Integrating human brain proteomes with genome-wide association data implicates new proteins in Alzheimer’s disease pathogenesis

Aliza P. Wingo1,2✉, Yue Liu3, Ekaterina S. Gerasimov3, Jake Gockley4, Benjamin A. Logsdon4, Duc M. Duong5, Eric B. Dammer5, Chloe Robins3, Thomas G. Beach6, Eric M. Reiman7, Michael P. Epstein8, Philip L. De Jager9, James J. Lah3, David A. Bennett10, Nicholas T. Seyfried5, Allan L. Levey3 and Thomas S. Wingo3,8✉

Genome-wide association studies (GWAS) have identified many risk loci for Alzheimer’s disease (AD)1,2, but how these loci confer AD risk is unclear. Here, we aimed to identify loci that confer AD risk through their effects on brain protein abundance to provide new insights into AD pathogenesis. To that end, we integrated AD GWAS results with human brain proteomes to perform a proteome-wide association study (PWAS) of AD, followed by Mendelian randomization and colocalization analysis. We identified 11 genes that are consistent with being causal in AD, acting via their cis-regulated brain protein abundance. Nine replicated in a confirmation PWAS and eight represent new AD risk genes not identified before by AD GWAS. Furthermore, we demonstrated that our results were independent of APOE e4. Together, our findings provide new insights into AD pathogenesis and promising targets for further mechanistic and therapeutic studies.

AD affects 35 million people worldwide, but there is no effective disease-modifying treatment for it1. To support the development of new AD therapeutics, genetic studies of AD, especially GWAS, have identified many risk loci1,2, but how these loci contribute to AD is unclear. To gain insight into how these loci contribute to AD pathogenesis, we integrated AD GWAS results1 with human brain proteomes3 to identify genes that confer AD risk through their effects on brain protein abundance.

In the discovery phase, we performed a PWAS by integrating AD GWAS results (n = 455,258)3 with 376 human brain proteomes profiled from the dorsolateral prefrontal cortex (dPFC; Supplementary Table 1a)3 using the FUSION pipeline3. Before integration, the proteomic profiles underwent quality control and the effects of clinical characteristics and technical factors were regressed out before we estimated the effects of genetic variants on protein abundance, referred to as protein weights. After quality control, the proteomic profiles included 8,356 proteins, of which 1,475 were heritable and their protein weights could be estimated for the PWAS. The PWAS identified 13 genes whose cis-regulated brain protein levels were associated with AD at a false discovery rate (FDR) of P < 0.05 (Fig. 1, Table 1, Extended Data Fig. 1a and Supplementary Table 2). A confirmation PWAS was performed using the same AD GWAS3 and an independent set of 152 human brain proteomes profiled from the dPFC (Supplementary Table 1b)3. After quality control, 8,168 proteins remained and 1,139 were heritable. Correlation between the protein weights in the discovery and confirmation datasets was high (median = 0.85, interquartile range = 0.21; Supplementary Table 3). Three of the 13 discovery PWAS-significant proteins could not be tested in the confirmation PWAS—one protein was not profiled and 2 were profiled but did not have significant heritability, which is likely due to the smaller sample size. Ten of these 13 proteins could be tested and all 10 proteins replicated in the confirmation PWAS (Table 1, Extended Data Fig. 1b and Supplementary Table 4).

Associations in the PWAS of AD may result when a variant is associated with protein expression (that is, the variant is a protein quantitative trait locus (pQTL)) and AD simultaneously, or from a coincidental overlap between pQTLs and sites in linkage disequilibrium with AD GWAS sites. The former is interpreted as evidence supporting either a pleiotropic or causal role for the gene (and will be referred to as consistent with being causal for simplicity) while the latter suggests a noncausal role. We investigated these possibilities using two independent but complementary approaches. First, using a Bayesian colocalization method, COLOC7, we examined the posterior probability for a shared causal variant between a pQTL and AD for the 13 discovery AD PWAS-significant genes. We found 9 of 13 genes consistent with being causal (Table 2 and Supplementary Table 5). Second, we used the summary data-based Mendelian randomization (SMR)8 and its accompanying heterogeneity in dependence instruments (HEIDI)4. The SMR results suggested that the cis-regulated protein abundance mediates the association between genetic variants and AD for all 13 these genes; however, the HEIDI results argued against a causal role for 4 genes due to linkage disequilibrium (Table 2 and Supplementary Table 6). Thus, 9 of the 13 genes have evidence consistent with a causal role in AD by SMR/HEIDI. In summary, we found seven genes with consistent results associated with AD at a false discovery rate (FDR) of P < 0.05 (Fig. 1, Table 1, Extended Data Fig. 1a and Supplementary Table 2).

1Division of Mental Health, Atlanta VA Medical Center, Decatur, GA, USA. 2Department of Psychiatry, Emory University School of Medicine, Atlanta, GA, USA. 3Department of Neurology, Emory University School of Medicine, Atlanta, GA, USA. 4Sage Bionetworks, Seattle, WA, USA. 5Department of Biochemistry, Emory University School of Medicine, Atlanta, GA, USA. 6Banner Sun Health Research Institute, Sun City, AZ, USA. 7Banner Alzheimer’s Institute, Arizona State University and University of Arizona, Phoenix, AZ, USA. 8Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, USA. 9Center for Translational and Computational Neuroimmunology, Department of Neurology and the Taub Institute for Research on Alzheimer’s Disease and the Aging Brain, Columbia University Medical Center, New York, NY, USA. 10Rush Alzheimer’s Disease Center, Rush University Medical Center, Chicago, IL, USA. ✉ e-mail: aliza.wingo@emory.edu; thomas.wingo@emory.edu

NATURE GENETICS | VOL 53 | FEBRUARY 2021 | 143–146 | www.nature.com/naturegenetics
for causality by both COLOC and SMR/HEIDI (CTSH, DOC2A, ICA1L, LACTB, PLEKHA1, SNX32 and STX4; Table 2) and 4 genes with conflicting results for causality by these two approaches (ACE, CARHSP1, RTFDC1 and STX6; Table 2). The results for EPHX2 and PVR argued against causality (Table 2).

