medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

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

Jeremy Schwartzentruber[1,2*], Sarah Cooper[2,3], Jimmy Z Liu[4], Inigo Barrio-Hernandez[1,2], Erica Bello[2,3], Natsuhiko Kumasaka[3], Toby Johnson[5], Karol Estrada[6], Daniel J. Gaffney[2,3,7], Pedro Beltrao[1,2], Andrew Bassett[2,3*]

*Corresponding:
Jeremy Schwartzentruber (jeremys@ebi.ac.uk)
Andrew Bassett (ab42@sanger.ac.uk)

Affiliations:

1. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), 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, 02142, USA
5. Target Sciences-R&D, GSK Medicines Research Centre, Stevenage, UK
6. BioMarin Pharmaceutical, San Rafael, CA 94901, USA
7. Genomics Plc, Oxford, OX1 1JD, UK

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Abstract

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

Introduction

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 GWAS of diagnosed late-onset Alzheimer's disease (AD) with GWAX for family history of AD in the UK Biobank[2,3], and reported a total of 12 novel disease-associated genomic loci. However, the causal genetic variants and genes which influence AD risk at these and previously discovered loci have only been clearly identified in a few cases. Discovering causal variants has led to deeper insight into molecular mechanisms of multiple diseases, including obesity[4], schizophrenia[5], and inflammatory bowel disease[6]. For AD, known causal variants include the ε4 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[7]. In addition, likely causal rare nonsynonymous variants have been discovered in *TREM2*[8], *PLCG2* and *ABI3*[9]. These findings have strengthened support for a causal role of microglial activation in AD.

Although non-synonymous variants are highly enriched in trait associations, most human trait-associated variants do not alter protein-coding sequence and are thought to mediate their effects via altered gene expression, which is likely to occur in a cell type-dependent manner. A growing number of studies have mapped genetic variants affecting gene expression traits, known as expression quantitative trait loci (eQTLs), in diverse tissues or

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

sorted cell types[10,11]. While it has become common to integrate GWAS results with eQTLs, this is often limited to a small number of datasets thought to be relevant.

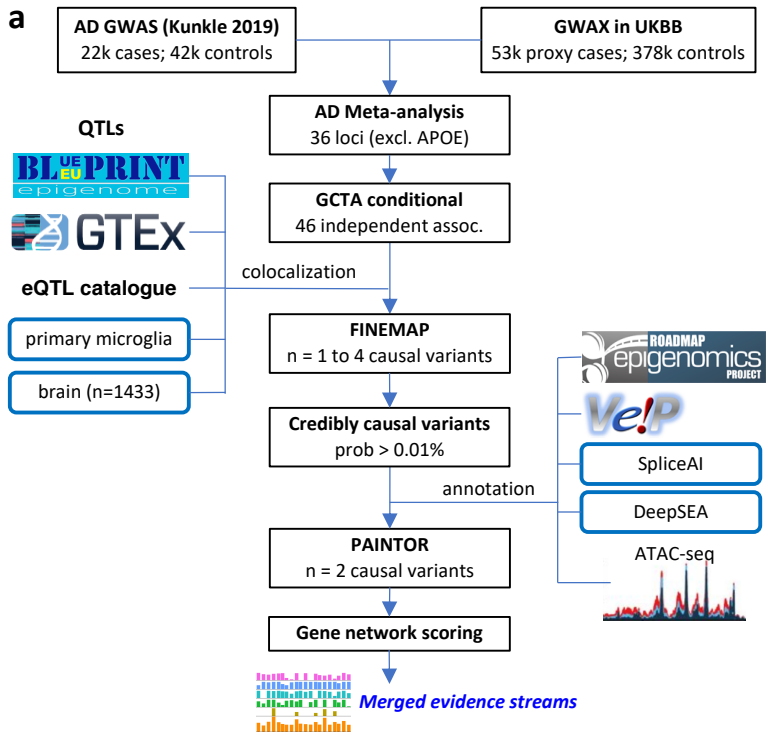
To identify putative causal genetic variants for AD, we performed a meta-analysis of GWAX in the UK Biobank with the latest GWAS for diagnosed AD[12], followed by fine-mapping using three alternative methods. Notably, this updated GWAS tested more genetic variants than the Lambert et al. study[13] used in the meta-analyses by Jansen[2,3] and Marioni[2] (11.5 vs. 7.1 million). The increased power from our meta-analysis enabled us to discover four additional AD risk loci at genome-wide significance, and the higher density genotype imputation identified new candidate causal variants at both novel and established risk loci. We also performed statistical colocalisation 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 find that multiple lines of evidence, including colocalisation, tissue- or cell type-specific expression and prioritization using information propagation in gene networks, converge on a set of likely causal AD genes.

