 315 of them have expression data available. Among 315 AD risk genes, we carried out survival analysis (using the 'survival' R package) to identify ones that may predict by their expression levels the disease prognosis among AD patients. For each AD risk gene, using its expression profile (described above), we first separated AD patients into two groups, with either high (≥ the median) or low (< the median) expression of the gene. We then used the Kaplan–Meier estimator to compute estimated survivor functions, plotted them to directly compare survival between these two groups, and used the log-rank test to formally test whether their survival curves are identical.

Prediction of AD risk variants. Risk variants contribute to the disease etiology and pathology mainly through two major mechanisms: coding variants may alter the function of gene products and non-coding variants may alter the transcription of genes by changing regulatory elements. Recently developed methods for variant annotation, such as LINSIGHT[13], ExPecto[14], and PrimateAI[15], can quantify or provide useful information about the functional impact of genetic variants in the human genome, regardless of their potential connection to a specific disease or trait. By integrating such functional annotation of variants as a part of the computational framework that we developed for the post-GWAS analysis (Fig. 1), we are able to not only identify disease risk genes but also further predict disease risk variants. As LINSIGHT[13], ExPecto[14], and PrimateAI[15] are so far the most completeness and up-to-date functional annotation of variants resources, we used them to prioritize the potential risk variant in this study. In this framework, we first use the variant type as the guide. For coding variants, we assume causality through the functional alteration of gene products and use their PrimateAI scores[15], on a scale from 0 (benign) to 1 (pathogenic) measuring pathogenicity, to prioritize them. For non-coding variants, we assume causality is through gene regulation and prioritize variants in promoters by their ExPecto scores[14], expression effect of fold change, and ones in enhancers by their LINSIGHT scores[13], ranging from 0 to 1 and higher score indicating deleterious fitness consequences of the variants.

We systematically fine-mapped the genomic neighborhoods of predicted risk variants by analyzing all genotyped and imputed variants from the latest AD GWAS[5]. Only variants mapped to risk region and coding sequences and regulatory elements (i.e., promoters and enhancers) of AD risk genes were selected for downstream analysis. We used the following criteria to predict risk variants: PrimateAI scores, which measure the pathogenicity of coding variants from 0 (less pathogenic) to 1 (more pathogenic), are greater than 0.7; ExPecto scores, which predict functional impact of non-coding variants in promoters, are greater than ln(fold change = 1.2); and LINSIGHT scores, which measure the probability of negative selection on non-coding sites and can be used to prioritize SNVs associated with genetic diseases, are greater than 0.9.

Transcription factor binding site motif analysis. The reference and the alternative alleles of each variant, along with ± 25 bp of flanking sequences, were analyzed using HOMER (v4.9.1)[98]. We used the findMotifs.pl program with the default parameters to find transcription factor binding site (TFBS) motifs in the reference and the alternative allele sequences, which were used as the background for each other to control the nucleotide context. A library of 364 vertebrate TFBS motifs in the format of position weight matrices was searched for matches. The matched motifs with scores greater than 7.5 were regarded as gain or loss of TFBS motifs.

Data availability

Only publicly available data were used in this study. See the 'Data sets' subsection above for their availability. Computer code for major steps of data processing is available from the GitHub (https://github.com/zhenwang19/AD).

Received: 18 November 2020; Accepted: 23 September 2021
Published online: 15 October 2021