By combining the evidence for replication and the results of the causality tests, we identified five genes (CTSH, DOC2A, ICA1L, LACTB and SNX32) with evidence for both replication and causality (Table 3). There were four genes with evidence for replication and mixed results supporting causality (ACE, CARHSP1, RTFDC1 and STX6; Table 3). Thus, among the 13 discovery PWAS-significant genes, 11 were consistent with being causal in AD and 9 of 11 replicated in the confirmation PWAS (Table 3).

Since the APOE e4 allele is strongly associated with AD, we investigated whether APOE e4 influenced our PWAS findings. To that end, we regressed out the effect of APOE e4 from the proteomes...
and used the regressed proteomic profiles to perform the PWAS of AD. That analysis found the 13 original PWAS-significant genes and 6 additional significant genes at an FDR of $P < 0.05$ (ACOT8, DDX58, ISLR2, PITPN1C1, TBC1D1 and TRIM65; Supplementary Table 7). All 13 genes had the same directions of association as those in the discovery PWAS. Moreover, results from the COLOC and SMR/HEIDI tests found the same evidence of causality as the original findings except that $ACE$ was now consistent with causality by both COLOC and SMR/HEIDI compared to the previous mixed findings (Supplementary Tables 8 and 9). The six additional genes were not consistent with being causal by COLOC (Supplementary Table 8). These observations suggest that our findings are unlikely to be influenced by $APOE e4$.

To understand the specificity of the AD PWAS results, we performed PWAS for other brain-relevant and biometric traits. We expected the degree of overlap of significant genes to roughly correspond to their genetic correlations. GWAS results from individuals of European ancestry for clinical AD ($n = 63,926$), amyotrophic lateral sclerosis (ALS; $n = 80,610$), Parkinson’s disease (PD; $n = 1,474,097$), neuroticism ($n = 390,278$), height ($n = 693,529$), body mass index (BMI; $n = 681,275$) and waist-to-hip ratio adjusting for BMI ($n = 694,649$) were combined with the discovery proteomic profiles ($n = 376$) to perform a PWAS of each trait. The PWAS of clinical AD identified 4 genes, ALS7 genes, PD 17 genes, neurotism 72 genes, height 662 genes, BMI 395 genes and waist-to-hip ratio adjusting for BMI ($n = 694,649$) were combined with the discovery proteomic profiles ($n = 376$) to perform a PWAS of each trait. The PWAS of clinical AD identified 4 genes, ALS7 genes, PD 17 genes, neurotism 72 genes, height 662 genes, BMI 395 genes and waist-to-hip ratio adjusting for BMI ($n = 694,649$) were combined with the discovery proteomic profiles ($n = 376$) to perform a PWAS of each trait. The PWAS of clinical AD identified 4 genes, ALS7 genes, PD 17 genes, neurotism 72 genes, height 662 genes, BMI 395 genes and waist-to-hip ratio adjusting for BMI ($n = 694,649$) were combined with the discovery proteomic profiles ($n = 376$) to perform a PWAS of each trait. The PWAS of clinical AD identified 4 genes, ALS7 genes, PD 17 genes, neurotism 72 genes, height 662 genes, BMI 395 genes and waist-to-hip ratio adjusting for BMI ($n = 694,649$).

For the 13 FDR-significant genes in the discovery AD PWAS, the result of COLOC $H_2$, which is the Bayesian posterior probability that a genetic variant is shared by both traits (that is, the genetically regulated protein level and AD), and $P$ values for the SMR and SMR HEIDI tests are given. *Genes not found in the confirmation PWAS. N/A (not applicable) indicates an undetermined result because the number of pQTL SNPs were too small for HEIDI to test. Genes were sorted by whether they were consistent with being a causal variant.

| Gene   | Chromosome | Discovery PWAS | Confirmation PWAS | COLOC | SMR |
|--------|------------|----------------|-------------------|-------|-----|
| CTSH   | 15         | Significant    | Replicated         | Yes   | Yes |
| DOC2A  | 16         | Significant    | Replicated         | Yes   | Yes |
| ICA1L  | 2          | Significant    | Replicated         | Yes   | Yes |
| LACTB  | 15         | Significant    | Replicated         | Yes   | Yes |
| SNX32  | 11         | Significant    | Replicated         | Yes   | Yes |
| ACE    | 17         | Significant    | Replicated         | Yes   | Yes |
| RTFDC1 | 20         | Significant    | Replicated         | Yes   | Yes |
| CARHSP1| 16         | Significant    | Replicated         | Yes   | Yes |
| STX6   | 1          | Significant    | Replicated         | Yes   | Yes |
| STX4*  | 16         | Significant    | Replicated         | Yes   | Yes |
| PVR*   | 19         | Significant    | Replicated         | Yes   | Yes |