Results

Meta-analysis discovers 37 loci associated with Alzheimer's disease risk

We performed a GWAX in the UK Biobank for family history of AD, based on 53,042 unique individuals who were either diagnosed with AD or who reported at least one first-degree relative (parent or sibling) having dementia, and 355,900 controls. This identified 13 risk loci at genome-wide significance ($p < 5x10^{-8}$), 10 of which have been reported previously. Three novel loci were located near genes *NCK2*, *PRL*, and *FAM135B*. Notably, PRL has been reported as a CSF biomarker of AD[15]. We next did a fixed-effects meta-analysis of these GWAX results with the Kunkle et al. stage 1 GWAS meta-analysis of 21,982 cases with diagnosed AD and 41,944 controls[12], across 10,687,126 overlapping variants (Figure 1). This revealed 34 AD risk loci ($p < 5x10^{-8}$), 22 of which were reported in the Kunkle et al. study, while 8 others were reported in either Jansen et al.[3] or Marioni et al[2]. Four loci were novel, located near genes *NCK2*, *TSPAN14*, *SPRED2*, and *CCDC6*. Notably, the *PRL* and *FAM135B* regions showed no evidence of association in Kunkle et al. ($p > 0.1$), and hence were not significant in meta-analysis. Three additional loci were found at suggestive significance ($p < 5x10^{-7}$) in the meta-analysis, near genes *IKZF1*, *TSPOAP1*, and *TMEM163*. We included these loci in our follow-up analyses, for a total of 37 loci (Figure 1, Supplementary Table 1).

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .



c

Locus	# signals	Lead p value
APOE	see note	~0
TREM2	3	1.83E-23
BIN1	2	1.10E-54
PTK2B-CLU	2	7.71E-26
ABCA7	2	2.41E-13
NCK2	2	1.28E-12
EPHA1	2	9.63E-12
ADAM10	2	2.67E-11
ACE	2	1.21E-08
APP-ADAMTS1	2	3.09E-08
PLCG2*	1	5.46E-08
CLNK	1	2.59E-11
SCIMP	1	1.35E-09
SPPL2A	1	1.74E-09
TSPAN14	1	2.74E-09
VKORC1	1	4.47E-09
APH1B	1	1.05E-08
SPRED2	1	1.54E-08
CCDC6	1	3.80E-08
ADAMTS4	1	4.30E-08
TMEM163	1	5.24E-08
IKZF1	1	7.68E-08
TSPOAP1	1	3.07E-07