References

1. Masters, C. L. *et al.* Alzheimer's disease. *Nat. Rev. Dis. Primers* **1**, 15056. https://doi.org/10.1038/nrdp.2015.56 (2015).
2. Cacace, R., Sleegers, K. & Van Broeckhoven, C. Molecular genetics of early-onset Alzheimer's disease revisited. *Alzheimers Dement.* **12**, 733–748. https://doi.org/10.1016/j.jalz.2016.01.012 (2016).
3. Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452-U1206. https://doi.org/10.1038/ng.2802 (2013).
4. Liu, C. C., Liu, C. C., Kanekiyo, T., Xu, H. & Bu, G. Apolipoprotein E and Alzheimer disease: Risk, mechanisms and therapy. *Nat. Rev. Neurol.* **9**, 106–118. https://doi.org/10.1038/nrneurol.2012.263 (2013).
5. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* https://doi.org/10.1038/s41588-018-0311-9 (2019).
6. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430. https://doi.org/10.1038/s41588-019-0358-2 (2019).
7. Novikova, G. *et al.* Integration of Alzheimer's disease genetics and myeloid genomics identifies disease risk regulatory elements and genes. *Nat. Commun.* https://doi.org/10.1038/s41467-021-21823-y (2021).
8. Zhao, T. Y., Hu, Y., Zang, T. Y. & Wang, Y. D. Integrate GWAS, eQTL, and mQTL data to identify Alzheimer's disease-related genes. *Front. Genet.* https://doi.org/10.3389/fgene.2019.01021 (2019).
9. Hao, S. C., Wang, R., Zhang, Y. & Zhan, H. Prediction of Alzheimer's disease-associated genes by integration of GWAS summary data and expression data. *Front. Genet.* https://doi.org/10.3389/fgene.2018.00653 (2019).

