ANALYSIS
https://doi.org/10.1038/s41588-020-00776-w

nature genetics

Check for updates

Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes

Jeremy Schwartzentruber[1,2,3]✉, Sarah Cooper[2,3], Jimmy Z. Liu[4], Inigo Barrio-Hernandez[1,2], Erica Bello[2,3], Natsuhiko Kumasaka[3], Adam M. H. Young[5], Robin J. M. Franklin[5], Toby Johnson[6], Karol Estrada[7], Daniel J. Gaffney[2,3,8], Pedro Beltrao[1,2] and Andrew Bassett[2,3]✉

Genome-wide association studies have discovered numerous genomic loci associated with Alzheimer's disease (AD); yet the causal genes and variants are incompletely identified. We performed an updated genome-wide AD meta-analysis, which identified 37 risk loci, including new associations near *CCDC6*, *TSPAN14*, *NCK2* and *SPRED2*. Using three SNP-level fine-mapping methods, we identified 21 SNPs with >50% probability each of being causally involved in AD risk and others strongly suggested by functional annotation. We followed this with colocalization analyses across 109 gene expression quantitative trait loci datasets and prioritization of genes by using protein interaction networks and tissue-specific expression. Combining this information into a quantitative score, we found that evidence converged on likely causal genes, including the above four genes, and those at previously discovered AD loci, including *BIN1*, *APH1B*, *PTK2B*, *PILRA* and *CASS4*.

Genome-wide association studies (GWAS) for family history of disease, known as GWAS by proxy (GWAX), are a powerful method for performing genetic discovery in large, unselected cohort biobanks, particularly for age-related diseases[1]. Recent meta-analyses have combined the GWAS of diagnosed late-onset AD with GWAX for a family history of AD in the UK Biobank (UKB)[2,3] and reported 12 new disease-associated genomic loci. However, the causal genetic variants and genes that influence AD risk at these and previously discovered loci have been clearly identified in only a few cases. Discovering causal variants has led to deeper insight into the molecular mechanisms of multiple diseases, including obesity[4], schizophrenia[5] and inflammatory bowel disease[6]. For AD, known causal variants include the $\varepsilon 4$ haplotype in *APOE*, the strongest genetic risk factor for late-onset AD, and a common nonsynonymous variant that strongly alters splicing of *CD33* exon 2 (ref. [7]). Likely causal rare nonsynonymous variants have also been discovered in *TREM2* (ref. [8]), *PLCG2* and *ABI3* (ref. [9]). These findings have strengthened support for a causal role of microglial activation in AD.

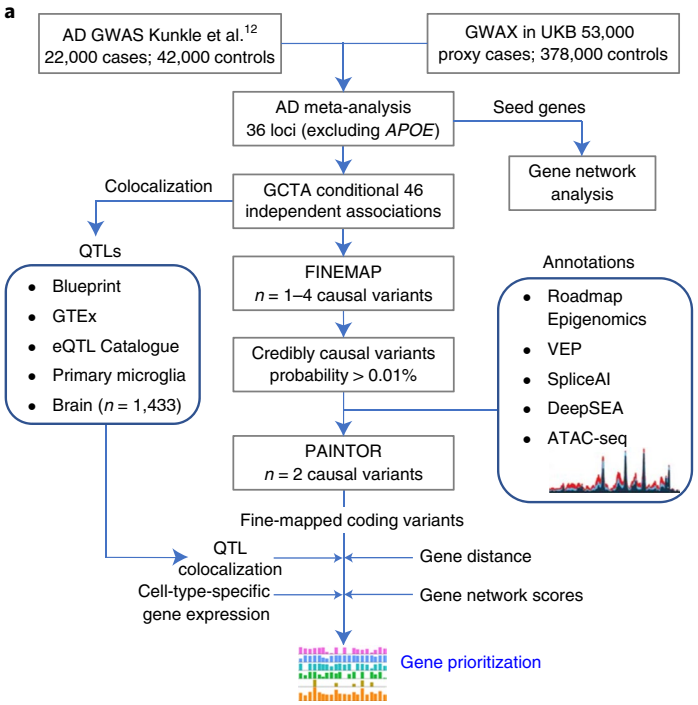
Although nonsynonymous variants are highly enriched in trait associations, most human trait-associated variants do not alter protein-coding sequences and are thought to mediate their effects via altered gene expression, which probably occurs in a cell-type-dependent manner. A growing number of studies have mapped genetic variants affecting gene expression, known as expression quantitative trait loci (eQTLs), in diverse tissues or sorted cell types[10,11]. While it is common to integrate GWAS results with eQTLs, this is often limited to a small number of datasets thought to be relevant.

To identify the putative causal genetic variants for AD, we performed a meta-analysis of GWAX in the UKB with the latest GWAS for diagnosed AD[12], followed by fine-mapping using three alternative methods. Notably, this updated GWAS tested more genetic variants than the Lambert et al. study[13] used in meta-analyses by Jansen et al.[3] and Marioni et al.[2] (11.5 versus 7.1 million). The increased power from our meta-analysis revealed four additional AD risk loci, and the higher density genotype imputation identified new candidate causal variants at both new and established loci. We also performed statistical colocalization analyses with a broad collection of eQTL datasets, including a recent study on primary microglia[14], to identify candidate genes mediating risk at AD loci. We found that multiple lines of evidence, including colocalization, tissue- or cell-type-specific expression and information propagation in gene networks, converge on a set of likely causal AD genes.

Results

Meta-analysis reveals 37 loci associated with AD risk. We performed a GWAX in the UKB for family history of AD, based on 53,042 unique individuals who were either diagnosed with AD or who reported a parent or sibling having dementia, and 355,900 controls. This identified 13 risk loci ($P < 5 \times 10^{-8}$), 10 of which have been reported previously. Three new loci were located near *NCK2*, *PRL* and *FAM135B*. Notably, *PRL* has been reported as a cerebrospinal fluid biomarker of AD[15]. We next performed a fixed-effects meta-analysis of these GWAX results with the Kunkle et al.[12] stage 1 GWAS meta-analysis of 21,982 cases with diagnosed AD and 41,944 controls across 10,687,126 overlapping variants (Fig. 1). This revealed 34 AD risk loci ($P < 5 \times 10^{-8}$), 22 of which were reported in Kunkle et al.[12], while 8 others were reported in either Jansen et al.[3] or Marioni et al.[2]. Four loci were new and located near *NCK2*, *TSPAN14*, *SPRED2* and *CCDC6*. Notably, the *PRL* and *FAM135B*

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, UK. [2]Open Targets, Wellcome Genome Campus, Cambridge, UK. [3]Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, UK. [4]Biogen, Cambridge, MA, USA. [5]Wellcome-Medical Research Council Cambridge Stem Cell Institute, Cambridge Biomedical Campus, University of Cambridge, Cambridge, UK. [6]Target Sciences-R&D, GSK Medicines Research Centre, Stevenage, UK. [7]BioMarin Pharmaceutical, San Rafael, CA, USA. [8]Genomics Plc, Oxford, UK. ✉e-mail: jeremys@ebi.ac.uk; ab42@sanger.ac.uk



c

Locus	Number of signals	Lead *P* value
APOE	See note	approximately 0
TREM2	3	1.83×10^{-23}
BIN1	2	1.10×10^{-54}
PTK2B-CLU	2	7.71×10^{-26}
ABCA7	2	2.41×10^{-13}
NCK2	2	1.28×10^{-12}
EPHA1	2	9.63×10^{-12}
ADAM10	2	2.67×10^{-11}
ACE	2	1.21×10^{-8}
APP-ADAMTS1	2	3.09×10^{-8}
*PLCG2**	1	5.46×10^{-8}
CLNK	1	2.59×10^{-11}
SCIMP	1	1.35×10^{-9}
SPPL2A	1	1.74×10^{-9}
TSPAN14	1	2.74×10^{-9}
VKORC1	1	4.47×10^{-9}
APH1B	1	1.05×10^{-8}
SPRED2	1	1.54×10^{-8}
CCDC6	1	3.80×10^{-8}
ADAMTS4	1	4.30×10^{-8}
TMEM163	1	5.24×10^{-8}
IKZF1	1	7.68×10^{-8}
TSPOAP1	1	3.07×10^{-7}

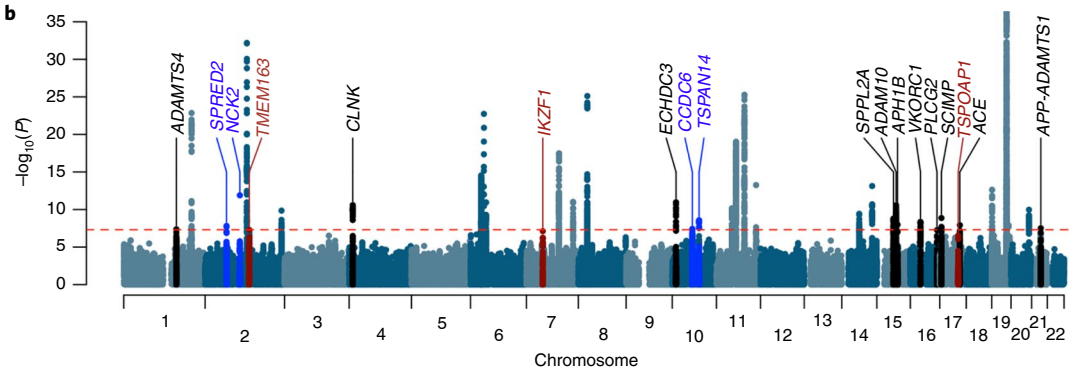


Fig. 1 | Analysis overview. a, Summary of the AD meta-analysis and data processing steps. **b**, Manhattan plot of the meta-analysis of GWAS for diagnosed AD and our GWAX in the UKB. New genome-wide-significant loci are labeled in blue, subthreshold loci in red and recently discovered loci[2,3,12] replicated in our analysis in black. **c**, Number of independent signals at each locus that is either recently discovered or that has more than one signal, as well as the meta-analysis *P* value for the lead SNP at the locus. The *PLCG2* locus was significant ($^*P < 5 \times 10^{-8}$) when including Kunkle et al.[12] stage 3 SNPs. Conditional analyses were not done at *APOE* due to the strength of the signal (Methods).

regions showed no evidence of association in Kunkle et al.[12] ($P > 0.1$) and hence were not significant in the meta-analysis. We included 37 loci in our follow-up analyses, which included three loci found at suggestive significance ($P < 5 \times 10^{-7}$) near *IKZF1*, *TSPOAP1* and *TMEM163* (Fig. 1 and Supplementary Table 1). Linkage disequilibrium (LD) score regression[16] showed that most of the inflation in summary statistics was due to the polygenicity of AD rather than confounding by population structure (lambda_GC = 1.140, intercept = 1.0285 with s.e. = 0.0069; Supplementary Table 2). Of our 37 loci, 16 were nominally replicated ($P < 0.05$) in either the GR@ACE study[17] (4,120 probable AD cases and 3,289 controls) or the FinnGen Biobank v.3 (3,697 cases and 131,941 controls) (Supplementary Table 3). Among our 4 new loci, only *TSPAN14* replicated with $P < 0.05$ (in FinnGen), although power was limited in these smaller datasets (estimated at 28–76%) and most of the alleles had concordant directions of effect. In a meta-analysis with all four datasets, support for most loci was strengthened (Extended Data Fig. 1), including the new loci *TSPAN14*, *CCDC6* and *NCK2*, but was weakened for *SPRED2* (meta-analysis $P = 1.3 \times 10^{-7}$). Although not included in the downstream analyses, four new loci were genome-wide significant, near *GRN*, *IGHG1*, *SHARPIN* and *SIGLEC11* (Supplementary Table 3).

Next, we applied stepwise conditioning using genome-wide complex trait analysis (GCTA)[18], with LD determined from UKB samples, to identify independent signals at the discovered loci. Apart from *APOE*, nine loci had two independent signals, while the *TREM2* locus had three signals (Fig. 1c). Interestingly, a number of the loci discovered recently[2,3,12] had multiple signals: *NCK2*, *EPHA1*, *ADAM10*, *ACE* and *APP-ADAMTS1*. To extract insight from both new and established AD GWAS discoveries, we performed comprehensive colocalization, annotation, fine-mapping and network analyses to identify causal genes and variants (Fig. 1a).