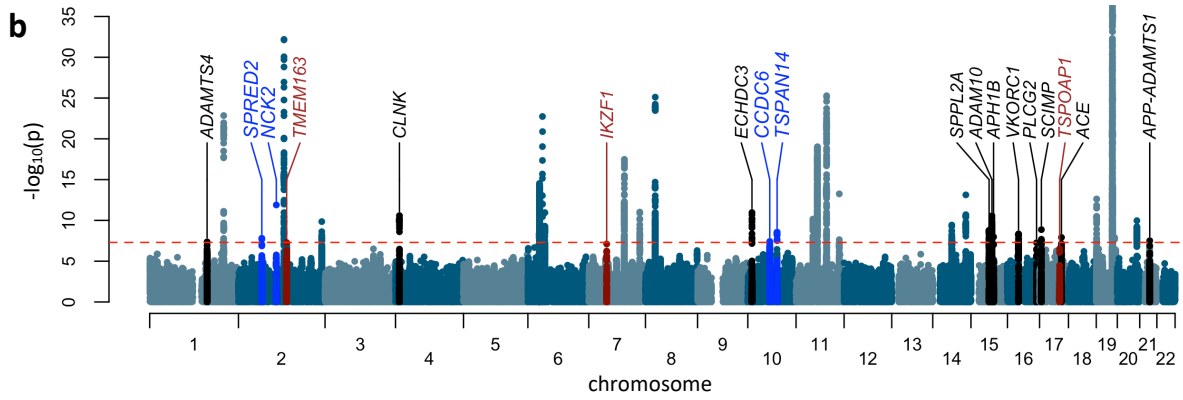


Figure 1: Analysis overview. (a) Summary of AD meta-analysis and data processing steps. (b) Manhattan plot of the meta-analysis of GWAS for diagnosed AD and our GWAX in UK Biobank. Novel genome-wide significant loci are labelled in blue, sub-threshold loci in red, and recently discovered loci[2,3,12] replicated in our analysis in black. (c) The number of independent signals at each locus which is either recently discovered or which has more than one signal. * The *PLCG2* locus was significant ($p < 5x10^{-8}$) when including Kunkle stage 3 SNPs. Conditional analyses were not done at *APOE* due to the strength of the signal.

Next, we applied stepwise conditioning using GCTA[16], with linkage disequilibrium (LD) determined from UK Biobank samples, to identify independent signals at the discovered loci. We excluded the *APOE* locus from conditional analyses and fine-mapping, because the strength of the association made these analyses unreliable (see methods). Apart from *APOE*, 9 loci had two independent signals, while the *TREM2* locus had three signals (Figure 1c). Interestingly, a number of the loci discovered recently[2,3,12] had multiple signals; specifically, *NCK2*, *EPHA1*, *ADAM10*, *ACE*, and *APP-ADAMTS1*. To extract insight from

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

both the new and established AD GWAS discoveries, we performed comprehensive colocalisation, annotation, fine-mapping and network analyses to identify causal genes and variants (Fig 1a).

Colocalisation between AD risk loci and gene expression traits

To identify genes whose expression may be altered by risk variants, we performed statistical colocalisation[17] between each of 36 risk loci (excluding *APOE*) and a set of 109 eQTL datasets representing a wide variety of tissues, cell types and conditions (Figure 2, Supplementary Table 2). The eQTL datasets include a study of primary microglia from 93