Table 3 | Summary of the 11 AD PWAS-significant genes with evidence for being consistent with a causal role in AD

| Gene   | Chromosome | Discovery PWAS | Confirmation PWAS | Evidence for causality | TWAS significant | New gene |
|--------|------------|----------------|-------------------|------------------------|------------------|----------|
| CTSH   | 15         | Significant    | Replicated         | Yes                    | Yes              | Suggestive |
| DOC2A  | 16         | Significant    | Replicated         | Yes                    | Yes              | Yes      |
| ICA1L  | 2          | Significant    | Replicated         | Yes                    | Yes              | N/A      |
| LACTB  | 15         | Significant    | Replicated         | Yes                    | Yes              | Suggestive |
| SNX32  | 11         | Significant    | Replicated         | Yes                    | Yes              | No       |
| ACE    | 17         | Significant    | Replicated         | Yes                    | Yes              | Yes      |
| RTFDC1 | 20         | Significant    | Replicated         | Yes                    | Yes              | No       |
| CARHSP1| 16         | Significant    | Replicated         | Yes                    | Yes              | Yes      |
| STX6   | 1          | Significant    | Replicated         | Yes                    | Yes              | Yes      |
| STX4*  | 16         | Significant    | Replicated         | Yes                    | Yes              | Yes      |
| PVR*   | 19         | Significant    | Replicated         | Yes                    | Yes              | N/A      |
The small overlap with biometric traits is not surprising given their estimates of genetic correlation with AD\(^1\). These results suggest the specificity of our AD PWAS findings.

Given the central dogma of molecular biology that DNA is transcribed into messenger RNA, which is translated into protein, we asked whether the identified 11 genes with evidence for being causal in AD at the protein level had similar evidence at the transcript level. We integrated the AD GWAS results\(^1\) with 888 human brain transcriptomes to perform a transcriptome-wide association study (TWAS) of AD using FUSION\(^2\). The 888 transcriptomes were mainly from the frontal cortex donated by participants of European ancestry (Supplementary Table 1c) and quality control was analogous to that of the proteomes to remove technical and clinical characteristics before estimating the effect of genetic variants on mRNA expression. Among the 13,650 transcripts after quality control, 6,670 were heritable. The AD TWAS identified 40 genes whose genetically regulated mRNA expression levels were associated with AD at an FDR of \(P < 0.05\) (Extended Data Fig. 3 and Supplementary Table 17). Among the 11 potentially causal genes identified at the protein level, 5 genes, \(ACE\), \(CARHSP1\), \(SNX32\), \(STX4\) and \(STX6\) showed at least nominal significance with similar directions of our PW AS findings are not simply the result of correlated protein levels, each belonged to a different module, which implies that proteins belonged to one of these modules while 5 did not. For these 6 proteins, only \(CTSH\) had evidence to suggest protein expression is mediated by mRNA expression level (Supplementary Table 18a,b). In summary, about half (5 of 11) of the genes with evidence consistent with being causal in AD at the protein level were also associated with AD at the transcript level.

We previously identified 31 modules of coexpressed proteins in Religious Orders Study/Memory and Aging Project (ROS/ MAP) reference proteomes by using Weighted Gene Co-expression Network Analysis\(^3\). We found that 6 of the 11 potential AD causal proteins belonged to one of these modules while 5 did not. For these 6 proteins, each belonged to a different module, which implies that our PWAS findings are not simply the result of correlated protein expression\(^4\).

Using human single-cell RNA-sequencing (RNA-seq) data profiled from the dPFC\(^5\), we found cell type-specific enrichment for expression of 6 of the 11 causal genes at an FDR of \(P < 0.05\) (adjusted for 17,775 genes). \(DOCA2\), \(ICA1L\), \(PLEKHA1\) and \(SNX32\) were enriched in excitatory neurons, whereas \(CARHSP1\) showed enrichment in oligodendrocytes and \(CTSH\) in astrocytes and microglia (Extended Data Fig. 4 and Supplementary Table 19).

Lastly, 8 of the 11 identified causal genes were not within 1Mb of AD genome-wide-significant sites\(^1\) while 3 were \((LACTB, RTFDC1\) and \(STX4\)), implying that these 8 genes were from new sites. The 8 genes were in regions with suggestive AD associations in GWAS (\(P\) values of \(5.3 \times 10^{-7}\) to \(1.9 \times 10^{-7}\)), which is in line with other TWAS studies\(^6\).

In conclusion, we identified 11 brain proteins that have evidence consistent with being causal in AD for future mechanistic studies to find new treatments for the disease.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-00773-z.

Received: 11 March 2020; Accepted: 21 December 2020; Published online: 28 January 2021