Colocalization between AD risk loci and gene expression traits. To identify genes whose expression might be altered by risk variants, we performed statistical colocalization[19] between each of 36

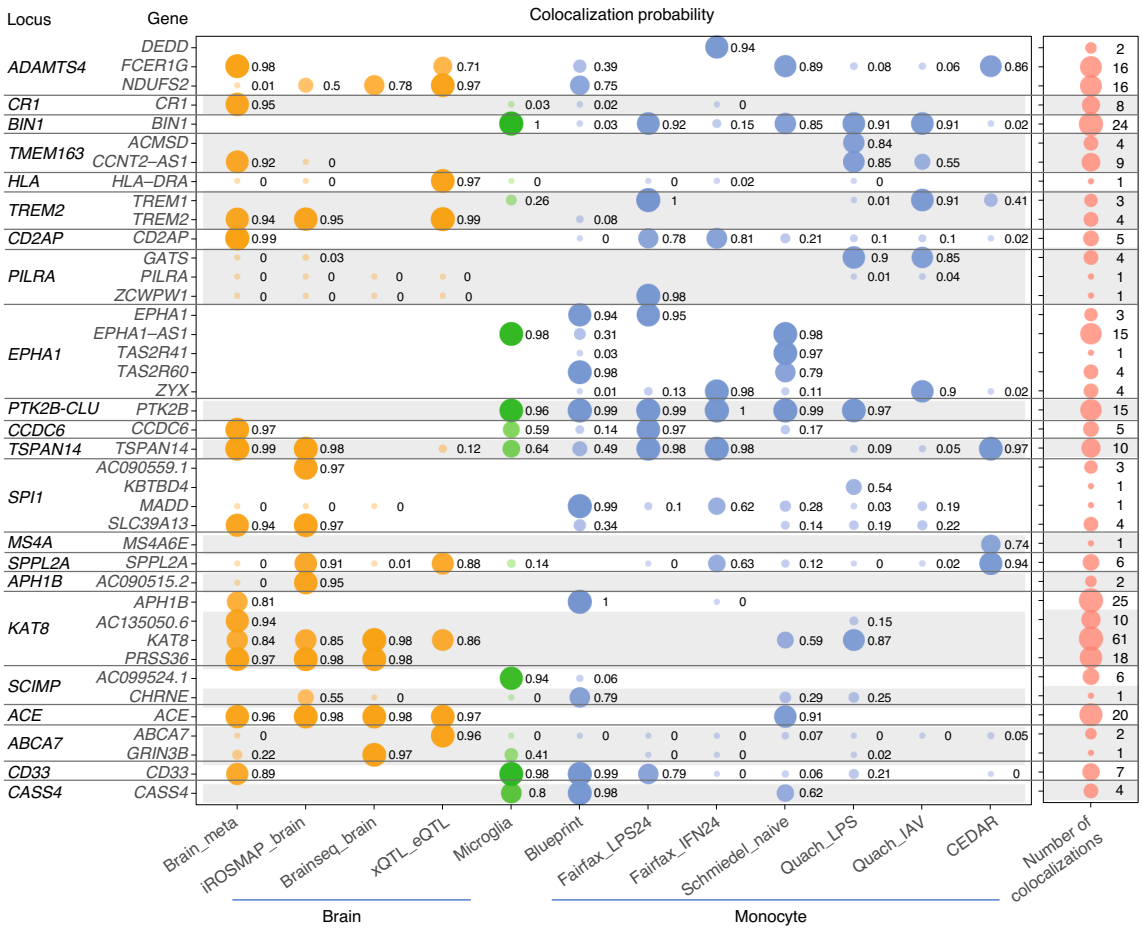


Fig. 2 | Colocalization with eQTLs. For genes with the top overall colocalization scores across AD risk loci, the colocalization probability (hypothesis 4) is shown for selected brain, microglia and monocyte eQTL datasets. For three loci with multiple signals (*BIN1*, *EPHA1* and *PTK2B-CLU*), the maximum score across the conditionally independent signals is shown. The last column shows, for each gene, the number of eQTL datasets with a colocalization probability above 0.8 (Supplementary Tables 5 and 6).

risk loci (excluding *APOE*) and a set of 109 eQTL datasets representing a wide variety of tissues, cell types and conditions (Fig. 2 and Supplementary Table 4). The eQTL datasets include a study of primary microglia from 93 brain surgery donors[14], a meta-analysis of 1,433 brain cortex samples[20], 49 tissues from GTEx[11] and 57 eQTL datasets uniformly reprocessed as part of the eQTL Catalogue[10]. The latter include multiple studies in tissues of potential relevance to AD, such as brain, as well as sorted blood immune cell types under different stimulation conditions[21–37]. For each gene, the colocalization analysis reports the probability that the GWAS and eQTL share a causal variant, referred to as hypothesis 4.

Some studies using colocalization have suggested that there is relatively limited overlap between GWAS associations and eQTLs above that expected by chance[6,38]. A possible reason is that colocalization analyses can have low sensitivity to detect shared causal variants between traits, which could occur for a number of reasons. First, when a locus has multiple causal variants and not all causal effects are shared between the two studies, colocalization may not be detected[19]. Second, if the relevant tissue, cell type or cellular context has not been assayed, then a colocalization may not be found. Third, differences in LD patterns between studies can reduce the likelihood of a positive colocalization. Lastly, low power in either study can further reduce the colocalization probability. To mitigate the first effect, we performed colocalizations separately for each conditionally independent AD signal, to model the case where not all causal variants are shared, as well as for the combined AD signal at each locus. Problems relating to power, LD mismatch or missing the relevant cell type or context are partially mitigated by our use of a large number of highly powered eQTL datasets, which include those with stimulated conditions.

Across the 36 loci, we found 391 colocalizations with at least 80% probability of a shared causal variant between AD and eQTL, representing 80 distinct genes at 27 loci (Supplementary Tables 5 and 6). The genes implicated by colocalization include many that have previously been investigated for roles in AD, such as *PTK2B*[39,40], *BIN1* (refs. [41,42]), *PILRA*[43], *CD33* (refs. [44,45]) and *TREM2* (refs. [46,47]), as well as new candidates including *FCER1G*, *TSPAN14*, *APH1B* and *ACE*. However, the presence of multiple genes with colocalization evidence within individual loci suggests that additional lines of evidence are important for prioritizing relevant genes.

Fine-mapping identifies credibly causal variants. Confirming the causal genes underlying AD risk will ultimately require experiments to identify the molecular mechanisms by which gene function is altered. Such experiments must be motivated by strong hypotheses regarding potentially causal variants and their possible effects. To identify candidate causal variants, we used three distinct fine-mapping methods: single causal variant fine-mapping[48] on each conditionally independent signal; FINEMAP[49], limiting the number of causal variants at each locus to the number of signals

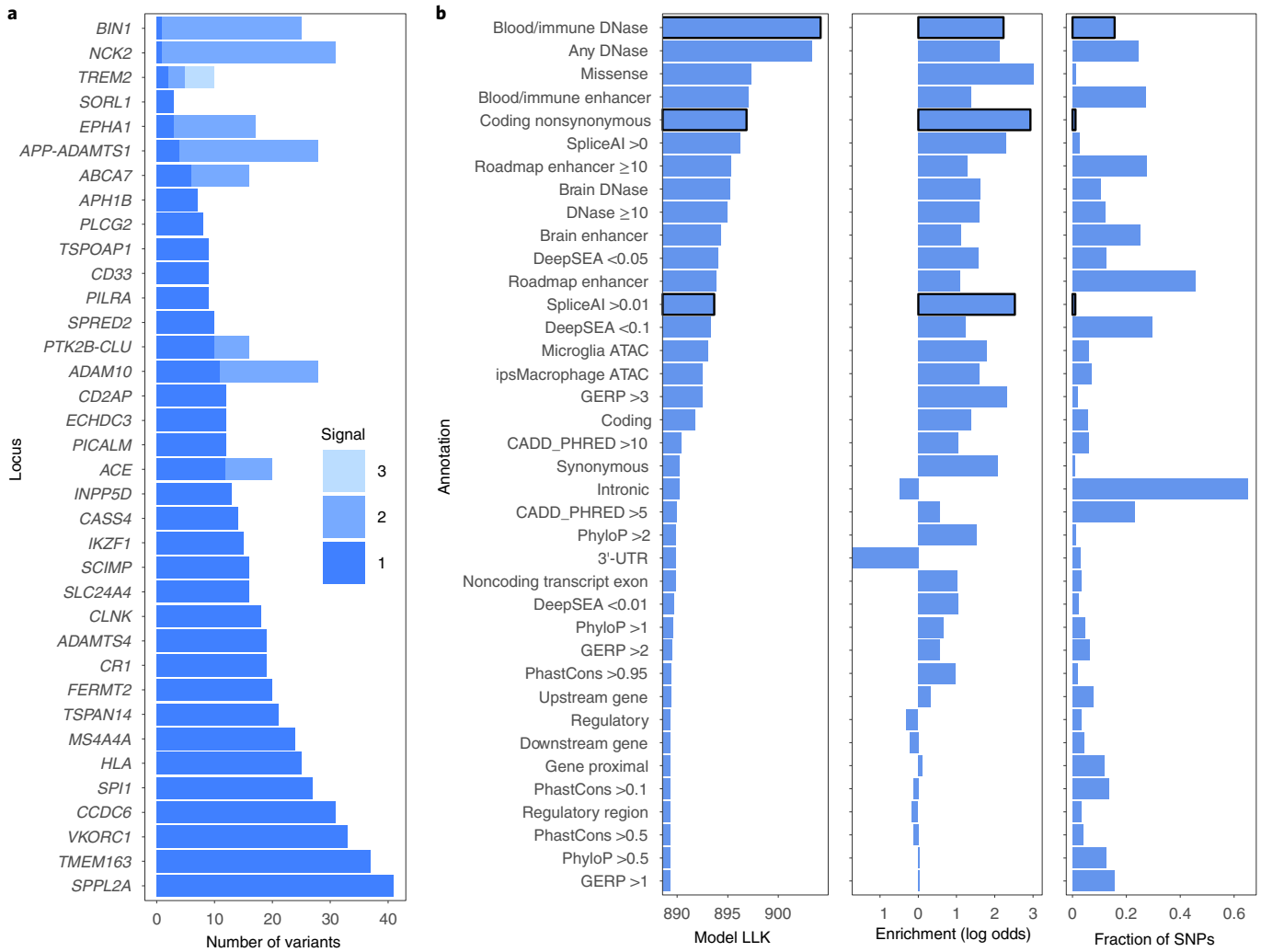


Fig. 3 | Fine-mapping summary. a, Number of variants with mean causal probability >1% for each independent signal. Variant counts for independent signals are shown in different shades of blue. **b**, PAINTOR outputs, showing the LLK of the model for each individual annotation (left); log odds enrichments for individual genomic annotations determined by PAINTOR (middle); fraction of SNPs that are in each annotation (among those selected by a FINEMAP probability >0.01%) (right). Annotations selected for the final model are shown with a black border.

determined by GCTA; and PAINTOR[50], a method that leverages enrichments in functional genomic annotations to improve causal variant identification (Methods).

As a reference panel for our analyses, we used LD computed from UKB participants. Previous work has shown that using reference panels that are either too small or poorly matched can result in spurious fine-mapping signals[51]. For this reason, we conducted a sensitivity analysis (described in the Supplementary Note) by using the same reference panel for conditional analysis and fine-mapping on the non-UKB portion of our meta-analysis (Kunkle et al.[12]). This gave comparable independent signals and SNP probabilities to the full meta-analysis, with the exception of a few loci, namely *ABCA7*, *HLA*, *EPHA1* and *ECHDC3* (Extended Data Fig. 2).

We used 44 annotations individually as input to PAINTOR (Supplementary Table 7); these included assay for transposase-accessible chromatin using sequencing (ATAC-seq) peaks from primary microglia[52] or induced pluripotent stem cell (iPSC)-derived macrophages[29], DNase peaks from the Roadmap Epigenomics Project[53], variant consequence annotations[54] and evolutionary conservation[55] (Fig. 3). We also used scores from DeepSEA[56] and SpliceAI[57], deep learning methods that predict the effects of variants on transcription factor binding or splicing. Missense mutations were the most enriched annotation, with 19.2-fold increased odds of being causal SNPs, but they comprised only 1% of input SNPs. Blood or immune DNase hypersensitivity peaks merged from 24 Roadmap Epigenomics Project tissues provided the highest model likelihood since these peaks covered 16% of SNPs, despite a lower 6.4-fold enrichment. Variants with a nonzero score from SpliceAI, which predicts changes to gene splicing, were also highly enriched (9.3-fold).

We next built a multi-annotation model in PAINTOR following a stepwise selection procedure, which identified a minimal but informative set of three annotations: blood and immune DNase; nonsynonymous coding variants; and variants with a SpliceAI score >0.01. We used probabilities from this PAINTOR model and computed the mean causal probability per variant across the three fine-mapping methods.

There were 21 variants with mean causal probability above 50% across the fine-mapping methods and 79 further variants with probabilities from 10 to 50% (Table 1 and Supplementary Table 8). These include SNPs near established AD risk genes, such as rs6733839 approximately 20 kilobases (kb) upstream of *BIN1*, which has recently been shown to alter a microglial MEF2C-binding site[14] and to regulate *BIN1* expression specifically in microglia[42].

Table 1 | Top candidate variants

Locus	SNP	*P* value	OR	Effect allele	Allele frequency	SNP probability	SpliceAI	DeepSEA	Note	References
ADAMTS4	rs2070902	1.64×10^{-6}	0.949	T	0.2580	0.384	0.107	0.140	Intronic in candidate gene *FCER1G*, with predicted splicing change	–
ADAMTS4	rs4575098	4.30×10^{-8}	1.063	A	0.2350	0.339	–	0.033	3′-UTR of *ADAMTS4*, open chromatin	2
SPRED2	rs268120	2.08×10^{-8}	1.063	A	0.2502	0.556	–	0.033	Strong DNase peak, predicted by DeepSEA to decrease	–
NCK2	rs143080277	1.28×10^{-12}	0.594	T	0.9957	1.000	–	0.086	Enhancer (Roadmap)	–
BIN1	rs6733839	1.10×10^{-54}	1.168	T	0.3915	0.998	–	0.027	Microglia ATAC peak. DeepSEA predicts decreased DNase hypersensitive site	14,40
INPP5D	rs10933431	1.41×10^{-10}	1.080	C	0.7817	0.833	–	0.022	–	80
PILRA	rs1859788	3.28×10^{-18}	0.914	A	0.3206	0.601	0.008	0.041	Known PILRA missense p.Gly78Arg	41
ECHDC3	rs7920721	1.08×10^{-11}	0.935	A	0.6195	0.641	–	0.026	DNase peak. DeepSEA predicts changed binding of USF, Max, Myc	81,82
TSPAN14	rs1870137	2.93×10^{-9}	0.932	C	0.2056	0.097	–	0.007	Top DeepSEA variant, predicting decreased binding of HNF4, FOXA1, SP1	–
TSPAN14	rs1870138	4.51×10^{-9}	0.933	A	0.2057	0.068	–	0.004	Highlighted in text; predicted loss of TAL1 binding	–
SORL1	rs11218343	5.59×10^{-14}	1.205	T	0.9630	1.000	–	0.209	–	83
SORL1	rs2298813	1.52×10^{-4}	1.089	A	0.0470	0.451	0.054	0.003	Secondary association. Missense; also top DeepSEA variant	–
APH1B	rs117618017	1.05×10^{-8}	1.089	T	0.1395	0.895	0.007	0.019	Highlighted in text; missense p.Thr27Ile	84
PLCG2	rs12444183	5.46×10^{-8}	0.948	A	0.3830	0.686	–	0.220	Near promoter of noncoding RNA AC099524.1, with strong microglia colocalization	2
PLCG2	rs72824905	6.35×10^{-6}	1.310	C	0.9924	0.492	0.018	0.006	Secondary association; known missense p.Pro522Arg. Top DeepSEA score	9
TSPOAP1	rs2632516	3.12×10^{-7}	0.952	C	0.4426	0.412	–	0.126	Overlaps ncRNA containing mir-142, important for hematopoietic development	82,85
TSPOAP1	rs2526377	8.45×10^{-7}	1.049	A	0.5579	0.169	–	0.006	Top DeepSEA variant (decreased DNase hypersensitivity) in microglial ATAC peak	86
ACE	rs4311	1.21×10^{-8}	0.947	T	0.4704	0.490	0.126	0.053	Strong predicted splicing change	87,58
ACE	rs3730025	2.58×10^{-7}	0.819	A	0.9828	0.416	0.002	0.021	Secondary association; low-frequency missense p.Tyr244Cys	–