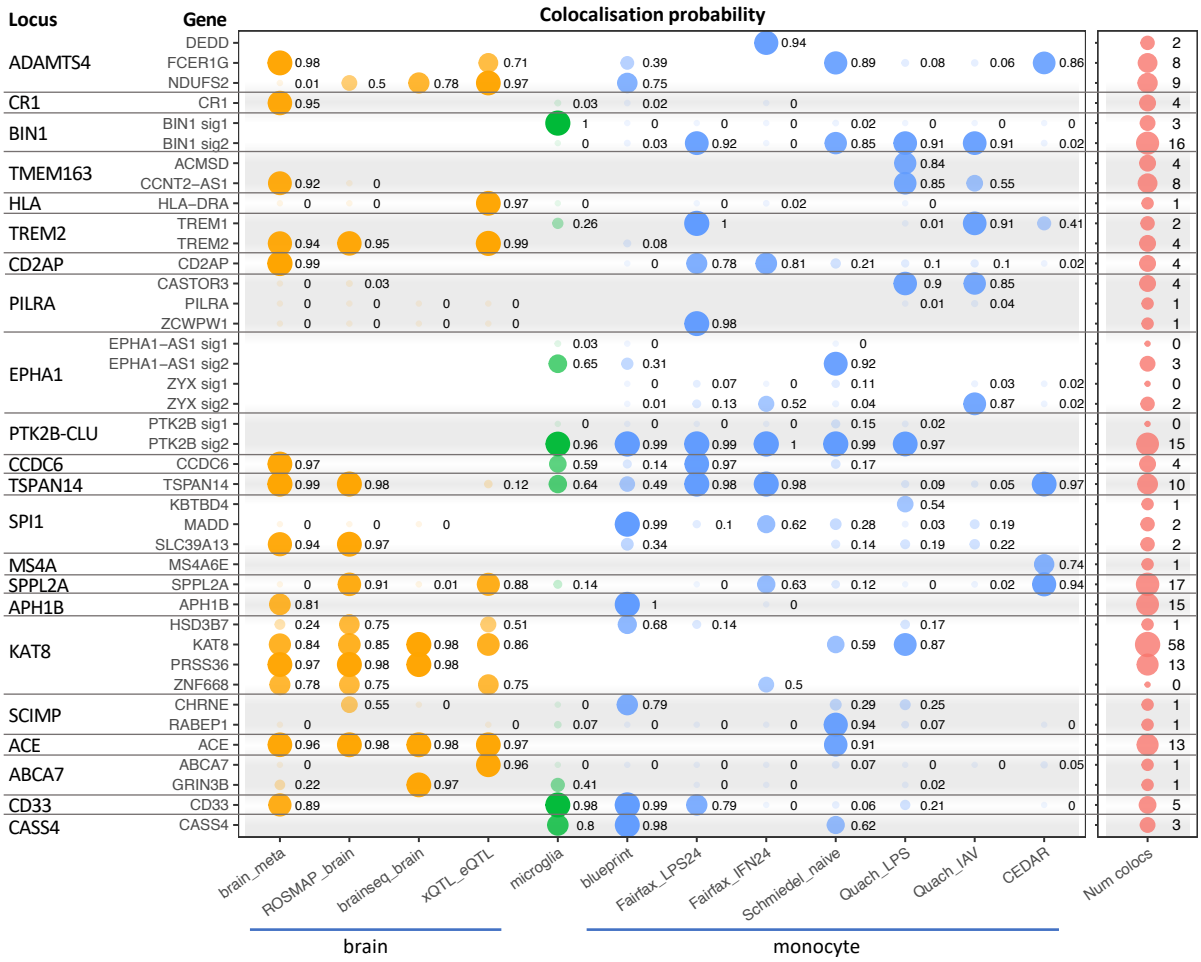


Figure 2: Colocalisation with eQTLs. For genes with the top overall colocalisation scores across AD risk loci, the colocalisation probability (H4) is shown for selected brain, microglia, and monocyte eQTL datasets. For three loci with multiple signals (*BIN1*, *EPHA1*, *PTK2B-CLU*), scores are shown separately for the conditionally independent signals. The last column shows, for each gene, the number of eQTL datasets with a colocalisation probability above 0.8 (Supplementary Tables 2-3).

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

brain surgery donors[14], a meta-analysis of 1433 brain cortex samples[18], as well as 49 tissues from the genotype-tissue expression project (GTEx) final release[10], and 57 eQTL datasets uniformly reprocessed as part of the eQTL catalogue[11]. 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[19–35].

Some studies using colocalisation have suggested that there is relatively limited overlap between GWAS associations and gene expression QTLs above that expected by chance[6,36]. A possible reason is that colocalisation analyses can suffer from a lack of 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 a pair of studies (e.g. GWAS and eQTL study), colocalisation may not be detected[17]. Second, differences in LD patterns in a pair of studies can reduce the likelihood of a positive colocalisation. Third, relatively low power in either study can further reduce the colocalisation probability. To mitigate the first effect, we performed colocalisations separately for each conditionally independent AD signal, to model the case where not all causal variants are shared, as well as for the main AD signal at each locus. Problems relating to power and LD mismatch are partially mitigated by our use of a large number of the most highly-powered eQTL datasets currently available.

Across the 36 loci, we found 391 colocalisations with at least 80% probability of a shared causal variant between AD and eQTL, representing 80 distinct genes at 27 loci (Supplementary Tables 3-4). The genes implicated by colocalisation include many which have alternative lines of evidence for roles in AD, such as *PTK2B*[37,38], *BIN1*[39,40], *PILRA*[41], *CD33*[42,43], and *TREM2*[44,45], as well as novel candidates including *FCER1G*, *TSPAN14*, *APH1B*, and *ACE*. However, the presence of multiple genes with colocalisation evidence within individual loci suggests that additional lines of evidence are important for prioritizing relevant genes.

Due to the large number of tissue datasets and colocalisation tests performed, we hypothesized that it would be important to upweight colocalisations in “relevant” tissues, as well as to accumulate colocalisation information across datasets. We therefore developed a weighted score which accumulates towards a maximum of 1.0 (inspired by scoring systems in STRING[46] and Open Targets[47]), with higher weight on colocalisations in microglia, brain, and immune cell types than in other tissues (see methods). We compared this with a score obtained by taking the maximum colocalisation probability across all datasets.