References
1. Jansen, I. E. et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer’s disease risk. Nat. Genet. 51, 404–413 (2019).
2. Kunkle, B. W. et al. Genetic meta-analysis of diagnosed Alzheimer’s disease identifies new risk loci and implicates AB\(_\beta\), tau, immunity and lipid processing. Nat. Genet. 51, 414–430 (2019).
3. Ballard, C. et al. Alzheimer’s disease. Lancet 377, 1019–1031 (2011).
4. Wingo, A. P. et al. Shared proteomic effects of cerebral atherosclerosis and Alzheimer’s disease on the human brain. Nat. Neurosci. 23, 696–700 (2020).
5. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. 48, 245–252 (2016).
6. Beach, T. G. et al. Arizona Study of Aging and Neurodegenerative Disorders and Brain and Body Donation Program. Neuropathology 35, 354–389 (2015).
7. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 10, e1004383 (2014).
8. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. 48, 481–487 (2016).
9. Nicolas, A. et al. Genome-wide analyses identify KIF5A as a novel ALS gene. Neuron 97, 1268–1283.e6 (2018).
10. Nalls, M. A. et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson’s disease: a meta-analysis of genome-wide association studies. Lancet Neurol. 18, 1091–1102 (2019).
11. Nagel, M. et al. Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. Nat. Genet. 50, 920–927 (2018).
12. Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700k00 individuals of European ancestry. Hum. Mol. Genet. 27, 3641–3649 (2018).
13. Pulkki, S. L. et al. Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. Hum. Mol. Genet. 28, 166–174 (2019).
14. Wu, Y. et al. Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. Nat. Commun. 9, 918 (2018).
15. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008).
16. Wainberg, M. et al. Opportunities and challenges for transcriptome-wide association studies. Nat. Genet. 51, 592–599 (2019).
17. Mathys, H. et al. Single-cell transcriptomic analysis of Alzheimer’s disease. Nature 570, 332–337 (2019).
18. Gusev, A. et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. Nat. Genet. 50, 538–548 (2018).
19. Huckins, L. M. et al. Gene expression imputation across multiple brain regions provides insights into schizophrenia risk. Nat. Genet. 51, 659–674 (2019).
20. Raj, T. et al. Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer’s disease susceptibility. Nat. Genet. 50, 1584–1592 (2018).

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

This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2021
Methods

Human brain proteomic and genetic data in the discovery PWAS. We generated human brain proteomes from the dPFC of postmortem brain samples donated by 400 participants of European ancestry from the ROS/MAP13. Participants in the ROS/ MAP studies gave informed consent for longitudinal assessments, agreed to an Anatomic Gift Act and consented to repurposing their data and biospecimens for future studies. The institutional review board of Rush University Medical Center approved the ROS/MAP studies.

We performed proteomic sequencing using isobaric tandem mass tag (TMT) peptide labeling and analyzed these peptides by liquid chromatography coupled to mass spectrometry (MS). Samples were randomized by age, sex, postmortem interval, cognitive diagnosis and pathologies into 50 batches before TMT labeling to minimize batch effects. Peptides from each individual sample (n = 400) and the global internal standard (GIS; n = 100) were labeled using the TMT10plex kit (Thermo Fisher Scientific); high pH fractionation was used to increase peptide deconvolution. Each of the 400 samples of the exactly described previously9 were included in each batch. We used the Proteome Discoverer suite v.2.3 (Thermo Fisher Scientific) and tandem mass spectrometry spectra searched against the canonical UniProtKB human proteome database (February 2019) with 20,338 total sequences to assign peptide spectral matches. Peptide spectral matches were filtered using percolator to an FDR of less than 1%; after spectral assignment, peptides were collated into proteins such that the combined probabilities of their constituent peptides achieved an FDR of less than 1%. Peptides shared among multiple proteins were assigned based on parsimony. Integration of ion quantification from tandem mass spectrometry or three-stage mass spectrometry scans with a tolerance of 20 parts per million at the most confident centroid setting was used to quantify reporter ions. After protein quantification, we identified proteins that were not reliably measured using the two GIS that were run in each batch. Proteins whose measurements fell outside the 95% confidence interval of the GIS for any batch were removed from further analysis. Proteomic analysis identified 12,691 proteins; after we excluded proteins with missing values in more than 50% of the 400 subjects, 8,356 proteins remained. To remove the effects of protein loading differences, we scaled each protein abundance with a sample-specific total protein abundance and log-transformed the abundance. Next, we identified and removed poorly performing samples using iterative principal component analysis to remove samples with greater than four standard deviations from the mean of either the first or second principal component. Subsequently, regression was used to estimate and remove the effects of proteomic sequencing batch, mass spectrometry reporter quantification mode, sex, age at death, postmortem interval, study (ROS versus MAP) and the final clinical diagnosis of cognitive status from the proteomic profile. Expanded details on the proteomic sequencing and quality control are published in Wing et al.

Genotyping was obtained from either whole-genome sequencing or genome-wide genotyping by either the Illumina OmniQuad Express or Affymetrix GeneChip 6.0 platforms as described in De Jager et al.23. Quality control of genotyping from either source was performed separately using Plink v.1.99b14. Whole-genome sequencing data were prefered over array-based genotyping in cases where individuals had genotyping data from both sources. Individuals with an overall genotyping missingness >5% were excluded. Variants were excluded if they had evidence of deviation from Hardy–Weinberg equilibrium (P < 1 × 10−5), missing genotype rate >5%, minor allele frequency <1% or were not an SNP. Next, we used to remove outliers or by removing proteins with more than 40% missingness greater than second-degree relatives. For array-based data, we imputed genotyping to the 1000 Genome Project Phase 3 (ref. 35) using the Michigan Imputation Server2; SNPs with imputation R2 > 0.3 were retained. Principal component analysis was performed to compare genetic ancestry of these individuals to CEU from the 100 Genomes Project (Extended Data Fig. 5 and Supplementary Table 20). All samples were kept for downstream analyses used only the 1,190,321汉Map SNPs present in the 489 individuals of European ancestry from the 1000 Genomes Project, which was provided by FUSION2 and commonly referred to as the linkage disequilibrium reference panel. After quality control, there were 376 subjects with both proteomic and genetic data for our discovery PWAS.