Continued

Table 1 | Top candidate variants (Continued)

Locus	SNP	*P* value	OR	Effect allele	Allele frequency	SNP probability	SpliceAI	DeepSEA	Note	References
ABCA7	rs12151021	2.41×10^{-13}	1.080	A	0.3258	0.713	0.013	0.312	Lead *ABCA7* variant	
ABCA7	rs4147918	7.63×10^{-7}	1.128	A	0.9587	0.552	0.071	0.045	Secondary association; missense p.Gln905Arg; predicted splicing change	59
CD33	rs12459419	2.02×10^{-8}	0.944	T	0.3256	0.662	0.001	0.070	Known missense p.Ala14Val; strong splicing QTL	7
CASS4	rs6014724	1.07×10^{-10}	1.116	A	0.9122	0.548	–	0.083	Lead *CASS4* variant	–
CASS4	rs17462136	1.01×10^{-9}	0.901	C	0.0872	0.067	–	0.001	5′-UTR of *CASS4*; global top DeepSEA variant predicting decreased transcription factor binding	–
ADAMTS1	rs2830489	3.09×10^{-8}	0.943	T	0.2749	0.718	–	0.077	Lead variant near *ADAMTS1*	–

This is a selected list of the most likely causal variants across loci, based on a combination SNP fine-mapping probabilities and annotations. The column 'SNP probability' indicates the mean fine-mapping probability for the SNP; the SpliceAI score is the maximum splicing probability for donor gain/loss or acceptor gain/loss, with nonzero values highly enriched for splicing effects; the DeepSEA functional significance score represents the magnitude of the predicted effect on chromatin features, combined with evolutionary conservation, with values closer to zero representing a stronger predicted change. References for specific SNPs are shown[2,7,9,14,42,43,80–89].

High-confidence variants also include a well-known missense SNP in *PILRA*[43] and a splice-altering missense SNP in *CD33* (ref. [7]). Missense SNP rs4147918 in *ABCA7* had 55% causal probability and *ABCA7* harbored 5 further missense SNPs with probabilities >0.01% at varying allele frequencies. Notably, rs4147918 and 6 other variants within *ABCA7*, including the lead SNP rs12151021, had positive SpliceAI scores. This is consistent with reports of a burden of deleterious variants at *ABCA7* associated with AD[58], as well as potential changes to splicing caused by intronic variable tandem repeats[59].

A number of newly identified AD risk genes had high-confidence fine-mapped variants. These include the *NCK2* rare intronic SNP rs143080277 (>99% probability, minor allele frequency (MAF) = 0.4%), *APH1B* missense SNP rs117618017 (90% probability), rs2830489 near *ADAMTS1* (72% probability) and rs268120 intronic in *SPRED2* (56% probability).

Manual review highlighted a number of candidate causal variants, where the annotation-based SNP probability was higher than that of the other two methods (Fig. 4). Within *TSPAN14*, rs1870137 and rs1870138 reside within a DNase hypersensitivity peak found broadly across tissues, which is also an ATAC peak in microglia (Fig. 4a). Of these, rs1870138 lies at the center of a chromatin immunoprecipitation followed by sequencing (ChIP–seq) peak for binding of multiple transcription factors, including FOS/Jun and GATA1. The AD risk allele rs1870138[G] alters an invariant position of a binding motif for *TAL1*, a gene highly expressed in microglia, which is a binding partner for GATA1. This allele is also associated with increased monocyte count[60] and increased risk for inflammatory bowel disease[61]. Notably, the AD signal in the region colocalizes with both an eQTL and a splicing QTL (sQTL) for *TSPAN14* in multiple datasets, and rs1870138[G] associates with higher *TSPAN14* expression in the brain and in microglia, but with lower expression in some GTEx tissues.

The missense SNP rs117618017 in exon 1 of *APH1B* (p.Thr27Ile) is the likely single causal variant at its locus, with a fine-mapping probability of 90% (Fig. 4b). APH1B is a component of the γ-secretase complex, other members of which (*PSEN1*, *PSEN2*) have rare variants associated with early-onset AD[62]. Interestingly, the AD signal colocalizes with an *APH1B* eQTL in monocytes, neutrophils and T cells, and rs117618017[T] associates with higher AD risk and higher *APH1B* expression across datasets. This allele introduces a motif for the transcriptional regulator YY1 and is predicted by DeepSEA to increase YY1 binding in multiple Encyclopedia of DNA Elements (ENCODE) cell lines. Therefore, it is an open question whether AD risk is mediated by altered APH1B protein structure or altered gene expression.

Finally, the AD association on chromosome 20 colocalizes with an eQTL for *CASS4* in Blueprint monocytes and in GTEx whole blood. While the intronic lead SNP rs6014724 (55% probability) shows no evidence of transcription factor binding in ENCODE data, rs17462136 (7% probability) lies in a region of dense transcription factor binding in the 5′-UTR of *CASS4* (Fig. 4c). The nucleotide position is highly conserved (genomic evolutionary rate profiling (GERP) score = 3.46) and overlaps an ATAC peak in microglia; the rs17462136[C] allele introduces a TEAD1 binding motif. In addition, rs17462136 is more strongly associated with *CASS4* expression in multiple eQTL datasets than is rs6014724.

Network evidence prioritizes genes within and beyond GWAS loci. As a further line of evidence, we developed a method that leverages gene network connectivity to prioritize genes at individual loci. We first constructed a gene interaction network combining information from the STRING, IntAct and BioGRID databases. Next, we nominated 32 candidate AD genes (Supplementary Table 9) based on our other evidence sources as well as literature reports and used these as seed genes similar to the approach used in the priority index for drug discovery[63]. For each locus in turn, we used as input all seed genes except those at the locus, and propagated information through the network with the PageRank algorithm. Thus, the 'networkScore' for a gene represents the degree to which the gene is supported by its interaction with top AD candidate genes at other loci, unbiased by any locus-specific features.

Across AD loci, our selected seed genes were highly enriched for having high network-based gene scores (one-tailed Wilcoxon rank-sum test, $P = 5\times10^{-9}$; Extended Data Fig. 3). At our four new AD loci, the nearest gene (*NCK2*, *TSPAN14*, *SPRED2* and *CCDC6*) in each case was one of the top two highest-scoring genes within 500 kb. Many established or recently discovered AD genes were also

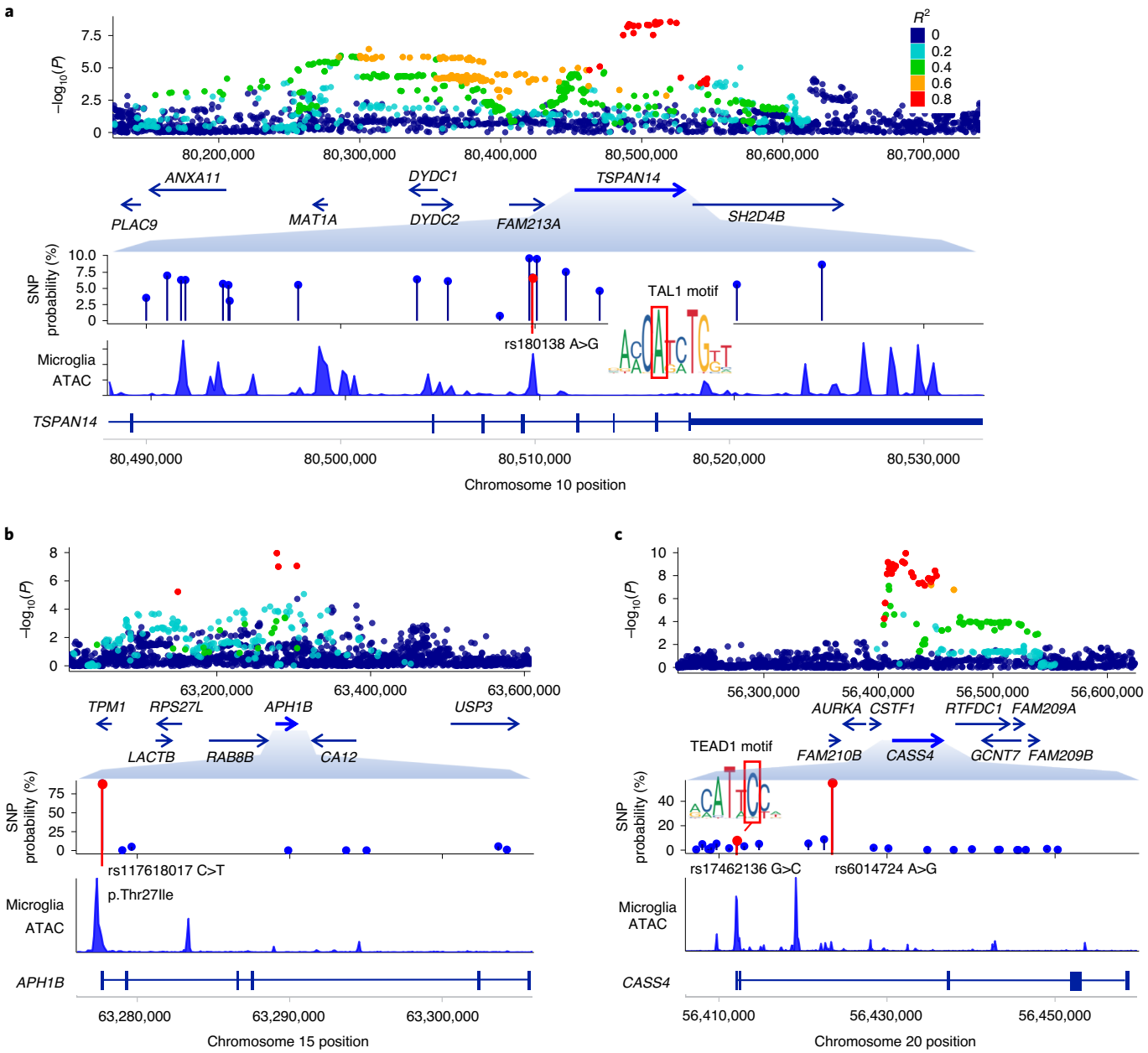


Fig. 4 | Fine-mapped variants. a, SNP rs1870138 in an intron of *TSPAN14* disrupts an invariant position of a TAL1 motif. **b**, Missense SNP rs117618017 in exon 1 of *APH1B*. **c**, SNP rs17462136 in the 5′-UTR of *CASS4* introduces a TEAD1 motif. Top, Locus plot with GWAS *P* values, the SNP color representing the LD to the lead SNP. Middle, Expanded view of a subregion showing the mean SNP probabilities from fine-mapping. Bottom, Read density of ATAC-seq assay from primary microglia[52].

the top gene within 500 kb by network score, including *ACE*, *BIN1*, *CASS4*, *CD2AP*, *PICALM*, *PLCG2* and *PTK2B*. At the *SLC24A4* locus, *RIN3* was strongly supported, whereas *SLC24A4* was not, in line with evidence from deleterious rare variants that *RIN3* may be causal[12].

Genes highly ranked by network propagation also include many outside of genome-wide-significant AD loci (Supplementary Table 10). Consistent with their involvement in AD, such genes tended to have SNPs with lower *P* values nearby than did remaining genes (Fig. 5a and Extended Data Fig. 3c), suggesting that numerous AD loci are yet to be discovered with larger GWAS sample sizes. Top network-ranked genes include *LILRB2* (nearby rs3855678, $P=9.8\times10^{-6}$), which encodes a leukocyte immunoglobulin-like receptor that recognizes multiple human leukocyte antigen (HLA) alleles and may also be involved in amyloid beta fibril growth[64]; *ABCA1* (rs59237458, $P=4\times10^{-6}$), involved in phospholipid transfer to apolipoproteins and previously associated with AD[65]; *SREBF1* (rs35763683, $P=2\times10^{-6}$), required for lipid homeostasis; and *AGRN* (rs2710871, $P=4\times10^{-6}$), involved in synapse formation in mature hippocampal neurons. Overall, genes with high network ranks were strongly enriched in biological processes and pathways that have previously been associated with AD, including clathrin-mediated endocytosis, activation of the immune response, phagocytosis, ephrin signaling and complement activation (Supplementary Table 11).