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

To evaluate the ability of these scores to identify relevant AD genes, we considered the top 40 genes at our AD loci that were prioritized via other lines of evidence (namely, expression, distance, coding variant changes, and network score, described further below; genes in Supplementary Table 5). The top 80 genes by weighted coloc score retrieved 18 of the 40 genes. Surprisingly, the top 80 genes by maximum coloc score retrieved 20 genes (16 overlapping; Supplementary Figure 1), and we therefore used this score in subsequent gene prioritization. These results suggest that strong colocalisations in tissues not thought to be relevant can still be informative for identifying likely causal genes, and that with our approach, upweighting "relevant" QTL datasets showed no benefit for gene prioritization.

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. We sought to identify candidate causal variants using three distinct fine-mapping methods. First, we used the WTCCC Bayesian fine-mapping method[48], which assumes a single causal variant, on each conditionally independent signal. Second, we used FINEMAP[49] at each locus, specifying the number of independent signals determined by GCTA as the maximum number of causal variants per locus. Third, we used PAINTOR[50], a method which estimates enrichments in functional genomic annotations to obtain a posterior probability of causality for each variant based on its annotations, and which also can account for multiple causal variants. For computational feasibility with PAINTOR, we only considered variants with at least 0.01% causal probability as determined by FINEMAP.

We used 43 annotations individually as input to PAINTOR (Supplementary Table 6); these included ATAC-seq peaks from primary microglia[51] or iPSC-derived macrophages[52], DNase peaks from cell type groups in the Roadmap Epigenomics project[53], variant consequence annotations[54] and evolutionary conservation[55] (Figure 3b). 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 a 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 tissues provided the highest model likelihood, as these peaks covered 16% of SNPs, despite a lower 6.4-fold enrichment. Variants with a nonzero score from the SpliceAI method, which predicts changes to gene splicing, were also highly enriched (9.3-fold), while variants with top DeepSEA scores were more modestly enriched.