Human brain proteomic and genetic data in the confirmation PWAS. The confirmation human brain proteomes were profiled from the dPFC of postmortem brain samples from 198 participants of European ancestry recruited by the Banner Sun Health Research Institute. Participants in this study were recruited from the retirement communities in greater Phoenix, Arizona, USA. All enrolled participants or their legal representatives signed an informed consent form, and the study was approved by the institutional review board of Banner Sun Health Research Institute. Participants consented to annual standardized medical, neurological and neuropsychological testing. Research diagnoses were made using approved research guidelines and a final clinico-pathological diagnosis was made after review of all clinical, medical records and assessments agreed to by an independent Anatomic Gift Act and consented to repurposing their data and biospecimens for future studies. The institutional review board of Rush University Medical Center approved the ROS/MAP studies.

We performed proteomic sequencing using isobaric tandem mass tag (TMT) peptide labeling and analyzed these peptides by liquid chromatography coupled to mass spectrometry (MS). Samples were randomized by age, sex, postmortem interval, cognitive diagnosis and pathologies into 50 batches before TMT labeling to minimize batch effects. Peptides from each individual sample (n = 400) and the global internal standard (GIS; n = 100) were labeled using the TMT10plex kit (Thermo Fisher Scientific); high pH fractionation was used to increase peptide deconvolution. Each of the 400 samples of the exactly described previously9 were included in each batch. We used the Proteome Discoverer suite v.2.3 (Thermo Fisher Scientific) and tandem mass spectrometry spectra searched against the canonical UniProtKB human proteome database (February 2019) with 20,338 total sequences to assign peptide spectral matches. Peptide spectral matches were filtered using percolator to an FDR of less than 1%; after spectral assignment, peptides were collated into proteins such that the combined probabilities of their constituent peptides achieved an FDR of less than 1%. Peptides shared among multiple proteins were assigned based on parsimony. Integration of ion quantification from tandem mass spectrometry or three-stage mass spectrometry scans with a tolerance of 20 parts per million at the most confident centroid setting was used to quantify reporter ions. After protein quantification, we identified proteins that were not reliably measured using the two GIS that were run in each batch. Proteins whose measurements fell outside the 95% confidence interval of the GIS for any batch were removed from further analysis. Proteomic analysis identified 12,691 proteins; after we excluded proteins with missing values in more than 50% of the 400 subjects, 8,356 proteins remained. To remove the effects of protein loading differences, we scaled each protein abundance with a sample-specific total protein abundance and log-transformed the abundance. Next, we identified and removed poorly performing samples using iterative principal component analysis to remove samples with greater than four standard deviations from the mean of either the first or second principal component. Subsequently, regression was used to estimate and remove the effects of proteomic sequencing batch, mass spectrometry reporter quantification mode, sex, age at death, postmortem interval, study (ROS versus MAP) and the final clinical diagnosis of cognitive status from the proteomic profile. Expanded details on the proteomic sequencing and quality control are published in Wing et al.

Genotyping was performed using the Affymetrix Precision Medicine Array using DNA extracted from the brain with the QIAGEN GenePure kit. We applied the same approach to quality control as described for the discovery dataset, including removing individuals based on data completeness or relatedness, removing sites with evidence of deviation from Hardy–Weinberg equilibrium, missingness above 5%, minor allele frequency below 1%, or variants that were not SNPs. Genotyping was imputed to the 1000 Genome Project Phase 3 using the Michigan Imputation Server. SNPs with an imputation R2 > 0.3 were retained. Finally, only sites included in the linkage disequilibrium reference panel were used in our confirmation PWAS, as recommended by the FUSION pipeline. After quality control, there were 152 individuals with both proteomic and genetic data to include in our confirmation analyses.

Brain transcriptomic and genetic data in the AD TWAS. The brain transcriptomes were profiled from postmortem brain samples donated by 783 individuals of European ancestry recruited by the ROS/MAP, Mayo and Mount Sinai Brain Bank studies20,21. Brain transcriptomes were also profiled from the dPFC and also from the frontal cortex, temporal cortex, inferior frontal gyrus, superior temporal gyrus and perirhinal gyrus. Details on alignment, quality control and normalization of the RNA-seq data have been described previously25. Briefly, Picard v.1.83 was used to convert BAM files to FASTQ format and STAR v.2.5.1b26 was used to align reads to the GRCh38 reference genome and compute gene expression. For each sample, we removed genes with >0.1% missingness in at least 50% of the samples and genes with missing gene length and percentage guanine–cytosine content. Next, we removed outlier samples. Then, we regressed out effects of batch, sex, postmortem interval, age at death, brain region and final diagnosis of cognitive status from the transcriptomic profiles before estimating minor weights.

For individuals with transcriptomic data, their genome-wide genotyping was generated as described previously22,23. Quality control of the genotyping data was performed as described above for the discovery ROS/MAP dataset. After quality control, there were 13,650 mRNAs quantified from 783 individuals using 888 transcriptomes. Genotyping was filtered to include only sites in the linkage disequilibrium reference panel provided by FUSION before estimating mRNA weights as described below.

AD GWAS summary association statistics. We used the summary association statistics from the latest GWAS of AD by Jansen et al.13,14, which had 455,258 participants of European ancestry, most of whom were from the UK Biobank.