10. Rappaport, N. *et al.* MalaCards: An amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res* **45**, D877–D887. https://doi.org/10.1093/nar/gkw1012 (2017).
11. Pletscher-Frankild, S., Palleja, A., Tsafou, K., Binder, J. X. & Jensen, L. J. DISEASES: Text mining and data integration of disease-gene associations. *Methods* **74**, 83–89. https://doi.org/10.1016/j.ymeth.2014.11.020 (2015).
12. Pinero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **45**, D833–D839. https://doi.org/10.1093/nar/gkw943 (2017).
13. Huang, Y. F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49**, 618–624. https://doi.org/10.1038/ng.3810 (2017).
14. Zhou, J. *et al.* Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171. https://doi.org/10.1038/s41588-018-0160-6 (2018).
15. Sundaram, L. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161. https://doi.org/10.1038/s41588-018-0167-z (2018).
16. Zhu, X. C. *et al.* CR1 in Alzheimer's disease. *Mol. Neurobiol.* **51**, 753–765. https://doi.org/10.1007/s12035-014-8723-8 (2015).
17. De Roeck, A., Van Broeckhoven, C. & Sleegers, K. The role of ABCA7 in Alzheimer's disease: Evidence from genomics, transcriptomics and methylomics. *Acta Neuropathol.* **138**, 201–220. https://doi.org/10.1007/s00401-019-01994-1 (2019).
18. Carmona, S. *et al.* The role of TREM2 in Alzheimer's disease and other neurodegenerative disorders. *Lancet Neurol.* **17**, 721–730. https://doi.org/10.1016/S1474-4422(18)30232-1 (2018).
19. Yin, R. H., Yu, J. T. & Tan, L. The role of SORL1 in Alzheimer's disease. *Mol. Neurobiol.* **51**, 909–918. https://doi.org/10.1007/s12035-014-8742-5 (2015).
20. Tan, M. S., Yu, J. T. & Tan, L. Bridging integrator 1 (BIN1): Form, function, and Alzheimer's disease. *Trends Mol. Med.* **19**, 594–603. https://doi.org/10.1016/j.molmed.2013.06.004 (2013).
21. da Huang, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57. https://doi.org/10.1038/nprot.2008.211 (2009).
22. Chapuis, J. *et al.* Increased expression of BIN1 mediates Alzheimer genetic risk by modulating tau pathology. *Mol. Psychiatr.* **18**, 1225–1234. https://doi.org/10.1038/mp.2013.1 (2013).
23. Allen, M. *et al.* Association of MAPT haplotypes with Alzheimer's disease risk and MAPT brain gene expression levels. *Alzheimers Res. Ther.* https://doi.org/10.1186/alzrt268 (2014).
24. Hirano, A. *et al.* A genome-wide association study of late-onset Alzheimer's disease in a Japanese population. *Psychiatr. Genet.* **25**, 139–146. https://doi.org/10.1097/Ypg.0000000000000090 (2015).
25. Crehan, H. *et al.* Complement receptor 1 (CR1) and Alzheimer's disease. *Immunobiology* **217**, 244–250. https://doi.org/10.1016/j.imbio.2011.07.017 (2012).
26. Strunz, M. *et al.* Modulation of SPARC/Hevin proteins in Alzheimer's disease brain injury. *J. Alzheimers Dis.* **68**, 695–710. https://doi.org/10.3233/JAD-181032 (2019).
27. Li, X., Long, J., He, T., Belshaw, R. & Scott, J. Integrated genomic approaches identify major pathways and upstream regulators in late onset Alzheimer's disease. *Sci. Rep.* **5**, 12393. https://doi.org/10.1038/srep12393 (2015).
28. Ciryam, P. *et al.* A transcriptional signature of Alzheimer's disease is associated with a metastable subproteome at risk for aggregation. *Proc. Natl. Acad. Sci. USA* **113**, 4753–4758. https://doi.org/10.1073/pnas.1516604113 (2016).
29. syn14237651. https://www.synapse.org/#!Synapse:syn14237651
30. Lin, J. R. *et al.* Integrated post-GWAS analysis sheds new light on the disease mechanisms of schizophrenia. *Genetics* **204**, 1587–1600. https://doi.org/10.1534/genetics.116.187195 (2016).