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

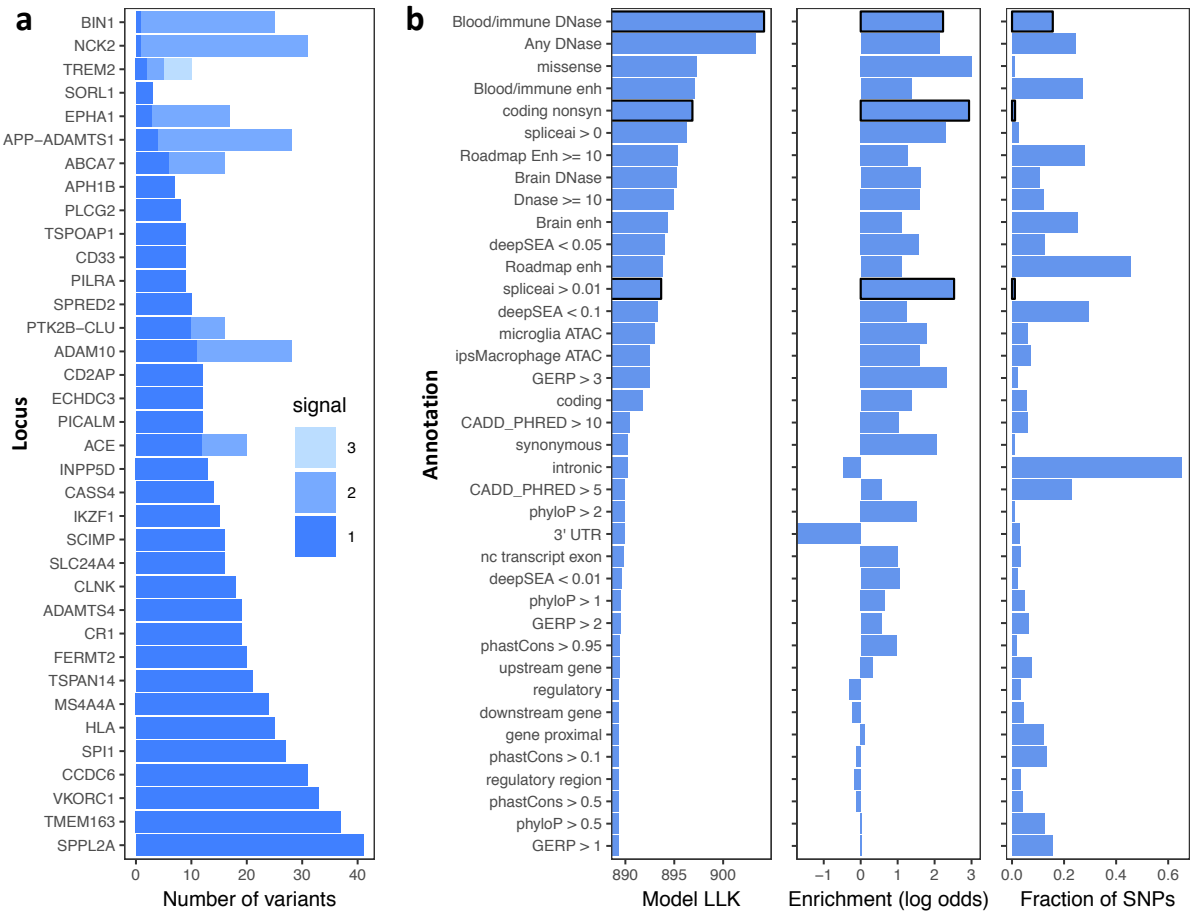


Figure 3: Fine-mapping summary. (a) Number of variants with mean causal probability > 1% for each independent signal. Variant counts for independent signals are shown in different shades. (b) PAINTOR outputs, showing (left) log-likelihood (LLK) of model for each individual annotation; (middle) log-odds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

We next built a multi-annotation model in PAINTOR (v3.1) 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 SpliceAI score greater than 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 a mean causal probability above 50% across the fine-mapping methods, and 79 further variants with probabilities from 10 - 50% (Table 1 and Supplementary Table 7). These include SNPs near established AD risk genes, such as rs6733839 ~20 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[40]. High-

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Gene	SNP	P value	Beta	Effect allele	Allele freq	SNP prob	SpliceAI	DeepSEA	Note	Refs
ADAMTS4	rs2070902	1.64E-06	-0.0522	T	0.2580	0.384	0.107	0.140	Intronic in candidate gene FCER1G, with predicted splicing change.	
ADAMTS4	rs4575098	4.30E-08	0.0609	A	0.2350	0.339		0.033	3' UTR of ADAMTS4, open chromatin	2
SPRED2	rs268120	2.08E-08	0.0612	A	0.2502	0.556		0.033	Strong DNase peak, predicted by DeepSEA to decrease.	
NCK2	rs143080277	1.28E-12	-0.5205	T	0.9957	1.000		0.086	Enhancer (Roadmap)	
BIN1	rs6733839	1.10E-54	0.1556	T	0.3915	0.998		0.027	Microglia ATAC peak. DeepSEA predicts decreased DNaseHS.	14,40
INPP5D	rs10933431	1.41E-10	0.0770	C	0.7817	0.833		0.022		60
PILRA	rs1859788	3.28E-18	-0.0897	A	0.3206	0.601	0.008	0.041	Known PILRA missense G78R.	41
ECHDC3	rs7920721	1.08E-11	-0.0669	A	0.6195	0.641		0.026	DNase peak. DeepSEA predicts changed binding of USF, Max, Myc.	61,62