Statistical approach. We used FUSION to estimate protein weights in the discovery and confirmation datasets, separately. For simplicity, we described the process for the discovery dataset and followed the same steps for the confirmation dataset. As mentioned above, we subset ROS/MAP genome-wide genotyping into an SNP-exome panel of 1,190,321 SNPs estimated using FUSION to minimize the influence of linkage disequilibrium on the estimated test statistics2. Next, the SNP-based heritability for each gene was estimated using the discovery proteomic and genetic data. For proteins with significant heritability (that is, heritability P < 0.01), we used FUSION to compute the effect of SNPs on protein abundance using multiple predictive models—top1, blup, lasso, enet and balm1. Protein weights from the most predictive model were selected. Subsequently, we used FUSION to combine the genetic effect of AD (AD GWAS z-score) with the protein weights by calculating the linear sum of z-score × weight for the independent SNPs at the locus to perform the PWAS of AD. Lambda (λz) and lambda 1,000 (λada), which is a standardized estimate of genetic influence scaled to a study of 1,000 cases and 1,000 controls27–29, were calculated for each PWAS. Lambda 1,000 was calculated using the following formula27–29: λz,000 = 1 + (λz − 1) × (1/1,000cases + 1/1,000controls)/(1/1,000cases + 1/1,000controls). They were found to be consistent with other studies using FUSION that calculated lambda2 (Extended Data Fig. 1). The slightly higher λz in the confirmation PWAS may reflect some difference in the heterogeneity of the datasets.

For the transcriptomic data, we calculated the transcript weights using FUSION with two modifications to accommodate individuals with transcriptomic profiles from more than one brain region. First, the flag --scale 1 was added to handle prescaled expression values. Second, the family ID in the Plink FAM file was used to ensure that samples from the same individual were always in the same fold within cross-validated records and that no fold differed by more than 5% in size from any other fold. RNA weights were estimated using all five models and the most predictive model was used. Next, we used FUSION to combine the genetic effect of AD (AD GWAS z-score) with the mRNA expression weights to perform the TWAS of AD.
For the colocalization test, we used COLOC v.3.2 software2 to estimate the posterior probability of the protein and AD sharing a causal variant, as well as the posterior probability of the protein and AD not sharing a causal variant, using the marginal association statistics. For SMR, SMR v.1.02 software7 was used to test whether the AD PWAS-significant genes (from FUSION) were associated with AD via their cis-regulated brain protein expression. We used PLink8 to estimate pQTLs in the discovery proteomic dataset by linear regression. Then, we applied SMR to the pQTL results and the AD GWAS summary statistics. We used the conservative unadjusted P < 0.05 from HEIDI to declare that presence of linkage likely influences the main SMR findings. For genes with both mRNA and protein abundance associated with AD, we applied SMR for two molecular traits7 to the expression quantitative trait loci summary statistics from Sieberts et al.22 and the pQTL summary statistics described above to determine if mRNA mediates the influence of SNP on protein.

We examined the cell type-specific expression of the 11 genes with evidence for a causal role in AD at the brain protein level by using human brain single-cell RNA-seq data profiled from the dPFC from Mathys et al.17. First, we performed data preprocessing and transformation on the raw single-cell RNA-seq data using the Seurat v.3.1.2 package23. We removed genes with fewer than 3 counts in a cell and cells with unique feature counts over 2,500 or less than 200. The RNA counts were then normalized and scaled using the NormalizeData and ScaleData functions. The RNA-seq data had 17,926 genes in 70,634 cells before and 17,775 genes in 53,083 cells after quality control and normalization. We focused on the five main cell types—excitatory neuron, inhibitory neuron, astrocyte, microglia and oligodendrocyte. For the 11 potential AD causal genes, we performed differential expression analysis to compare their expression levels in one cell type versus the rest of the other cell types to determine if they were highly expressed in a particular cell type. Multiple testing correction applied to this analysis was corrected for all 17,775 genes. To determine the novelty of the genes identified in the discovery PWAS, we asked whether each gene was within a 1-Mb window of the 2,358 significant AD GWAS sites (P < 5 × 10−8) that corresponded to the 29 independent risk loci.

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

**Data availability** Phenotypic, proteomic and transcriptomic data used in this manuscript are available via the AD Knowledge Portal (https://adknowledgeportal.org). The AD Knowledge Portal is a platform for accessing data, analyses and tools generated by the Accelerating Medicines Partnership (AMP-AD) Target Discovery Program and other National Institute on Aging (NIA)-supported programs to enable open-sciences practices and accelerate translational learning. The data, analyses and tools are shared early in the research cycle without a publication embargo on secondary use. Data are available for general research use according to the following requirements for data access and data attribution (https://adknowledgeportal.org/DataAccess/Instructions). Results of the pQTL analysis, protein weights and transcript weights described in this manuscript can be accessed at https://doi.org/10.7303/syn23627957.

**References**

21. Bennett, D. A. et al. Religious Orders Study and Rush Memory and Aging Project. *J. Alzheimers Dis.* 64, S161–S189 (2018).
22. Mertins, P. et al. Reproducible workflow for multiplexed deep-scale proteome and phosphoproteome analysis of tumor tissues by liquid chromatography–mass spectrometry. *Nat. Protoc.* 13, 1632–1661 (2018).
23. De Jager, P. L. et al. A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Sci. Data* 5, 180142 (2018).
24. Purcell, S. et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* 81, 559–575 (2007).
25. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* 26, 2867–2873 (2010).
26. Abecasis, G. R. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65 (2012).
27. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* 48, 1284–1287 (2016).
28. Allen, M. et al. Human whole genome genotype and transcriptome data for Alzheimer’s and other neurodegenerative diseases. *Sci. Data* 3, 160089 (2016).
29. Wang, M. et al. The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer’s disease. *Sci. Data* 5, 180185 (2018).
30. Wan, Y. W. et al. Meta-analysis of the alzheimer’s disease human brain transcriptome and functional dissection in mouse models. *Cell Rep.* 32, 107908 (2020).
31. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21 (2013).
32. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* 55, 997–1004 (1999).
33. Freedman, M. L. et al. Assessing the impact of population stratification on genetic association studies. *Nat. Genet.* 36, 388–393 (2004).
34. Wu, L. et al. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nat. Genet.* 50, 968–978 (2018).
35. Sieberts, S. K. et al. Large eQTL meta-analysis reveals differing patterns across cerebral cortical and cerebellar brain regions. *Sci. Data* 7, 340 (2020).
36. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420 (2018).