31. Holton, P. *et al.* Initial assessment of the pathogenic mechanisms of the recently identified Alzheimer risk Loci. *Ann. Hum. Genet.* **77**, 85–105. https://doi.org/10.1111/ahg.12000 (2013).
32. N'Songo, A. *et al.* African American exome sequencing identifies potential risk variants at Alzheimer disease loci. *Neurol. Genet.* https://doi.org/10.1212/NXG.0000000000000141 (2017).
33. Tao Yang, P. T., Zhao, S. & Ye, J. Identifying Genetic Risk Factors via Sparse Group Lasso with Group Graph Structure. *arXiv preprint* http://arxiv.org/abs/1709.03645 (2017).
34. Jun, G. *et al.* A novel Alzheimer disease locus located near the gene encoding tau protein. *Mol. Psychiatry* **21**, 108–117. https://doi.org/10.1038/mp.2015.23 (2016).
35. van Arensbergen, J. *et al.* High-throughput identification of human SNPs affecting regulatory element activity. *Nat. Genet.* **51**, 1160–1169. https://doi.org/10.1038/s41588-019-0455-2 (2019).
36. syn17015233. https://www.synapse.org/#!Synapse:syn17015233
37. Rogaeva, E. *et al.* The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer disease. *Nat. Genet.* **39**, 168–177. https://doi.org/10.1038/ng1943 (2007).
38. Hashimoto, Y. *et al.* Transforming growth factor beta2 autocrinally mediates neuronal cell death induced by amyloid-beta. *J. Neurosci. Res.* **83**, 1039–1047. https://doi.org/10.1002/jnr.20804 (2006).
39. von Bernhardi, R., Cornejo, F., Parada, G. E. & Eugenin, J. Role of TGFbeta signaling in the pathogenesis of Alzheimer's disease. *Front. Cell Neurosci.* **9**, 426. https://doi.org/10.3389/fncel.2015.00426 (2015).
40. Tong, Y., Xu, Y., Scearce-Levie, K., Ptacek, L. J. & Fu, Y. H. COL25A1 triggers and promotes Alzheimer's disease-like pathology in vivo. *Neurogenetics* **11**, 41–52. https://doi.org/10.1007/s10048-009-0201-5 (2010).
41. Kaur, G. & Levy, E. Cystatin C in Alzheimer's disease. *Front. Mol. Neurosci.* **5**, 79. https://doi.org/10.3389/fnmol.2012.00079 (2012).
42. Yeates, E. F. & Tesco, G. The endosome-associated deubiquitinating enzyme USP8 regulates BACE1 enzyme ubiquitination and degradation. *J. Biol. Chem.* **291**, 15753–15766. https://doi.org/10.1074/jbc.M116.718023 (2016).
43. Farh, K. K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343. https://doi.org/10.1038/nature13835 (2015).
44. Eikelenboom, P. & Veerhuis, R. The role of complement and activated microglia in the pathogenesis of Alzheimer's disease. *Neurobiol. Aging* **17**, 673–680 (1996).
45. Shen, Y. & Meri, S. Yin and Yang: complement activation and regulation in Alzheimer's disease. *Prog. Neurobiol.* **70**, 463–472 (2003).
46. Morgan, B. P. Complement in the pathogenesis of Alzheimer's disease. *Semin. Immunopathol.* **40**, 113–124. https://doi.org/10.1007/s00281-017-0662-9 (2018).
47. Zhu, X., Lee, H. G., Raina, A. K., Perry, G. & Smith, M. A. The role of mitogen-activated protein kinase pathways in Alzheimer's disease. *Neurosignals* **11**, 270–281. https://doi.org/10.1159/000067426 (2002).
48. Wang, L. *et al.* Gonadotropin-releasing hormone receptor system: Modulatory role in aging and neurodegeneration. *CNS Neurol. Disord. Drug Targets* **9**, 651–660 (2010).
49. Zhang, B. *et al.* Neuroprotective effects of salidroside through PI3K/Akt pathway activation in Alzheimer's disease models. *Drug Des. Dev. Ther.* **10**, 1335–1343. https://doi.org/10.2147/Dddt.S99958 (2016).
50. Mitre, M., Mariga, A. & Chao, M. V. Neurotrophin signalling: Novel insights into mechanisms and pathophysiology. *Clin. Sci.* **131**, 13–23. https://doi.org/10.1042/Cs20160044 (2017).
51. Berridge, M. J. Calcium signalling and Alzheimer's disease. *Neurochem. Res.* **36**, 1149–1156. https://doi.org/10.1007/s11064-010-0371-4 (2011).