TSPAN14	rs1870137	2.93E-09	-0.0702	C	0.2056	0.097		0.007	Top DeepSEA variant, predicting decreased binding of HNF4, FOXA1, SP1.	
TSPAN14	rs1870138	4.51E-09	-0.0693	A	0.2057	0.068		0.004	Highlighted in text; predicted loss of TAL1 binding.	
SORL1	rs11218343	5.59E-14	0.1864	T	0.9630	1.000		0.209		63
SORL1	rs2298813	1.52E-04	0.0850	A	0.0470	0.451	0.054	0.003	Secondary assoc. Missense; also top DeepSEA variant.	
APH1B	rs117618017	1.05E-08	0.0850	T	0.1395	0.895	0.007	0.019	Highlighted in text; missense Thr27Ile.	64
PLCG2	rs12444183	5.46E-08	-0.0533	A	0.3830	0.686		0.220	Near promoter of ncRNA AC099524.1, with strong microglia coloc.	2
PLCG2	rs72824905	6.35E-06	0.2703	C	0.9924	0.492	0.018	0.006	Secondary association; known missense Pro522Arg. Top DeepSEA score.	9
TSPOAP1	rs2632516	3.12E-07	-0.0493	C	0.4426	0.412		0.126	Overlaps ncRNA containing mir-142, important for hematopoietic dev't.	62,65
TSPOAP1	rs2526377	8.45E-07	0.0477	A	0.5579	0.169		0.006	Top DeepSEA variant (decreased DNaseHS) in microglial ATAC peak.	66
ACE	rs4311	1.21E-08	-0.0543	T	0.4704	0.490	0.126	0.053	Strong predicted splicing change.	67,68
ACE	rs3730025	2.58E-07	-0.1994	A	0.9828	0.416	0.002	0.021	Secondary association; low-freq missense Tyr244Cys.	
ABCA7	rs12151021	2.41E-13	0.0773	A	0.3258	0.713	0.013	0.312	Lead ABCA7 variant.	
ABCA7	rs4147918	7.63E-07	0.1201	A	0.9587	0.552	0.071	0.045	Seondary association; missense Gln905Arg; predicted splicing change.	69
CD33	rs12459419	2.02E-08	-0.0576	T	0.3256	0.662	0.001	0.070	Known missense Ala14Val; strong splicing QTL.	7
CASS4	rs6014724	1.07E-10	0.1094	A	0.9122	0.548		0.083	Lead CASS4 variant.	
CASS4	rs17462136	1.01E-09	-0.1038	C	0.0872	0.067		0.001	5' UTR of CASS4; global top DeepSEA variant predicting decreased TF binding.	
ADAMTS1	rs2830489	3.09E-08	-0.0590	T	0.2749	0.718		0.077	Lead variant near ADAMTS1.	