**Acknowledgements** We thank the participants of the ROS, MAP, Mayo, Mount Sinai Brain Bank and Banner Sun Health Research Institute Brain and Body Donation Program for their time and participation. The following National Institutes of Health (NIH) grants supported this work: P30 AG086611 (A.L.I.); P50 AG101611 (D.A.B.); R01 AG055688 (A.L.I.); R01 AG158319 (D.A.B.); R01 AG107917 (D.A.B.); R01 AG053960 (N.T.S.); R01 AG056533 (T.S.W., A.P.W.); R01 AG057911 (N.T.S.); R01 AG061800 (N.T.S.); R56 AG060757 (T.S.W.); R56 AG062256 (T.S.W.); RC2 AG036457 (D.A.B.); RF1 AG057470 (T.S.W.); U01 AG046415 (P.L.D.); U01 AG046161 (A.L.I.); U01 AG061356 (P.L.D.); U01 AG061357 (A.L.I.); and U01 MH15484 (A.P.W.). NIH grants include those that supported the Accelerating Medicine Partnership for AD, the National Institute of Neurological Disorders and Stroke Emory Neuroscience Core and Goizueta Alzheimer’s Disease Research Center (ADRC) at Emory University, the Rush University ADRC and Arizona State University ADRC that made this work possible. The following Veterans Administration grants supported this work: I01 BX003853 (A.P.W.) and IK4 BX005219 (A.P.W.). The Brain and Body Donation Program has been supported by the NIH, the Arizona Department of Health Services, the Arizona Biomedical Research Commission and the Michael J. Fox Foundation for Parkinson’s Research. Additional support includes grants from the Alzheimer’s Association (N.T.S.), Alzheimer’s Research UK (N.T.S.), the Michael J. Fox Foundation for Parkinson’s Research (N.T.S.) and the Weston Brain Institute Biomarkers Across Neurodegenerative Diseases Grant 11060 (N.T.S.). The views expressed in this work do not necessarily represent the views of the Veterans Administration or the United States Government.

**Author contributions** A.P.W. and T.S.W. conceptualized and designed the study. A.P.W., D.M.D., E.B.D., T.G.B., E.M.R., P.L.D., J.J.L., D.A.B., N.T.S., A.I.L. and T.S.W. acquired the data. A.P.W., Y.L., E.S.G., J.G., B.A.L. and T.S.W. conducted the analyses. A.P.W., Y.L., E.S.G., J.G., B.A.L., E.B.D., C.R., M.P.E., J.J.L., D.A.B., N.T.S., A.I.L. and T.S.W. interpreted the data. A.P.W., Y.L., E.S.G., J.G., B.A.L., M.P.E., J.J.L., D.A.B., N.T.S., A.I.L. and T.S.W. wrote the first draft of the manuscript. All authors critically revised and reviewed the manuscript.

**Competing interests** The authors declare no competing interests.

**Additional information**

Extended data is available for this paper at https://doi.org/10.1038/s41588-020-00773-z.

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41588-020-00773-z.

**Correspondence and requests for materials** should be addressed to A.P.W. or T.S.W.

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

**Reprints and permissions information** is available at www.nature.com/reprints.
Extended Data Fig. 1 | Quantile-quantile plots for the discovery and replication PWAS of AD. Quantile-quantile plot for a, the discovery PWAS of AD ($\lambda = 1.36; \lambda_{1000} = 1.003$) and b, confirmatory PWAS of AD ($\lambda = 1.39; \lambda_{1000} = 1.003$).
Extended Data Fig. 2 | Overlap of significant genes between AD and other traits. Overlap between results of the AD PWAS and PWAS for other traits. All the PWAS used the discovery ROS/MAP proteomic dataset (n = 376) and GWAS summary results from Caucasian individuals. The following outcomes were tested: clinical AD GWAS (N = 63,926), amyotrophic lateral sclerosis (ALS; N = 80,610), body mass index (BMI; N = 681,275), height (N = 693,529), neuroticism (N = 390,278), Parkinson’s disease (PD; N = 1,474,097), and waist-to-hip ratio adjusting for BMI (WHRadjBMI; N = 694,649). Significant genes considered for overlap are those with FDR p < 0.05.
Extended Data Fig. 3 | Quantile-quantile plot for the TWAS of AD. Quantile-quantile plot for the TWAS of AD ($\lambda = 1.22; \lambda_{1000} = 1.002$).
Extended Data Fig. 4 | Single cell-type expression. Single-cell type expression for AD PWAS-significant genes with evidence of causality in AD. Using human brain single-cell RNA-sequencing data profiled from the dPFC, we found that 6 genes (of the 11 genes) had evidence of enrichment in a cell type at FDR p < 0.05. Enrichment testing was performed using Wilcoxon rank sum test, as implemented by the Seurat package, and multiple testing was accounted for by FDR adjusted for 17,775 tested genes. CARHSP1 showed enrichment in oligodendrocytes. CTSR showed enrichment in astrocytes and microglia. DOC2A, ICA1L, PLEKHA1, and SNX32 were enriched in excitatory neurons.
Extended Data Fig. 5 | Genetic principal components of genetic ancestry for each dataset. Genetic principal components of genetic ancestry for each dataset. The first two genetic principal components for individuals in each dataset are plotted (grey boxes) with individuals from the 1000 Genomes CEU dataset (purple triangles) for a, the discovery proteomic dataset, b, the replication proteomic dataset, and c, the transcriptomic dataset.
Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection
An Orbitrap Fusion mass spectrometer (ThermoFischer Scientific) was used to collect raw proteomic files using built-in software.