AD risk is enriched near genes with high microglial gene expression. To understand the contribution of cell-type-specific gene expression to AD risk, we used functional genome-wide

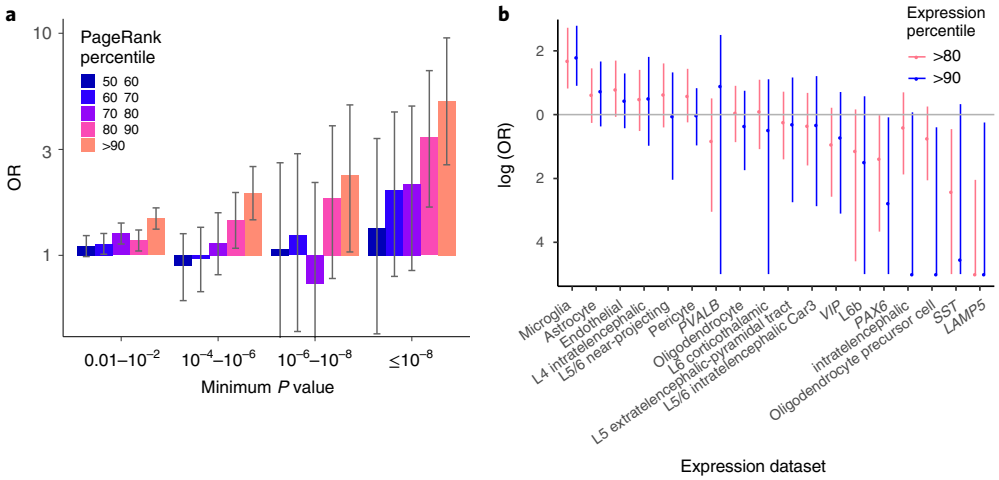


Fig. 5 | Genome-wide network and gene expression enrichments. a, Enrichment of low GWAS *P* values within 10 kb of genes having high versus low network PageRank percentile (low defined as below the 50th percentile). The whiskers represent the 95% confidence intervals based on Fisher's exact test for $n = 18{,}055$ genes. **b**, Enrichment of AD risk near genes with high expression in each brain cell type (above the 80th or 90th percentile) relative to the other cell types. Cell types are defined based on the single-cell clusters defined in Hodge et al.[67]. Neuronal cells are defined either by cortical layer (L4, L5, L6) and/or by projection target (intratelencephalic, corticothalamic, extratelencephalic-pyramidal tract, near-projecting), or by binary marker genes (*LAMP5*, *PAX6*, *PVALB*, *VIP* and *SST*). The whiskers represent the 95% confidence intervals as determined by fgwas[66].

association analysis (fgwas)[66] to assess the genome-wide enrichment of SNPs near genes highly expressed in specific cell types, based on a single-nucleus sequencing dataset of 49,495 nuclei from six human brain cortical areas[67,68]. Out of 18 broad cell type clusters, only microglia showed clear enrichment of AD risk (odds ratio (OR) = 6.0) near genes with expression above the 90th percentile across cell types (Fig. 5b). We performed a similar analysis looking at bulk gene expression across human tissues from GTEx, along with a small number of additional RNA sequencing datasets, including sorted primary microglia from brain surgeries[14] (Extended Data Fig. 4 and Supplementary Table 12). This gave consistent results, with microglia showing strong enrichment (OR = 4.4), followed by tissues rich in immune cells, including spleen (OR = 3.6) and whole blood (OR = 3.2). Notably, iPSC-derived microglia showed similar enrichment to primary microglia, while bulk brain tissues (including hippocampus) showed no enrichment.

Integrative gene prioritization from five lines of evidence. Determining the genes responsible for AD risk across GWAS loci is challenging, in part because few genes have been definitively confirmed as having a causal role. Therefore, we developed a comprehensive gene prioritization score, which incorporates quantitative information based on five lines of evidence: gene distance to lead SNPs; colocalization; network score; bulk and single-cell gene expression; and the sum of fine-mapped probability for any coding SNPs within a gene (Fig. 6, Extended Data Figs. 5 and 6 and Supplementary Table 13).

We first explored how best to use colocalization information. We found that genes with maximum colocalization probability above 0.9 had higher prioritization scores based on the other 4 predictors, but this was not the case for genes with weaker colocalization evidence (Extended Data Fig. 5a). We also examined colocalizations in different cell types or tissue groups, such as brain, microglia and other GTEx tissues. There was little evidence that colocalizing genes within any specific groups had higher total scores than other groups (Extended Data Fig. 5b), although this conclusion was limited by the low number of studies in some cell types, such as microglia. Therefore, we based our colocalization score on the maximum colocalization probability across tissues (>0.9) and normalized this to the 0–1 range.

A priori, we do not know which lines of evidence are most important for prioritizing genes. Therefore, we sought a systematic way to identify appropriate weights for the predictors. Although we do not know the causal AD genes, we selected two independent, unbiased sets of candidate genes for use in supervised learning: genes nearest to the GWAS peaks; and genes with high network scores (>80th percentile). To identify weights for our predictive features, we defined two models to discriminate these two gene sets from others within 500 kb, in each case using cross-validated lasso-regularized logistic regression with the remaining variables as predictors. As expected, when predicting genes nearest GWAS peaks, the highest-weight predictor was fine-mapped coding variants; however, only a few loci have such variants. The most informative predictor, determined based on a change in mean squared error (MSE) when the predictor is left out, was colocalization, followed by coding variants and then network score (Supplementary Table 14). When predicting high network score genes, the most informative predictor was distance to GWAS peak, followed by microglial gene expression; neither colocalization nor coding variant predictors improved the model. For both models, including hippocampus expression (GTEx) or single-cell astrocyte expression resulted in worse models (increased MSE).

We defined our gene prioritization 'model score' as the average of the predictions from our two models. The model score identified as top-ranked many AD candidate genes previously suggested as causal (Fig. 6). Exemplifying the importance of integrating genetic evidence sources, *ABCA7*, *SORL1* and *CR1* were top-ranked by overall score at their respective loci, despite having only moderate network-based scores, while *SORL1*, *PICALM* and *SPI1* were top-ranked despite having limited eQTL colocalization evidence.

While our prioritization further supports many established AD candidate genes, it also implicates new genes. Among these are *FCER1G*, which has been reported as a hub gene in microglial gene modules associated with neurodegeneration[69,70] and has been experimentally shown to influence microglial phagocytosis[71]. Another candidate is *ZYX*, which receives a top network score, is highly expressed in microglia and was recently nominated as an AD risk gene based on chromatin interactions between the *ZYX* promoter and AD risk variants in a *ZYX* enhancer[72].

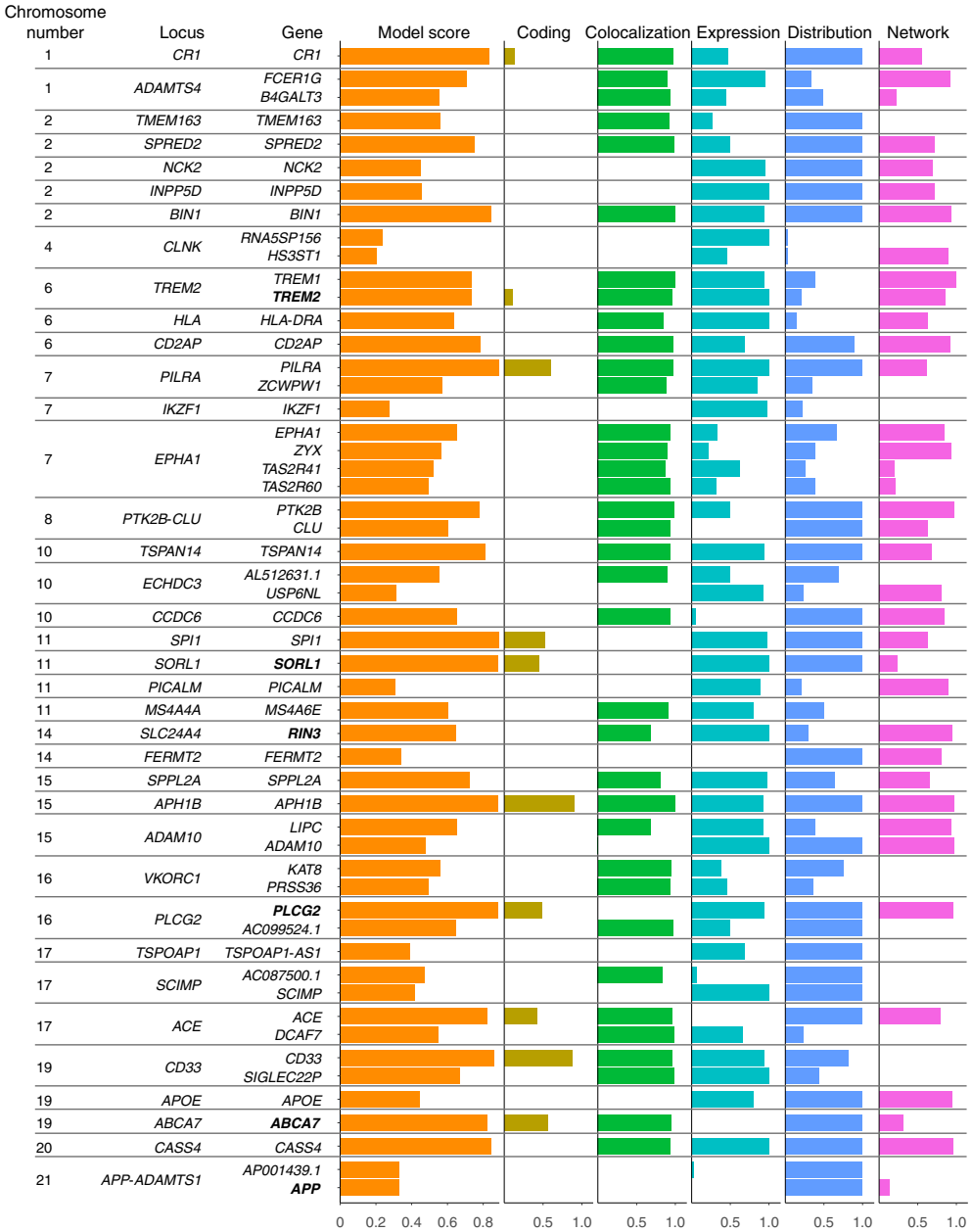


Fig. 6 | Gene evidence summary. The top gene at each locus is shown, as well as the next 13 top genes by model score; for 3 loci where a noncoding gene was the top-scoring gene, we also show the top-scoring protein-coding gene. Score components for each gene are indicated by colored bars and the points show the distribution of scores for all genes within 500 kb at the locus. Bold gene names are those with evidence of causality based on rare variants from other studies. Scores for all genes are listed in Supplementary Table 13.

Discussion

Identifying therapeutic targets for human diseases is a key goal of human genetics research and is particularly important for neurodegenerative diseases such as AD, for which no disease-modifying therapies yet exist. However, identifying the causal genes and genetic variants from GWAS is challenging since noncoding associations can act via the regulation of distal genes. We approached this challenge for AD by performing comprehensive fine-mapping, eQTL colocalization, network analysis and quantitative gene prioritization.

Our meta-analysis identified four new associations near *NCK2*, *SPRED2*, *TSPAN14* and *CCDC6*. Each of these was the nearest gene to the association peak and was supported by both eQTL colocalization and network ranking. Yet, despite the large number of eQTL datasets we used, colocalization of likely AD risk genes was sometimes found in only one or a few datasets; this was the case for *SPRED2* (TwinsUK lymphoblastoid cell line colocalization probability = 0.99), *RIN3* (GTEx frontal cortex probability = 0.94) and *PILRA* (Fairfax 2-h lipopolysaccharide monocyte colocalization probability = 0.99). Many factors could account for dataset-specific colocalizations, such as biological differences in sample state, differences in LD match between the GWAS and eQTL datasets and technical differences in the transcriptome annotations used for eQTL discovery. As a result, absence of colocalization provides only weak evidence for lack of an effect in a given tissue type, whereas positive colocalization provides strong support for a shared genetic effect.

Therefore, it is useful to look broadly across eQTL studies for colocalization, which will be facilitated by resources that simplify access to these datasets, such as the eQTL Catalogue[11].

One of our most confidently prioritized genes was *APH1B*, encoding a γ-secretase complex component involved in APP processing. *APH1B* harbors the likely causal missense variant p.Thr27Ile, yet it also has strong colocalization evidence that higher expression correlates with higher AD risk. One possibility is that impaired function of *APH1B* due to the missense variant leads to upregulation of *APH1B* transcription. This interpretation would be consistent with evidence from both mice[73] and humans[74] that loss of APH1B and γ-secretase function leads to AD. It is noteworthy, however, that recent experiments failed to find an effect of the T27I variant on γ-secretase activity in HEK cells[75].

Among our new associations, TSPAN14 has a role in defining the localization of ADAM10 (ref. [76]), another recently discovered AD protein that is a key component of the α-secretase complex and that could thus mediate AD risk via processing of amyloid precursor protein. However, ADAM10 also cleaves the microglia-associated protein TREM2 to generate its soluble ligand-binding domain[77]. Our fine-mapping showed that the risk SNP rs1870138 is also associated with higher risk for inflammatory bowel disease, an immune-mediated disease, and with higher monocyte count in UKB participants. Since *TSPAN14* is expressed more highly in immune cell types, including microglia, than in brain tissue, it is also plausible that AD risk is mediated by its effect on either immune cell count or activation. Recently proposed AD candidate genes supported by our analyses include *RIN3*, *HS3ST1* and *FCER1G*. As noted above, FCER1G is a microglial master regulator[69–71]; RIN3 interacts with both BIN1 and CD2AP in the early endocytic pathway[78]; and HS3ST1 is involved in the cellular uptake of tau[79] and was recently associated with AD in an independent Norwegian sample[59].

In summary, our study reports quantitative gene prioritization for 36 AD-associated regions as well as AD-specific gene network scores beyond these loci. Our genetic findings highlight the presence of diverse mechanisms in AD pathogenesis and suggest candidate targets for therapeutic development.

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/s41588-020-00776-w.