52. Slack, C. Ras signaling in aging and metabolic regulation. *Nutr. Healthy Aging* **4**, 195–205. https://doi.org/10.3233/NHA-160021 (2017).
53. Wang, Y. P. & Mandelkow, E. Tau in physiology and pathology. *Nat. Rev. Neurosci.* **17**, 5–21. https://doi.org/10.1038/nrn.2015.1 (2016).
54. Selkoe, D. J. & Hardy, J. The amyloid hypothesis of Alzheimer's disease at 25 years. *Embo Mol. Med.* **8**, 595–608. https://doi.org/10.15252/emmm.201606210 (2016).
55. Lampron, A., Gosselin, D. & Rivest, S. Targeting the hematopoietic system for the treatment of Alzheimer's disease. *Brain Behav. Immun.* **25**, S71–S79. https://doi.org/10.1016/j.bbi.2010.12.018 (2011).
56. Heppner, F. L., Ransohoff, R. M. & Becher, B. Immune attack: the role of inflammation in Alzheimer disease. *Nat. Rev. Neurosci.* **16**, 358–372. https://doi.org/10.1038/nrn3880 (2015).
57. Jiang, T. *et al.* Upregulation of TREM2 ameliorates neuropathology and rescues spatial cognitive impairment in a transgenic mouse model of Alzheimer's disease. *Neuropsychopharmacology* **39**, 2949–2962. https://doi.org/10.1038/npp.2014.164 (2014).
58. Wang, Y. *et al.* TREM2 lipid sensing sustains the microglial response in an Alzheimer's disease model. *Cell* **160**, 1061–1071. https://doi.org/10.1016/j.cell.2015.01.049 (2015).
59. Andersen, O. M. *et al.* Neuronal sorting protein-related receptor sorLA/LR11 regulates processing of the amyloid precursor protein. *Proc. Natl. Acad. Sci. USA* **102**, 13461–13466. https://doi.org/10.1073/pnas.0503689102 (2005).
60. Li, H., Karl, T. & Garner, B. Understanding the function of ABCA7 in Alzheimer's disease. *Biochem. Soc. Trans.* **43**, 920–923. https://doi.org/10.1042/BST20150105 (2015).
61. Dourlen, P. *et al.* Functional screening of Alzheimer risk loci identifies PTK2B as an in vivo modulator and early marker of Tau pathology. *Mol. Psychiatr.* **22**, 874–883. https://doi.org/10.1038/mp.2016.59 (2017).
62. Moreau, K. *et al.* PICALM modulates autophagy activity and tau accumulation. *Nat. Commun.* **5**, 4998. https://doi.org/10.1038/ncomms5998 (2014).
63. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. USA* **112**, 7285–7290. https://doi.org/10.1073/pnas.1507125112 (2015).
64. Wang, P., Zhao, D., Lachman, H. M. & Zheng, D. Enriched expression of genes associated with autism spectrum disorders in human inhibitory neurons. *Transl. Psychiatry* **8**, 13. https://doi.org/10.1038/s41398-017-0058-6 (2018).
65. Hansen, D. V., Hanson, J. E. & Sheng, M. Microglia in Alzheimer's disease. *J. Cell Biol.* **217**, 459–472. https://doi.org/10.1083/jcb.201709069 (2018).
66. Kelleher, R. J. & Soiza, R. L. Evidence of endothelial dysfunction in the development of Alzheimer's disease: Is Alzheimer's a vascular disorder?. *Am. J. Cardiovasc. Dis.* **3**, 197–226 (2013).
67. Sagare, A. P. *et al.* Pericyte loss influences Alzheimer-like neurodegeneration in mice. *Nat. Commun.* **4**, 2932. https://doi.org/10.1038/ncomms3932 (2013).
68. Cai, Z. & Xiao, M. Oligodendrocytes and Alzheimer's disease. *Int. J. Neurosci.* **126**, 97–104. https://doi.org/10.3109/00207454.2015.1025778 (2016).
69. Cuberos, H. *et al.* Roles of LIM kinases in central nervous system function and dysfunction. *FEBS Lett.* **589**, 3795–3806. https://doi.org/10.1016/j.febslet.2015.10.032 (2015).
70. Cano, C. E. *et al.* Tumor protein 53-induced nuclear protein 1 is a major mediator of p53 antioxidant function. *Cancer Res.* **69**, 219–226. https://doi.org/10.1158/0008-5472.Can-08-2320 (2009).
71. Driver, J. A. *et al.* Inverse association between cancer and Alzheimer's disease: Results from the Framingham Heart Study. *Brit. Med. J.* https://doi.org/10.1136/bmj.e1442 (2012).
72. Miller, L. M. *et al.* Synchrotron-based infrared and X-ray imaging shows focalized accumulation of Cu and Zn co-localized with beta-amyloid deposits in Alzheimer's disease. *J. Struct. Biol.* **155**, 30–37. https://doi.org/10.1016/jjsb.2005.09.004 (2006).
73. Bencze, J. *et al.* Biological function of Lemur tyrosine kinase 2 (LMTK2): Implications in neurodegeneration. *Mol. Brain* **11**, 20. https://doi.org/10.1186/s13041-018-0363-x (2018).
74. Matarin, M. *et al.* A genome-wide gene-expression analysis and database in transgenic mice during development of amyloid or tau pathology. *Cell Rep.* **10**, 633–644. https://doi.org/10.1016/j.celrep.2014.12.041 (2015).
75. Friedmann, E. *et al.* SPPL2a and SPPL2b promote intramembrane proteolysis of TNFalpha in activated dendritic cells to trigger IL-12 production. *Nat. Cell Biol.* **8**, 843–848. https://doi.org/10.1038/ncb1440 (2006).
76. Condello, C., Yuan, P., Schain, A. & Grutzendler, J. Microglia constitute a barrier that prevents neurotoxic protofibrillar A beta 42 hotspots around plaques. *Nat. Commun.* https://doi.org/10.1038/ncomms7176 (2015).
77. Mattson, M. P. Apoptosis in neurodegenerative disorders. *Nat. Rev. Mol. Cell Biol.* **1**, 120–129. https://doi.org/10.1038/35040009 (2000).
78. Shen, K., Sidik, H. & Talbot, W. S. The Rag-Ragulator complex regulates lysosome function and phagocytic flux in microglia. *Cell Rep.* **14**, 547–559. https://doi.org/10.1016/j.celrep.2015.12.055 (2016).
79. Majumder, P., Roy, K., Singh, B. K., Jana, N. R. & Mukhopadhyay, D. Cellular levels of Grb2 and cytoskeleton stability are correlated in a neurodegenerative scenario. *Dis. Model Mech.* **10**, 655–669. https://doi.org/10.1242/dmm.027748 (2017).
80. Russo, C. *et al.* Signal transduction through tyrosine-phosphorylated C-terminal fragments of amyloid precursor protein via an enhanced interaction with Shc/Grb2 adaptor proteins in reactive astrocytes of Alzheimer's disease brain. *J. Biol. Chem.* **277**, 35282–35288. https://doi.org/10.1074/jbc.M110785200 (2002).
81. Gotz, J., Bodea, L. G. & Goedert, M. Rodent models for Alzheimer disease. *Nat. Rev. Neurosci.* **19**, 583–598. https://doi.org/10.1038/s41583-018-0054-8 (2018).
82. Lin, J. R. *et al.* PGA: post-GWAS analysis for disease gene identification. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btx845 (2017).
83. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901. https://doi.org/10.1093/nar/gkw1133 (2017).
84. Logue, M. W. *et al.* A comprehensive genetic association study of Alzheimer disease in African Americans. *Arch. Neurol.* **68**, 1569–1579. https://doi.org/10.1001/archneurol.2011.646 (2011).
85. Wang, Z. *et al.* HEDD: Human enhancer disease database. *Nucleic Acids Res.* **46**, D113–D120. https://doi.org/10.1093/nar/gkx988 (2018).
86. Consortium, G. T. The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585. https://doi.org/10.1038/ng.2653 (2013).
87. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421. https://doi.org/10.1038/nature13595 (2014).
88. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30. https://doi.org/10.1093/nar/28.1.27 (2000).
89. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* **25**, 25–29. https://doi.org/10.1038/75556 (2000).
90. Skene, N. G. & Grant, S. G. Identification of vulnerable cell types in major brain disorders using single cell transcriptomes and expression weighted cell type enrichment. *Front. Neurosci.* **10**, 16. https://doi.org/10.3389/fnins.2016.00016 (2016).
91. Lake, B. B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80. https://doi.org/10.1038/nbt.4038 (2018).