Data analysis
Proteome Discoverer suite (version 2.3 ThermoFisher Scientific) was used to analyze raw proteomic data. All statistical analyses were performed using R version 3.5.1 (2018-07-02) with the following R packages: ‘dplyr’ version 0.8.0.1, ‘ggplot2’ version 3.1.0, ‘grid’ version 3.5.1, ‘readxl’ version 1.3.0, ‘stringi’ version 1.3.1, ‘ggrepel’ version 0.8.0, ‘reshape2’ version 3.5.1, and ‘ggcorrplot’ version 0.1.2. Plink version v1.99b was used. KING version 2.2 was used. Picard version 1.83 was used. STAR version 2.5.1b was used. FUSION commit c677be0 from https://github.com/gusevlab/fusion_twas was used. COLOC version 3.2 was used. SMR version 1.02 was used. Seurat version 3.1.2 was used.

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Phenotypic, proteomic, and transcriptomic data used in this manuscript are available via the AD Knowledge Portal (https://adknowledgeportal.org). The AD Knowledge Portal is a platform for accessing data, analyses, and tools generated by the Accelerating Medicines Partnership (AMP-AD) Target Discovery Program and other National Institute on Aging (NIA)-supported programs to enable open-science practices and accelerate translational learning. The data, analyses and tools are shared early in the research cycle without a publication embargo on secondary use. Data is available for general research use according to the following
Field-specific reporting

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

- **Life sciences**
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](http://nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

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

**Sample size**

No sample size calculations were used. All available samples with brain proteomic and phenotypic data were used for the analysis.

**Data exclusions**

There was no pre-established exclusion criterion. A total of 9 individuals were removed from the discovery dataset (ROS/MAP cohorts) proteomic analysis. This was done through an iterative process of detecting outliers by principal component analysis of the proteomic data and excluding all individuals who were greater than 4 standard deviations from the mean of the first two principal components. A total of 3 rounds of principal component analysis was used until no further outliers were identified. Then, of these 391 samples, 376 had both proteomic and genome-wide genotyping data for the PWAS.

**Replication**

The independent Banner dataset was used for confirmation since this dataset included individuals with both deep brain proteomic and genome-wide genetic data. 152 samples had both proteomic and genome-wide genetic data and were used for the analysis. Replication was successful.

**Randomization**

Both the ROS/MAP and Banner samples were randomized by age, sex, PMI, cognitive diagnosis, and pathology outcomes into batches of 8 samples for proteomic analysis to minimize the batch effects.

**Blinding**

The individuals preparing samples for proteomic sequencing for the discovery (ROS/MAP) and confirmation (Banner) datasets were blinded to all phenotypic information. This is not a clinical trial so blinding is not applicable to the study.

Reporting for specific materials, systems and methods

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

| Materials & experimental systems | Methods |
|-------------------------------|---------|
| n/a  | Involved in the study | Involved in the study |
| ☒ | Antibodies | ☒ | ChIP-seq |
| ☒ | Eukaryotic cell lines | ☒ | Flow cytometry |
| ☒ | Palaeontology | ☒ | MRI-based neuroimaging |
| ☒ | Animals and other organisms | ☒ | |
| ☒ | Human research participants | ☒ | |
| ☒ | Clinical data | ☒ | |

Human research participants

Policy information about [studies involving human research participants](#)

**Population characteristics**

The discovery reference proteomic dataset (ROS/MAP cohorts) is community-based cohorts focused on understanding aging and age-related cognitive changes. The confirmation proteomic dataset (Banner Sun Health Research Institute) includes participants from the retirement communities in Phoenix, AZ. Participants in these cohorts were older adults, with a higher percentage of females, and had either normal cognition, mild cognitive impairment, or dementia at death.

**Recruitment**

The ROS study recruits older Catholic priests, nuns, and monks throughout the USA. The MAP study recruits older lay persons from the greater Chicago area. Both studies involve detailed annual cognitive and clinical evaluations and brain autopsy. Participants provided informed consent, signed an Anatomic Gift Act, and a repository consent to allow their data and biospecimens to be repurposed. Likewise, the Banner Sun Health Institute study recruits participants from the retirement communities in Phoenix, AZ and performs standardized medical, neurological, and neuropsychological testing annually. Participants were recruited from the communities and there is no known selection bias.

**Ethics oversight**

The ROS/MAP studies were approved by an Institutional Review Board of Rush University Medical Center. The Banner Sun Health Institute study was approved by its Institutional Review Board.
Note that full information on the approval of the study protocol must also be provided in the manuscript.