Received: 16 January 2020; Accepted: 23 December 2020;
Published online: 15 February 2021

References

1. Liu, J. Z., Erlich, Y. & Pickrell, J. K. Case-control association mapping by proxy using family history of disease. *Nat. Genet.* **49**, 325–331 (2017).
2. Marioni, R. E. et al. GWAS on family history of Alzheimer's disease. *Transl. Psychiatry* **8**, 99 (2018).
3. 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).
4. Claussnitzer, M. et al. FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* **373**, 895–907 (2015).
5. Sekar, A. et al. Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
6. Huang, H. et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173–178 (2017).
7. Malik, M. et al. CD33 Alzheimer's risk-altering polymorphism, CD33 expression, and exon 2 splicing. *J. Neurosci.* **33**, 13320–13325 (2013).
8. Guerreiro, R. et al. TREM2 variants in Alzheimer's disease. *N. Engl. J. Med.* **368**, 117–127 (2013).
9. Sims, R. et al. Rare coding variants in *PLCG2*, *ABI3*, and *TREM2* implicate microglial-mediated innate immunity in Alzheimer's disease. *Nat. Genet.* **49**, 1373–1384 (2017).
10. Kerimov, N. et al. eQTL Catalogue: a compendium of uniformly processed human gene expression and splicing QTLs. Preprint at *bioRxiv* https://doi.org/10.1101/2020.01.29.924266 (2020).
11. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
12. Kunkle, B. W. et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).
13. Lambert, J. C. et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
14. Young, A. M. H. et al. A map of transcriptional heterogeneity and regulatory variation in human microglia. Preprint at *bioRxiv* https://doi.org/10.1101/2019.12.20.874099 (2019).
15. Leung, Y. Y. et al. Identifying amyloid pathology-related cerebrospinal fluid biomarkers for Alzheimer's disease in a multicohort study. *Alzheimers Dement. (Amst)* **1**, 339–348 (2015).
16. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
17. Moreno-Grau, S. et al. Genome-wide association analysis of dementia and its clinical endophenotypes reveal novel loci associated with Alzheimer's disease and three causality networks: the GR@ACE project. *Alzheimers Dement.* **15**, 1333–1347 (2019).
18. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
19. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
20. Sieberts, S. K. et al. Large eQTL meta-analysis reveals differing patterns between cerebral cortical and cerebellar brain regions. *Sci. Data* **7**, 340 (2020).
21. Ng, B. et al. An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat. Neurosci.* **20**, 1418–1426 (2017).
22. Schmiedel, B. J. et al. Impact of genetic polymorphisms on human immune cell gene expression. *Cell* **175**, 1701–1715.e16 (2018).
23. Jaffe, A. E. et al. Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat. Neurosci.* **21**, 1117–1125 (2018).
24. Buil, A. et al. Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat. Genet.* **47**, 88–91 (2015).
25. Fairfax, B. P. et al. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* **44**, 502–510 (2012).
26. Fairfax, B. P. et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).
27. Naranbhai, V. et al. Genomic modulators of gene expression in human neutrophils. *Nat. Commun.* **6**, 7545 (2015).
28. Chen, L. et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* **167**, 1398–1414.e24 (2016).
29. Alasoo, K. et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.* **50**, 424–431 (2018).
30. Gutierrez-Arcelus, M. et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife* **2**, e00523 (2013).
31. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
32. Kilpinen, H. et al. Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**, 370–375 (2017).
33. Nédélec, Y. et al. Genetic ancestry and natural selection drive population differences in immune responses to pathogens. *Cell* **167**, 657–669.e21 (2016).
34. Quach, H. et al. Genetic adaptation and Neandertal admixture shaped the immune system of human populations. *Cell* **167**, 643–656.e17 (2016).
35. Schwartzentruber, J. et al. Molecular and functional variation in iPSC-derived sensory neurons. *Nat. Genet.* **50**, 54–61 (2018).
36. van de Bunt, M. et al. Transcript expression data from human islets links regulatory signals from genome-wide association studies for type 2 diabetes and glycemic traits to their downstream effectors. *PLoS Genet.* **11**, e1005694 (2015).
37. Momozawa, Y. et al. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nat. Commun.* **9**, 2427 (2018).
38. Chun, S. et al. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.* **49**, 600–605 (2017).

39. Salazar, S. V. et al. Alzheimer's disease risk factor Pyk2 mediates amyloid-β-induced synaptic dysfunction and loss. *J. Neurosci.* **39**, 758–772 (2019).
40. 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).
41. Calafate, S., Flavin, W., Verstreken, P. & Moechars, D. Loss of Bin1 promotes the propagation of Tau pathology. *Cell Rep.* **17**, 931–940 (2016).
42. Nott, A. et al. Brain cell type-specific enhancer–promoter interactome maps and disease-risk association. *Science* **366**, 1134–1139 (2019).
43. Rathore, N. et al. Paired immunoglobulin-like type 2 receptor alpha G78R variant alters ligand binding and confers protection to Alzheimer's disease. *PLoS Genet.* **14**, e1007427 (2018).
44. Chan, G. et al. CD33 modulates TREM2: convergence of Alzheimer loci. *Nat. Neurosci.* **18**, 1556–1558 (2015).
45. Raj, T. et al. CD33: increased inclusion of exon 2 implicates the Ig V-set domain in Alzheimer's disease susceptibility. *Hum. Mol. Genet.* **23**, 2729–2736 (2014).
46. Jonsson, T. et al. Variant of TREM2 associated with the risk of Alzheimer's disease. *N. Engl. J. Med.* **368**, 107–116 (2013).
47. Claes, C. et al. Human stem cell-derived monocytes and microglia-like cells reveal impaired amyloid plaque clearance upon heterozygous or homozygous loss of TREM2. *Alzheimers Dement.* **15**, 453–464 (2019).
48. Aerts, J. et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).
49. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
50. Kichaev, G. et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722 (2014).
51. Benner, C. et al. Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *Am. J. Hum. Genet.* **101**, 539–551 (2017).
52. Gosselin, D. et al. An environment-dependent transcriptional network specifies human microglia identity. *Science* **356**, eaal3222 (2017).
53. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
54. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
55. Davydov, E. V. et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).
56. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
57. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e24 (2019).
58. Steinberg, S. et al. Loss-of-function variants in *ABCA7* confer risk of Alzheimer's disease. *Nat. Genet.* **47**, 445–447 (2015).
59. De Roeck, A. et al. An intronic VNTR affects splicing of *ABCA7* and increases risk of Alzheimer's disease. *Acta Neuropathol.* **135**, 827–837 (2018).
60. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nat. Genet.* **50**, 1593–1599 (2018).
61. Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
62. Lanoiselée, H.-M. et al. *APP*, *PSEN1*, and *PSEN2* mutations in early-onset Alzheimer disease: a genetic screening study of familial and sporadic cases. *PLoS Med.* **14**, e1002270 (2017).
63. Fang, H. et al. A genetics-led approach defines the drug target landscape of 30 immune-related traits. *Nat. Genet.* **51**, 1082–1091 (2019).
64. Amin, L. & Harris, D. A. Aβ receptors specifically recognize molecular features displayed by fibril ends and neurotoxic oligomers. Preprint at *bioRxiv* https://doi.org/10.1101/822361 (2019).
65. Nordestgaard, L. T., Tybjærg-Hansen, A., Nordestgaard, B. G. & Frikke-Schmidt, R. Loss-of-function mutation in *ABCA1* and risk of Alzheimer's disease and cerebrovascular disease. *Alzheimers Dement.* **11**, 1430–1438 (2015).
66. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).
67. Hodge, R. D. et al. Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).
68. Bakken, T. E. et al. Evolution of cellular diversity in primary motor cortex of human, marmoset monkey, and mouse. Preprint at *bioRxiv* https://doi.org/10.1101/2020.03.31.016972 (2020).
69. Mukherjee, S., Klaus, C., Pricop-Jeckstadt, M., Miller, J. A. & Struebing, F. L. A microglial signature directing human aging and neurodegeneration-related gene networks. *Front. Neurosci.* **13**, 2 (2019).
70. Zhang, B. et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153**, 707–720 (2013).
71. Patel, K. R. et al. Single cell-type integrative network modeling identified novel microglial-specific targets for the phagosome in Alzheimer's disease. Preprint at *bioRxiv* https://doi.org/10.1101/2020.06.09.143529 (2020).
72. Novikova, G. et al. Integration of Alzheimer's disease genetics and myeloid genomics reveals novel disease risk mechanisms. Preprint at *bioRxiv* https://doi.org/10.1101/694281 (2019).
73. Biundo, F., Ishiwari, K., Del Prete, D. & D'Adamio, L. Deletion of the γ-secretase subunits *Aph1B/C* impairs memory and worsens the deficits of knock-in mice modeling the Alzheimer-like familial Danish dementia. *Oncotarget* **7**, 11923–11944 (2016).
74. Nicolas, G. et al. Somatic variants in autosomal dominant genes are a rare cause of sporadic Alzheimer's disease. *Alzheimers Dement.* **14**, 1632–1639 (2018).
75. Zhang, X. et al. Negative evidence for a role of *APH1B T27I* variant in Alzheimer's disease. *Hum. Mol. Genet.* **29**, 955–966 (2020).
76. Matthews, A. L. et al. Regulation of leukocytes by TspanC8 tetraspanins and the 'molecular scissor' ADAM10. *Front. Immunol.* **9**, 1451 (2018).
77. Schlepckow, K. et al. An Alzheimer-associated TREM2 variant occurs at the ADAM cleavage site and affects shedding and phagocytic function. *EMBO Mol. Med.* **9**, 1356–1365 (2017).
78. Juul Rasmussen, I., Tybjærg-Hansen, A., Rasmussen, K. L., Nordestgaard, B. G. & Frikke-Schmidt, R. Blood–brain barrier transcytosis genes, risk of dementia and stroke: a prospective cohort study of 74,754 individuals. *Eur. J. Epidemiol.* **34**, 579–590 (2019).
79. Zhao, J. et al. Rare 3-*O*-sulfation of heparan sulfate enhances Tau interaction and cellular uptake. *Angew. Chem. Int. Ed. Engl.* **59**, 1818–1827 (2020).
80. Jun, G. R. et al. Transethnic genome-wide scan identifies novel Alzheimer's disease loci. *Alzheimers Dement.* **13**, 727–738 (2017).
81. Andersen, O. M., Rudolph, I.-M. & Willnow, T. E. Risk factor *SORL1*: from genetic association to functional validation in Alzheimer's disease. *Acta Neuropathol.* **132**, 653–665 (2016).
82. Sassi, C. et al. Influence of coding variability in APP-Aβ metabolism genes in sporadic Alzheimer's disease. *PLoS ONE* **11**, e0150079 (2016).
83. Lu, Q. et al. Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset Alzheimer's disease. *PLoS Genet.* **13**, e1006933 (2017).
84. Ghanbari, M. et al. A functional variant in the miR-142 promoter modulating its expression and conferring risk of Alzheimer disease. *Hum. Mutat.* **40**, 2131–2145 (2019).
85. Chung, C.-M. et al. Fine-mapping angiotensin-converting enzyme gene: separate QTLs identified for hypertension and for ACE activity. *PLoS ONE* **8**, e56119 (2013).
86. Nylocks, K. M. et al. An angiotensin-converting enzyme (ACE) polymorphism may mitigate the effects of angiotensin-pathway medications on posttraumatic stress symptoms. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **168B**, 307–315 (2015).
87. Kamboh, M. I. et al. Genome-wide association study of Alzheimer's disease. *Transl. Psychiatry* **2**, e117 (2012).
88. Bernstein, A. I. et al. 5-Hydroxymethylation-associated epigenetic modifiers of Alzheimer's disease modulate Tau-induced neurotoxicity. *Hum. Mol. Genet.* **25**, 2437–2450 (2016).
89. Witoelar, A. et al. Meta-analysis of Alzheimer's disease on 9,751 samples from Norway and IGAP study identifies four risk loci. *Sci. Rep.* **8**, 18088 (2018).

Methods

GWAS on the family history of AD. Sample quality control, variant quality control and imputation were performed on all UKB participants as described in Bycroft et al.[90]. After genotype imputation, 93,095,623 variants across 487,409 individuals were available for analysis. To exclude individuals of non-European ancestry, we extracted 'white British' ancestry participants as described in Bycroft et al.[90]. These individuals self-reported their ethnic background as 'British' and have similar genetic ancestry based on principal component analysis. To extract additional individuals of European ancestry, we followed a similar approach to Bycroft et al.[90] and applied the R package Aberrant[91] v.1.0 on principal components 1v2, 3v4 and 5v6 across the individuals who self-reported as 'Irish' or 'any other white background'. We identified first-degree relatives by applying KING[92] v.2.0 to 147,522 UKB participants who had at least one relative identified in Bycroft et al.[90] (UKB field 22021). For each first-degree relative pair, we prioritized AD cases and proxy cases (see below) for inclusion and otherwise excluded one of the pair at random. We also excluded variants with low imputation quality (INFO score < 0.3) and/or those with MAFs < 0.0005, resulting in 25,647,815 variants available for analysis.

AD cases were extracted from the UKB self-report (field 20002), International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10) diagnoses (fields 41202 and 41204) and ICD-10 cause of death (fields 40001 and 40002) data. UKB participants were asked whether they had a biological father, mother or sibling who had AD/dementia (UKB fields 20107, 20110 and 20111, respectively). We extracted all participants with at least one affected relative as proxy cases. Participants who answered 'do not know' or 'prefer not to answer' were excluded from the analyses. All remaining individuals were denoted as controls.

There were 898 AD cases, 52,791 AD proxy cases and 355,900 controls in the combined white British and white non-British cohorts. For the association analyses, we lumped the true and proxy cases together (53,042 unique affected individuals) and used the linear mixed model implemented in BOLT-LMM[93] v.2.3.2.

AD meta-analysis. To enable meta-analysis combining the UKB cohorts with external case-control studies, we transformed the AD proxy BOLT linear-model effect sizes to a log odds-ratio scale by dividing the β and s.e. values from BOLT by $f(1-f)$:

$$\log\mathrm{OR} \approx \beta_{\mathrm{LMM}}/(f(1-f))$$

with the s.e.:

$$\mathrm{s.e.} \approx \mathrm{s.e.}_{\mathrm{LMM}}/(f(1-f))$$

where β_{LMM} and $\mathrm{s.e.}_{\mathrm{LMM}}$ are the SNP effect sizes and standard errors respectively from BOLT-LMM and f is the fraction of cases in the sample[94]. Since the affected individuals in our analysis include both true and proxy cases, we then multiplied the transformed logORs and standard errors by 1.897 to approximate the logORs obtained from a true case-control study[1].