92. Benita, Y. *et al.* Gene enrichment profiles reveal T-cell development, differentiation, and lineage-specific transcription factors including ZBTB25 as a novel NF-AT repressor. *Blood* **115**, 5376–5384. https://doi.org/10.1182/blood-2010-01-263855 (2010).
93. Murtagh, F. & Legendre, P. Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion?. *J. Classif.* **31**, 274–295. https://doi.org/10.1007/s00357-014-9161-z (2014).
94. Hawrylycz, M. J. *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391–399. https://doi.org/10.1038/nature11405 (2012).
95. Colantuoni, C. *et al.* Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature* **478**, 519–523. https://doi.org/10.1038/nature10524 (2011).
96. syn7391833. https://www.synapse.org/#!Synapse:syn7391833
97. Ping, L. *et al.* Global quantitative analysis of the human brain proteome in Alzheimer's and Parkinson's disease. *Sci. Data* **5**, 180036. https://doi.org/10.1038/sdata.2018.36 (2018).
98. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589. https://doi.org/10.1016/j.molcel.2010.05.004 (2010).

Acknowledgements

We thank Yousin Suh of the Department of Genetics at the Albert Einstein College of Medicine and Michael G. Rosenfeld of the School of Medicine at the University of California, San Diego, for their expert advice on the biology of enhancers.

Author contributions

Z.D.Z. conceived and designed this study. Z.W. carried out the analyses and also participated in the study design. Z.W. and Z.D.Z. wrote the manuscript. Q.Z., J.R.L., M.R.J., J.M., N.N. provided useful input for the analyses and helped edit the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the National Institutes of Health grants R01 HG008153 from the National Human Genome Research Institute and R01 AG057909 from the National Institute on Aging to Z.D.Z.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-99352-3.

Correspondence and requests for materials should be addressed to Z.D.Z.