We combined the transformed UKB white British cohort, UKB white non-British cohort and the stage 1 summary statistics from Kunkle et al.[12] using a fixed-effects (inverse variance-weighted) meta-analysis across 10,687,126 overlapping variants. For display purposes (Supplementary Table 8), we used CrossMap[95] v.0.2.5 to convert variant positions from GRCh37 to GRCh38.

Replication. To assess replication of our discovered signals, we downloaded the publicly available summary statistics for the GR@ACE study of AD[17] from the GWAS catalog and for the FinnGen GWAS of phenotypes 'Alzheimer's disease, wide definition' and 'Alzheimer's disease (late onset)' from FinnGen release 3. We extracted summary results for our lead SNPs or a partner in strong LD when the lead SNP was not found and present these in Supplementary Table 3. We estimated power to detect our four new loci at nominal significance ($P < 0.05$) using the genetic power calculator (http://zzz.bwh.harvard.edu/gpc/cc2.html) with the genotype relative risks estimated from our meta-analysis and the allele frequency and case-control count from the GWAS study of interest (GR@ACE or FinnGen), assuming a disease prevalence of 5%. We performed an inverse variance-weighted meta-analysis of all four studies (Kunkle et al.[12], UKB, GR@ACE and FinnGen 'AD wide'), similar to our discovery meta-analysis.

Conditional analysis and statistical fine-mapping. To run GCTA, we prepared Plink input files with genotypes from 10,000 randomly sampled UKB individuals at variants within ±5 megabases (Mb) from each lead SNP. We excluded variants with an INFO score < 0.85 or which had a P value from Cochran's Q test for study heterogeneity <0.001. We also excluded variants with an MAF in the UKB <0.1% since LD estimates are unreliable at low allele counts. We selected these thresholds after manual examination of fine-mapping results, where we found that more lenient cutoffs led either FINEMAP or PAINTOR to select implausible causal variants at a few loci, such as pairs of very weakly associated rare variants to explain a common variant signal. We ran GCTA v.1.92.1 --cojo-slct with a threshold of $P < 10^{-5}$ to identify secondary signals at each locus and then retained only loci with a lead $P < 5 \times 10^{-8}$. For the HLA locus, we used a GCTA P value threshold of 5×10^{-8}. We also retained the loci *TSPOAP1*, *IKZF1* and *TMEM163* since they had $P < 5 \times 10^{-8}$ in an earlier version of our analysis. We excluded the *APOE* locus from conditional analysis and fine-mapping because the strength of association in the region would require a more perfect LD panel match to avoid spurious signals.

We then ran FINEMAP v.1.3 at each locus with --n-causal-snps given as the number of independent SNPs determined by GCTA. For FINEMAP, we excluded variants with an MAF < 0.2%. For loci with multiple signals, we also used GCTA --cojo-cond to condition on each independent SNP identified in the previous analysis and retained SNPs within 500 kb of any conditionally independent SNP at the locus. To fine-map based on the GCTA conditional signals, we converted beta and standard error values to approximate Bayesian factors[96] using a prior of W = 0.1 (in Wakefield notation) and used the Wellcome Trust Case Control Consortium (WTCCC) single causal variant method[48], probability = SNP Bayesian factor/sum (all SNP Bayesian factors).

To assess the sensitivity of the results to our choice of reference panel, we applied the same steps (GCTA + FINEMAP) to the summary statistics from the Kunkle et al.[12] substudy, which are described further in the Supplementary Note.

Colocalization with eQTLs. For eQTL colocalization, we downloaded the summary statistics and determined eQTL genes at a false discovery rate of 5% for each dataset in a uniform manner, first using Bonferroni correction of lead SNP nominal P values based on the number of variants tested for the gene and using the Benjamini–Hochberg method to compute the false discovery rate. QTL calling for primary microglia was performed with RASQUAL[97] v.0.1 with the --no-posterior-update option. For datasets in GRCh38 coordinates, we first used CrossMap[95] to convert back to GRCh37 coordinates to match variants between eQTL and GWAS. We used the coloc package[19] v.4.04 with default priors to perform colocalization tests between GWAS and eQTLs having lead variants within 500 kb of each other and passed to coloc all variants within 200 kb of each lead variant. We also ran coloc using P values for each conditionally independent GWAS signal, obtained with GCTA as described above.

Functional annotations. We used the Ensembl VEP online Web tool (http://www.ensembl.org/info/docs/tools/vep/index.html)[54] to predict variant consequences and add selected annotations (Supplementary Table 7). We downloaded BED files based on imputed data for Roadmap Epigenomics DNase and 25 state genome segmentations for 127 epigenomes[53]. We grouped these into groups 'all', 'brain' (epigenomes 7, 9, 10, 53, 54, 67, 68, 69, 70, 71, 72, 73, 74, 81, 82, 125) and 'blood and immune' (epigenomes 33, 34, 37, 38, 39, 40, 41, 42, 43, 44, 45, 47, 48, 62, 29, 30, 31, 32, 35, 36, 46, 50, 51, 116). We considered nine genome segmentation states to represent enhancers: TxReg; TxEnh5; TxEnh3; TxEnhW; EnhA1; EnhA2; EnhAF; EnhW1; EnhW2. We used bedtools[98] v.2.27.1 to determine overlaps and counted the number of overlaps for each variant with peaks in the above groups. We downloaded FANTOM5 (ref. [99]) permissive enhancer annotations from https://fantom.gsc.riken.jp/5/data/. We downloaded precomputed SpliceAI scores[57] for variants within genes from https://github.com/Illumina/SpliceAI (accessed 7 March 2019). We merged filtered whole-genome and exome scores together; for each AD variant, we annotated the maximum score across splice donor gain, donor loss, acceptor gain and acceptor loss. We used DeepSEA[56] (http://deepsea.princeton.edu/job/analysis/create/) to annotate variants selected for functional fine-mapping with DeepSEA's 'functional significance' score. BigWig files with PhastCons, PhyloP and GERP rejected substitution scores were downloaded from University of California, Santa Cruz. We downloaded microglial ATAC-seq based on the study by Gosselin et al.[52], aligned reads to GRCh37 with Burrows–Wheeler Aligner v.0.7.15 (ref. [100]) and called multisample peaks across all 15 datasets using MACS2 (ref. [101]) v.2.1.2. We prepared BigWig files from the alignments by using bedtools genomecov, followed by bedGraphToBigWig. To visualize the microglia ATAC-seq tracks, we adapted code from wiggleplotr[102] v.1.10.1.

Annotation-based fine-mapping. For fine-mapping with PAINTOR, we first restricted the number of considered variants for computational feasibility by selecting 3,207 variants which had (1) a FINEMAP probability ≥0.01% based on the GCTA-identified number of causal variants at the locus or (2) had a FINEMAP probability ≥1% when run with either 1 or 2 causal variants or (3) were among the top 20 variants at the locus by FINEMAP probability. We defined binary annotations for input to PAINTOR based on the features described above, thresholding certain scores at multiple levels (for example, combined annotation-dependent depletion ≥ 5, 10, 20). For Roadmap annotations, we included a category based on whether a variant was in a peak or enhancer in ≥10 epigenomes. We ran PAINTOR v.3.1 once for each of the 43 annotations (Fig. 2 and Supplementary Table 7), allowing 2 causal variants per locus.

We built a multi-annotation model using forward stepwise selection. We selected the best annotation by log-likelihood (LLK), blood and immune DNase, then ran PAINTOR again for each combination of this annotation and the 42 remaining annotations. We added a top-ranking annotation at each iteration until the model LLK improvement was <1. This occurred at iteration 4 and so we kept the first three annotations in the combined model. We computed the mean

causal probability for each SNP as the mean of the three fine-mapping methods at loci with two or more signals or as the mean of the FINEMAP and PAINTOR probabilities for loci with one signal since FINEMAP gives approximately the same results as WTCCC fine-mapping for a single causal variant.

Network analysis. For network analysis, we created a gene interaction network based on selecting all edges between protein-coding genes from systematic studies (>1,000 interactions) in the IntAct[103] v.2019-05-02 and BioGRID databases[104] v.3.5.172 and edges from STRING v.10.5 (ref. [105]) with an edge score >0.75. This combined network included 18,055 genes and 540,421 edges. We identified 28 top candidate genes across AD loci (Supplementary Table 9) to use as seed genes and assigned weight to these as the $-\log_{10}(P)$ of the locus lead SNP. We added four genes from the literature (*MAPT*, *PSEN1*, *PSEN2* and *ABI3*), with a weight (equivalent $-\log_{10}(P)$) of 15. For three loci, the nearest gene was not present in the network (*ECHDC3*, *TMEM163* and *SCIMP*). For each locus, we used all seed genes as input except those at the same locus, and propagated information through the network with the personalized PageRank algorithm[106], included in the igraph R package[107] v.1.2.4.2. Since a gene's resulting PageRank was highly correlated with its node degree, we compared the PageRank of each gene to the distribution of PageRanks obtained for the same gene in 1,000 iterations of network propagation, where the same number of seed genes were randomly selected. We computed the percentile of a gene's true PageRank relative to the 1,000 network propagations with randomized inputs. Although the distribution of PageRank percentile was fairly uniform, we further normalized this to a uniform distribution across genes, so that a PageRank percentile of 90% indicates that a gene's PageRank relative to permutations is above that of 90% of genes. To determine gene set enrichment, we used the top 1,000 genes by network rank as input to gProfiler[108] with default settings, with the set of all genes ranked by the network as a background set. To determine enrichment of low *P* value AD SNPs near genes in specific bins of the PageRank percentile (Fig. 5a), we first determined for each gene the minimum SNP *P* value within 10 kb of the gene's footprint. We excluded genes within 1 Mb of *APOE*. Then, for genes in each PageRank percentile bin, we used Fisher's exact test to determine the OR for a gene in that bin (relative to genes with a PageRank percentile <50%) to have a minimum SNP *P* value in the given bin (relative to genes with a minimum SNP $P > 0.01$).

Gene expression. Gene expression values for all tissues were determined in units of transcripts per million (TPM). Both GTEx v.8 and the eQTL Catalogue provide tables of the median TPM expression across samples for each tissue and gene. For primary microglia, we obtained a table of read counts per gene computed using featureCounts v.1.5.3 as described by Young et al.[14], from which we computed the median TPM. For use in gene prioritization and enrichment analyses, we first selected four GTEx brain tissues (cortex, hippocampus, substantia nigra, cerebellum) to avoid overrepresenting the brain, and then the remaining 41 GTEx tissues, as well primary microglia and in-house expression data from iPSC-derived microglia, iPSC-derived NGN2 cortical neurons and iPSC-derived neurons from growth factor differentiation. For each gene, we determined the TPM expression relative to all tissues/cell types.

Single-cell gene expression data were obtained from the Allen Institute for Brain Research as a gene-by-cell counts table based on SMART (switching mechanism at the 5′ end of RNA template) sequencing of six human brain cortical areas[109]. For each cell type 'subclass' as defined in the metadata (but excluding vascular and leptomeningeal cells for having too few cells and the outlier subclass labeled 'exclude'), counts were summed across cells and then normalized to TPM within each subclass. We determined each gene's TPM expression in each subclass relative to all 18 subclasses.

Genome-wide enrichment. We determined the GRCh37 coordinates of 18,055 genes present in the gene network using the R package annotables v.0.1.91. For each AD GWAS SNP, excluding the *APOE* region (chromosome 19: 44–47 Mb), we determined the nearest gene. We defined annotation inputs for FGWAS labeling an SNP 1 if it was nearest to a gene with network score in a given percentile bin (50–60, 60–70, 70–80, 80–90, 90–95, >95) and 0 otherwise. We ran FGWAS[75] (-cc) with all network annotations as input so that enrichments were with respect to SNPs nearest to genes with a network score <50th percentile. For every bulk gene expression dataset selected above, we defined an annotation for the nearest genes to SNPs with relative expression above the 80th (or 90th) percentile and similarly for cell types from single-cell gene expression. We ran FGWAS once for each expression annotation to determine the enrichment of SNPs near high-expression genes relative to remaining genes (Supplementary Table 12).

Gene prioritization. Five predictors were used for gene prioritization: (1) the coding score is the sum of the mean fine-mapping probability for missense or loss of function variants in a gene; (2) the expression score is the sum of component scores for bulk and single-cell microglial expression and rewards genes with an expression percentile above the 50th: exprScore = (bulkExprScore + singleCellExprScore)/2 bulkExprScore = max(0, bulk_microglia_pctile − 50)/50 singleCellExprScore = max(0, sc_microglia_pctile − 50)/50; genes without measured expression in a given dataset (bulk, single-cell) are assigned an expression score of zero for that dataset. Recent evidence from both eQTLs[110] and metabolite GWAS[111] suggests that genomic distance from the association peak is a strong predictor of causal target genes; (3) the distance score is defined to give reasonable scores over the main range of interest of 0–200 kb: $\text{distScore} = (\log_{10}(\text{maxDist}) - \log_{10}((\text{abs}(x) + \text{distBias})))/(\log_{10}(\text{maxDist}) - \log_{10}(\text{distBias}))$ where x is the minimum distance from the gene's footprint to the region defined by independent lead SNPs at a GWAS locus, maxDist is 500,000 and distBias is 100 (Extended Data Fig. 6); (4) the coloc score is defined based on the maximum value across QTL datasets of the 'hypothesis 4' probability and rewards colocalization probabilities >0.9: colocScore = max(0, max(QTL dataset hypothesis 4) − 0.9); (5) the network score is determined based on the PageRank percentile for a gene relative to permutations: networkScore = max(0, (PageRank pctile − 50)/50). Genes not present in the network are assigned a network score of zero. The total score for a gene is the sum of these five scores.

To give appropriate weight to each component, we trained lasso-regularized logistic regression models with cross-validation using glmnet[112]. As input we used all protein-coding genes within 500 kb of our AD GWAS peaks, excluding the *APOE* region due to lack of colocalization information and excluding genes not present in the network. For the distance model, genes within 10 kb of each GWAS peak (40 genes) were set as positives, genes 10–100 kb were excluded and genes >100 kb (394 genes) were set as negatives. These were predicted using the four non-distance predictors. For the network model, genes with a PageRank percentile >80% (143 genes) were set as positives, those with a PageRank percentile 50–80% were excluded and the 230 other genes were set as negatives; the latter were predicted using the four non-network predictors. In each case, we selected the model that minimized the MSE, shown in Supplementary Table 14, and used those parameters to generate predictions (in the range 0–1) for all genes at the AD loci. We defined the model score for a gene as the average prediction from the two models. To determine the importance of the predictors to each model (apart from looking at regression coefficients), we ran glmnet models excluding each predictor in turn. If the MSE was lower with a predictor excluded, then we removed it from the final model. For each model, we compared the MSE when using our quantitative predictors as defined above or using categorical predictors by thresholding the predictors into 2–4 bins. For both models, the quantitative predictors gave improved MSE. We also examined models that included as predictors expression scores from astrocytes (based on the single-cell data) and from brain hippocampus (based on the GTEx data), but for both models this resulted in higher MSE and the regularization set the coefficients to zero.

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

Data availability

Summary statistics from the meta-analysis are available through the National Human Genome Research Institute-European Bioinformatics Institute GWAS catalog under accession nos. GCST90012877 and GCST90012878 (https://www.ebi.ac.uk/gwas/downloads/summary-statistics). The eQTL Catalogue is available at http://www.ebi.ac.uk/eqtl/. The GTEx Portal is available at https://www.gtexportal.org/home/. The Roadmap Epigenomics is available at http://www.roadmapepigenomics.org/. DeepSEA is available at http://deepsea.princeton.edu/job/analysis/create/. SpliceAI is available at https://github.com/Illumina/SpliceAI. FANTOM5 enhancers are available at https://fantom.gsc.riken.jp/5/data/. GERP is available at hgdownload.cse.ucsc.edu/gbdb/hg19/bbi/All_hg19_RS.bw. PhyloP is available at hgdownload.cse.ucsc.edu/goldenpath/hg19/phyloP100way. PhastCons is available at hgdownload.cse.ucsc.edu/goldenpath/hg19/phastCons100way. The brain eQTL meta-analysis summary statistics are avilable at https://www.synapse.org/#!Synapse:syn16984815. Primary microglia eQTL summary statistics are available under EGA accession no. EGAD00001005736. Primary microglia ATAC-seq data are available under dbGaP accession no. phs001373.v1.p1. The Allen Brain Institute is available at http://portal.brain-map.org/atlases-and-data/rnaseq. The IntAct molecular interaction database is available at https://www.ebi.ac.uk/intact/. The BioGRID database is available at https://thebiogrid.org/. The STRING database is available at https://string-db.org/.

Code availability

The code for the analyses described in this article can be found at https://github.com/jeremy37/AD_finemap.

References

90. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
91. Bellenguez, C. et al. A robust clustering algorithm for identifying problematic samples in genome-wide association studies. *Bioinformatics* **28**, 134–135 (2012).
92. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

93. Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
94. Pirinen, M., Donnelly, P. & Spencer, C. C. A. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *Ann. Appl. Stat.* **7**, 369–390 (2013).
95. Zhao, H. et al. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).
96. Wakefield, J. Bayes factors for genome-wide association studies: comparison with *P*-values. *Genet. Epidemiol.* **33**, 79–86 (2009).
97. Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* **48**, 206–213 (2016).
98. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
99. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
100. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
101. Zhang, Y. et al. Model-based analysis of ChIP–Seq (MACS). *Genome Biol.* **9**, R137 (2008).
102. Alasoo, K. wiggleplotr: Make read coverage plots from BigWig files. R package version 1.10.1 https://bioconductor.org/packages/release/bioc/html/wiggleplotr.html (2019).
103. Orchard, S. et al. The MIntAct project: IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358–D363 (2014).
104. Chatr-Aryamontri, A. et al. The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* **45**, D369–D379 (2017).
105. Szklarczyk, D. et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362–D368 (2017).
106. Fogaras, D., Rácz, B., Csalogány, K. & Sarlós, T. Towards scaling fully personalized PageRank: algorithms, lower bounds, and experiments. *Internet Math.* **2**, 333–358 (2005).
107. Csardi, G., Nepusz, T. The igraph software package for complex network research. *InterJ. Complex Syst.* **1695**, 1–9 (2006).
108. Raudvere, U. et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).
109. Battle, A. et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14–24 (2014).
110. Barbeira, A. N. et al. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. Preprint at *bioRxiv* https://doi.org/10.1101/814350 (2019).
111. Stacey, D. et al. ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Res.* **47**, e3 (2019).
112. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).

Acknowledgements

This work was funded by Open Targets (OTAR037). We thank J. Barrett for guidance during the initiation of the project and K. Alasoo for early access to the eQTL Catalogue. We thank A. Ruiz for support in using the summary results from the GR@ACE study. We thank the participants and investigators of the FinnGen study and UK Biobank. R.J.M.F. is supported by grants from the UK Multiple Sclerosis Society (MS 50), the Adelson Medical Research Foundation and a core support grant from the Wellcome Trust and Medical Research Council (MRC) to the Wellcome-MRC Cambridge Stem Cell Institute (no. 203151/Z/16/Z). A.M.H.Y. is supported by a Wellcome Trust PhD for Clinicians fellowship.

Author contributions

J.S. planned and conducted the analyses and wrote the paper. J.Z.L. performed the GWAX and meta-analysis. S.C. and E.B. assisted with fine-mapping, variant and gene prioritization. I.B.-H. and P.B. performed and supervised the gene network analysis. R.J.M.F. designed and A.M.H.Y. performed the isolation of human microglia from brain biopsies. N.K. performed the microglia eQTL mapping. A.B., T.J., D.J.G. and K.E. conceived and supervised the study.

Competing interests

J.Z.L. was an employee of Biogen at the time of the study and is now an employee of GSK. D.J.G. is an employee of Genomics PLC. T.J. is an employee of GSK. K.E. is an employee of BioMarin Pharmaceutical.

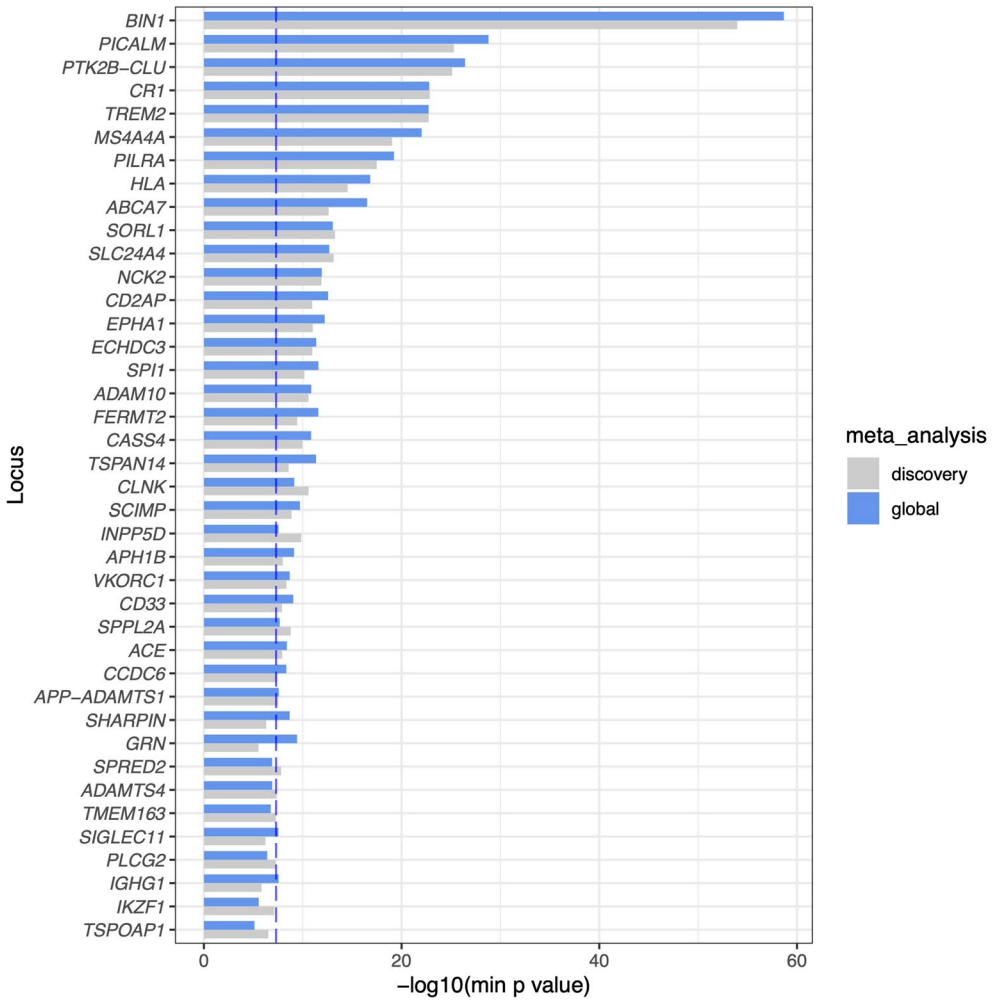
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-020-00776-w.

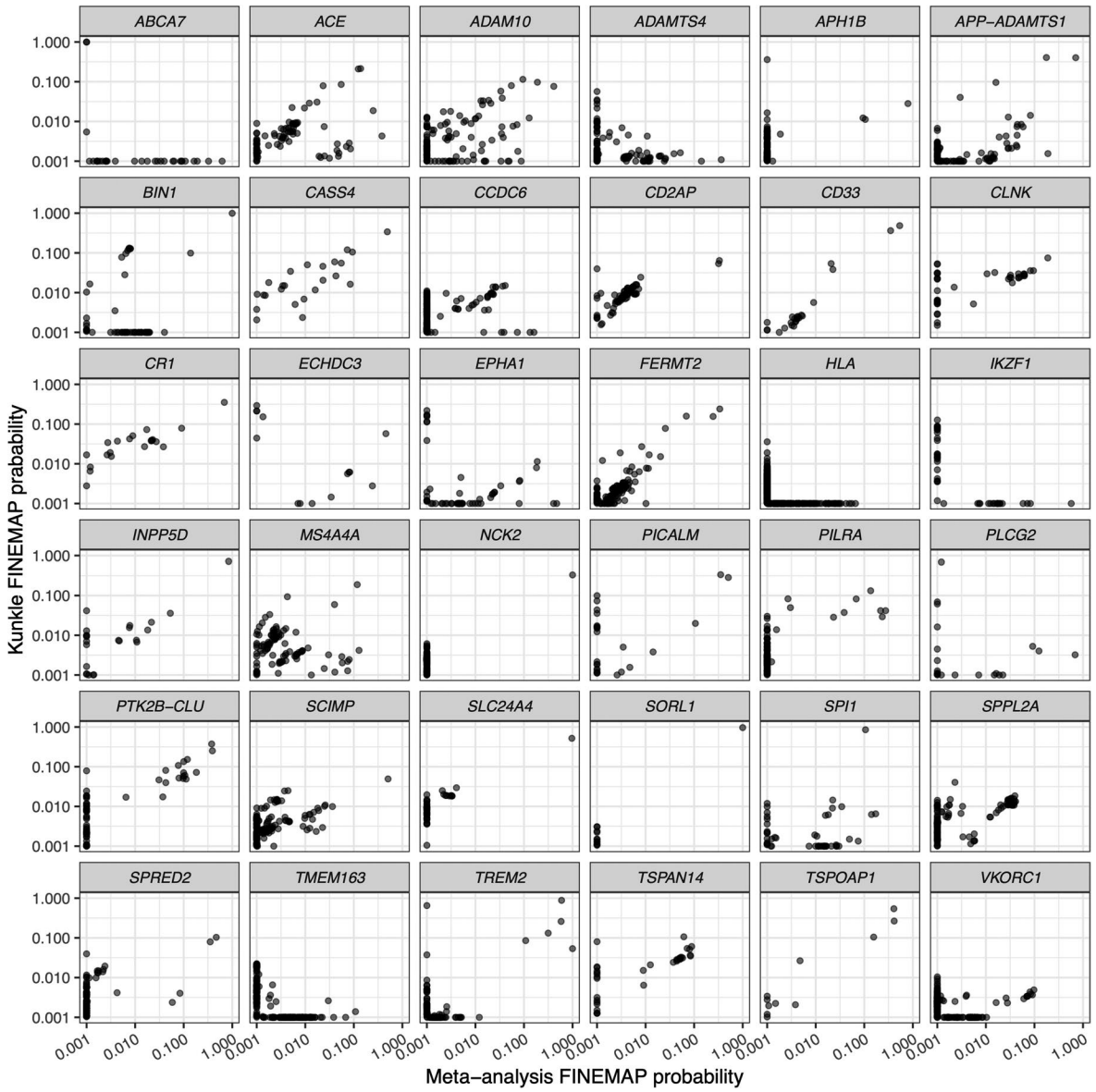
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-020-00776-w.

Correspondence and requests for materials should be addressed to J.S. or A.B.

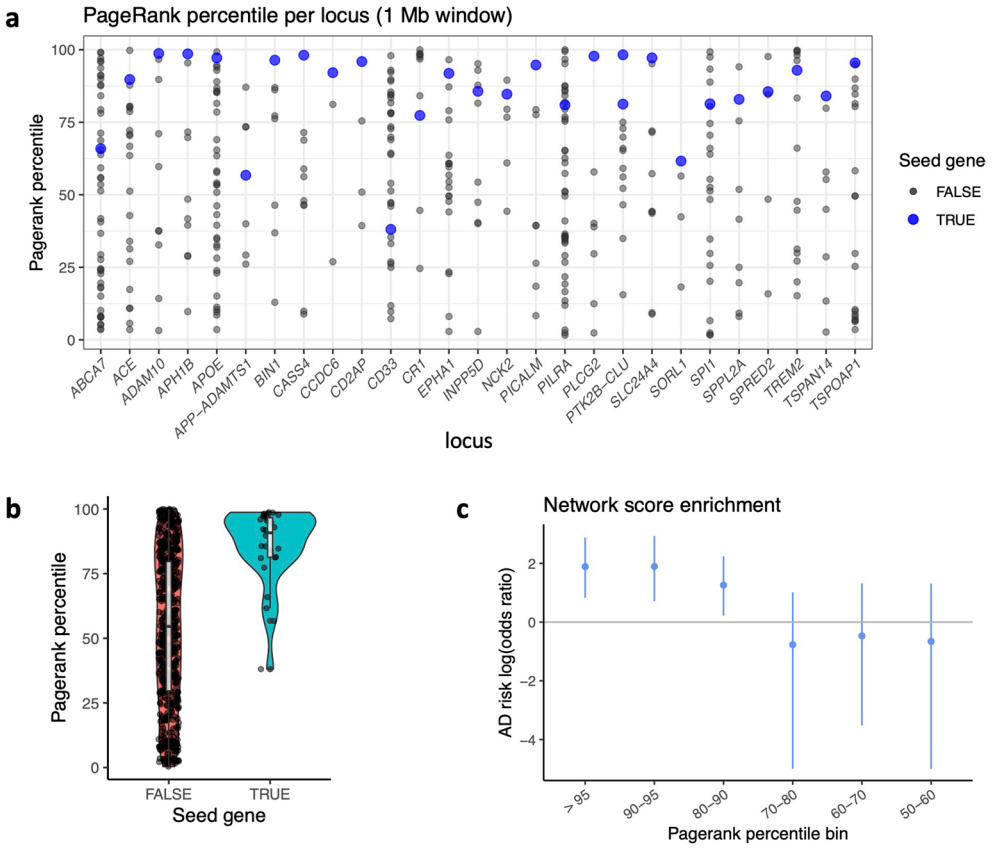
Reprints and permissions information is available at www.nature.com/reprints.



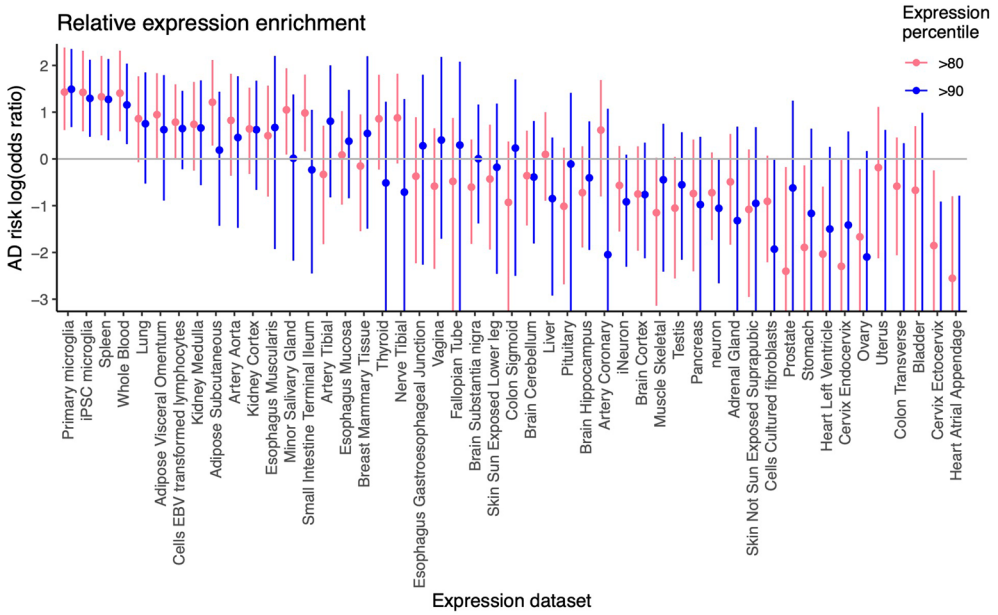
Extended Data Fig. 1 | Association of AD loci in discovery + replication ('global') meta-analysis. Association of AD loci in discovery + replication dataset ('global') meta-analysis. For most loci, association significance is increased in the global meta-analysis (blue bars) relative to the discovery analysis (grey bars). The dashed vertical line shows $P = 5 \times 10^{-8}$. *P*-values were computed by inverse variance weighted meta-analysis, and bars show the $-\log_{10}(P)$ for the SNP with minimum *P* value at the locus in either the discovery or global meta-analysis.



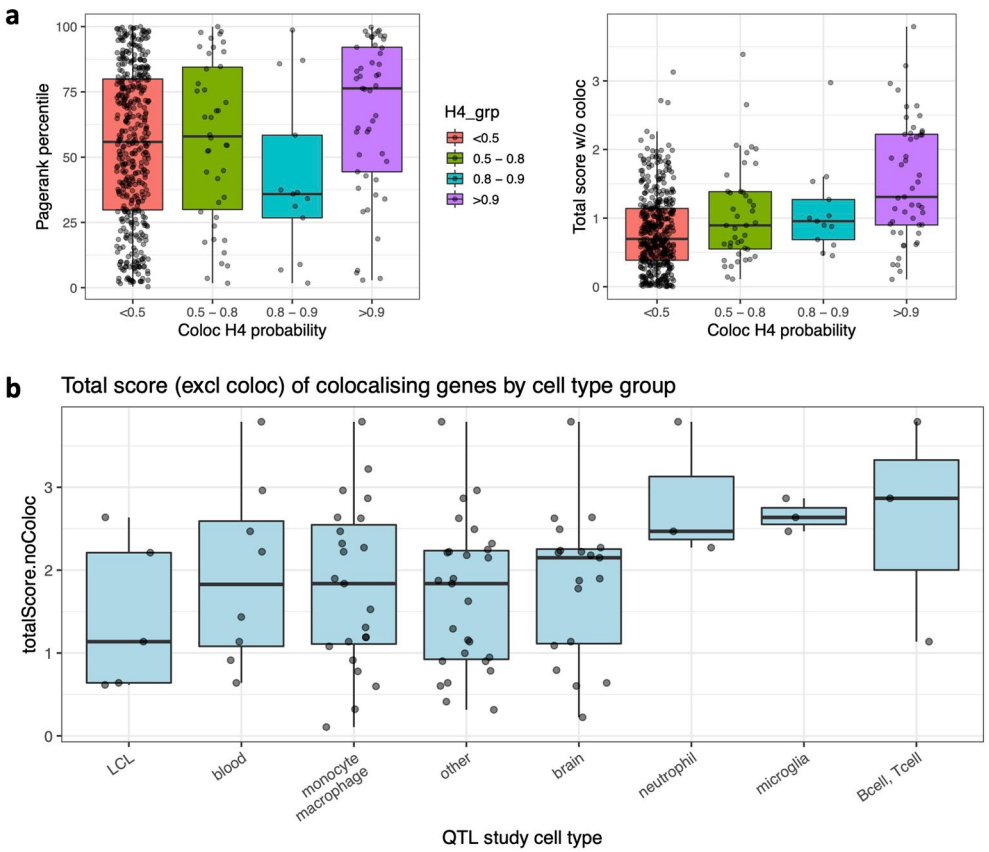
Extended Data Fig. 2 | Comparison of fine-mapping in the meta-analysis vs. Kunkle et al. Comparison of fine-mapping in the meta-analysis vs. Kunkle et al. Scatterplots showing, for each locus, SNP probabilities from FINEMAP applied to either the Kunkle et al. + UK Biobank meta-analysis (*x*-axis), or to only Kunkle et al. The number of causal variants at each locus was set to the number detected by GCTA in the meta-analysis. For most of the 36 loci, SNP probabilities are well correlated. For a few loci that are well powered in Kunkle et al., this is not the case, namely *ABCA7*, *EPHA1*, *ECHDC3*, and *HLA*. For these loci, fine-mapping results should be interpreted with caution. Six other loci are not well correlated (*ADAMTS4*, *APH1B*, *IKZF1*, *PLCG2*, *TMEM163*, and *VKORC1*), but these loci are poorly powered in Kunkle et al. (lead *P* values 2.1×10^{-6} to 2.1×10^{-3}).



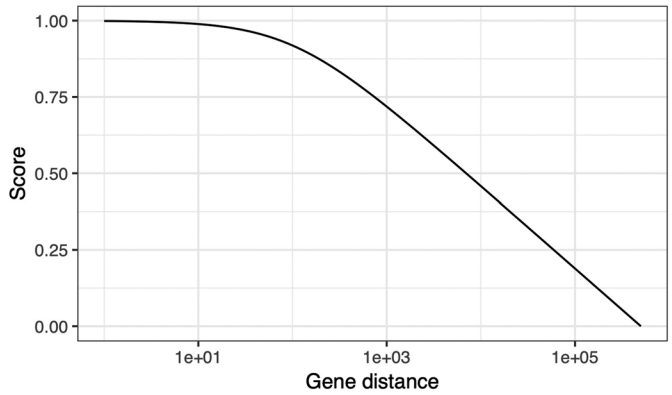
Extended Data Fig. 3 | Network enrichment. a, The Pagerank percentile of all genes (within 500 kb) at each AD GWAS locus containing a seed gene is shown, with seed genes highlighted in blue. **b**, A violin/boxplot shows that seed genes have a markedly higher network Pagerank percentile than remaining genes ($P = 2.4 \times 10^{-9}$, one-tailed Wilcoxon rank sum test). **c**, Log odds ratio enrichment of AD risk among SNPs nearest to genes with network Pagerank percentile in different bins, determined using fgwas (whiskers represent 95% confidence intervals).



Extended Data Fig. 4 | Gene expression enrichments. Expression enrichments for GTEx + microglia. Shown are the log odds ratio enrichments of AD risk among SNPs with relative gene expression in each tissue above the 80th (or 90th) percentile across tissues. Whiskers represent 95% confidence intervals determined by fgwas.



Extended Data Fig. 5 | Colocalization scores. a, Genes with maximum colocalization H4 probability >0.9 have higher Pagerank percentile (left boxplot) and higher total score (sum of the four non-coloc predictors, right boxplot) than do genes without colocalisation (<0.5). Genes with intermediate colocalisation evidence (bins 0.5 - 0.8 and 0.8 - 0.9) show little evidence of having higher scores by the other metrics. Based on this, we chose a maxColoc probability of 0.9 as the lower bound for our colocalization score. **b**, Boxplot of the total score (excluding coloc) for genes that have a colocalisation probability > 0.9 in at least one QTL dataset within each tissue group. The most significant difference is between totalScore for genes with microglial colocalizations vs. the genes with colocalization in 'other' tissues (non-immune GTEx tissues), but the difference is weak ($P = 0.041$, Wilcoxon rank sum test). In all cases, boxplots show the 25th, median, and 75th percentile of the distribution, with whiskers extending to the largest (and smallest) value no further than 1.5 times the interquartile range from the boxplot hinge.



Extended Data Fig. 6 | Gene distance score. The distance score assigned to genes near an AD GWAS peak, which decreases approximately linearly (past a distance of 1 kb) with increasing log-scaled distance up to 500 kb.

nature research

Corresponding author(s): Jeremy Schwartzentruber

Last updated by author(s): Dec 10, 2020

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 our Editorial Policies and the Editorial Policy Checklist.

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	Analysis code is located at http://github.com/jeremy37/AD_finemap Additional software used for analysis was: KING v2.0 Aberrant (Bellenguez et al. 2012) BOLT-LMM (Loh et al. 2015) FINEMAP v1.3 PAINTOR v3.1 GCTA v1.92.1 bwa 0.7.15 MACS2 FeatureCounts 1.5.3 RASQUAL

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

Summary statistics from the meta-analysis are available through the NHGRI-EBI GWAS Catalog under accession GCST90000652:
www.ebi.ac.uk/gwas/downloads/summary-statistics
eQTL Catalogue: www.ebi.ac.uk/eqtl
GTEx: www.gtexportal.org
Roadmap Epigenomics: www.roadmapepigenomics.org
DeepSEA: deepsea.princeton.edu
SpliceAI: github.com/Illumina/SpliceAI
GERP: hgdownload.cse.ucsc.edu/gbdb/hg19/bbi/All_hg19_RS.bw
PhyloP: hgdownload.cse.ucsc.edu/goldenpath/hg19/phyloP100way
PhastCons: hgdownload.cse.ucsc.edu/goldenpath/hg19/phastCons100way
Brain eQTL meta-analysis summary statistics: www.synapse.org/#!Synapse:syn16984815
Primary microglia eQTL summary statistics, EGA Accession ID: EGAD00001005736
Primary microglia ATAC-seq, dbGaP Study Accession: phs001373.v1.p1
Allen Brain Institute: portal.brain-map.org/atlases-and-data/rnaseq
IntAct database: www.ebi.ac.uk/intact
BioGRID database: thebiogrid.org
STRING database: string-db.org

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	For the proxy-GWAS, all available UK Biobank samples were used, apart from exclusion of first-degree relatives, and these numbers are specified in the methods section.
Data exclusions	As described in the methods section: For the proxy-GWAS, one member of each first-degree relative pair in UK Biobank was excluded. Also, participants who answered "Do not know" or "Prefer not to answer" regarding whether they have a first-degree relative affected by dementia were excluded from analyses. For fine-mapping analyses we excluded rare variants (< 0.2% frequency, 10% of all variants) as well as those failing imputation filters (INFO < 0.85, 3% of variants) or study heterogeneity filters (p_het < 0.001, 0.4% of variants).
Replication	We searched for nominal replication of our genome-wide significant signals in two other AD studies: the Gr@ace cohort and FinnGen v3.
Randomization	There were no experimental treatments and hence no randomization.
Blinding	The investigators were not involved in data collection for UK Biobank, and all individual-level data are anonymized. There is no group allocation, and so blinding is not relevant to this 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

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

Methods

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