Table 1: Top candidate variants. A selected list of the most likely causal variants across loci, based on a combination SNP fine-mapping probabilities and annotations. Column 'SNP prob' 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 significance above expectation for chromatin feature changes, as well as evolutionary conservation, with lower values more significant. References for specific SNPs are shown[2,7,9,14,40,41,60–69].

confidence variants also include a well-known missense SNP in *PILRA*[41], and a splice-altering missense SNP in *CD33*[7]. Missense SNP rs4147918 in *ABCA7* had 55% causal probability, and *ABCA7* harbored 5 further missense SNPs with FINEMAP probabilities greater than 0.01%, at varying allele frequencies. Notably, rs4147918 as well as 6 other variants within *ABCA7*, including the lead SNP rs12151021, had positive SpliceAI scores for predicted changes to gene splicing. 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].

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

A number of newly identified AD risk genes had high-confidence fine-mapped variants. These include the *NCK2* rare intronic SNP rs143080277 (>99% probability, MAF 0.4%), *APH1B* missense SNP rs117618017 (90% probability), rs2830489 at the *APP-ADAMTS1* locus (72% probability), rs61182333 intronic in *SCIMP* (61% probability), and rs268120 intronic in *SPRED2* (56% probability).

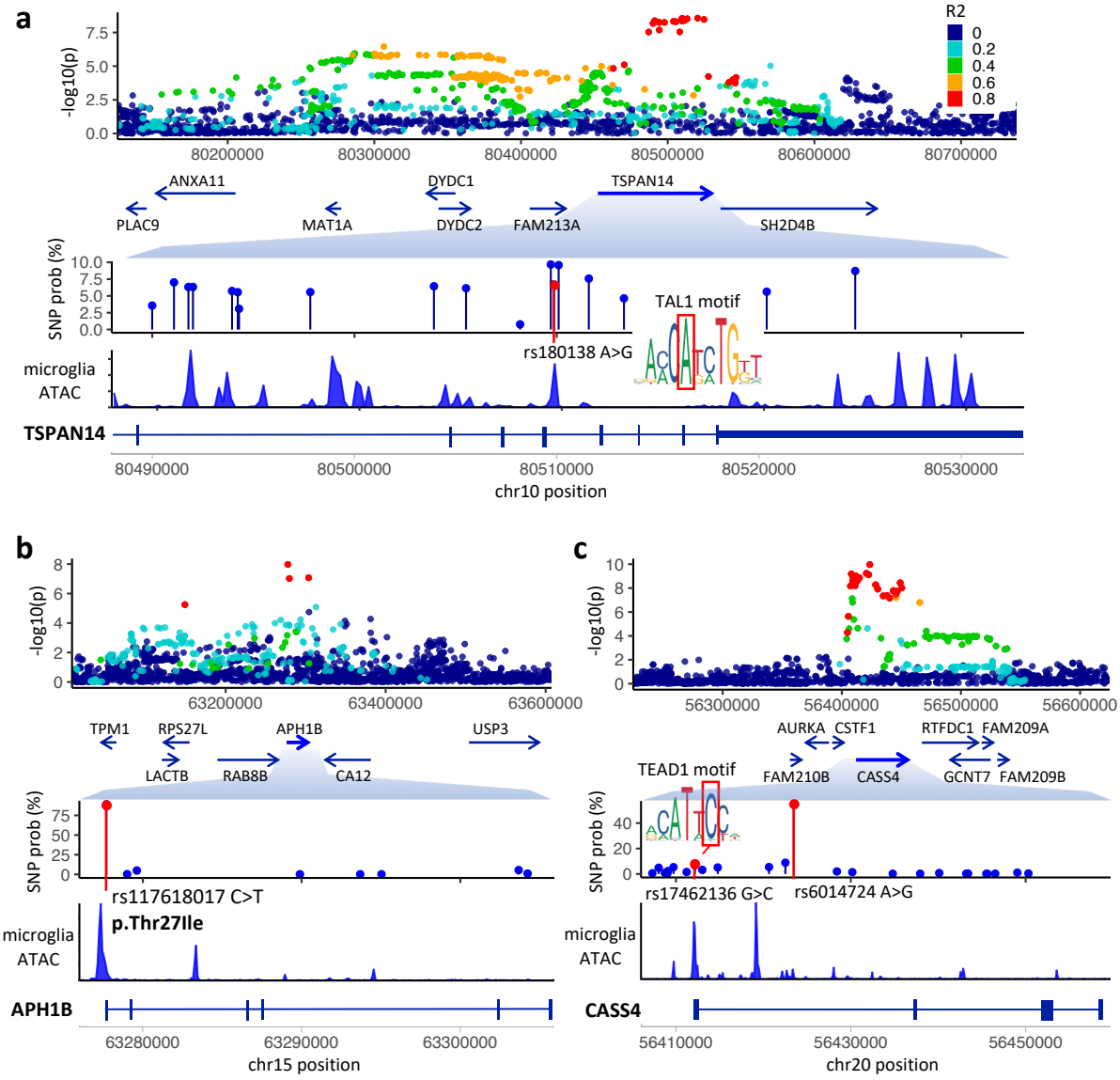


Figure 4: Fine-mapped variants. (a) SNP rs1870138 in an intron of *TSPAN14* disrupts an invariant position of a TAL1 motif. (b) Missense SNP rs117618017 in exon 1 of *APH1B*. (c) SNP rs17462136 in the 5' UTR of *CASS4* introduces a TEAD1 motif.

Annotation-based fine-mapping highlighted a number of candidate causal variants, which were not always the highest probability SNPs at the locus (Figure 4). Within *TSPAN14*, rs1870137 and rs1870138 reside within a DNase hypersensitivity peak found broadly across tissues, which is also an ATAC peak in microglia. Of these, rs1870138 lies at the centre of a

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

ChIP-seq peak for binding of multiple transcription factors, including FOS/JUN and GATA1, and is within a FOS/JUN motif, albeit at a relatively low information content position. However, the alt allele rs1870138-G alters an invariant position of a binding motif for *TAL1*, a gene highly expressed in microglia, and which is a binding partner for GATA1. The AD risk allele, rs1870138-G, is also associated with increased monocyte count[70] and increased risk for inflammatory bowel disease[71], and in both cases is among the top associated variants. Notably, the AD signal in the region colocalises with both an eQTL and a splicing QTL for *TSPAN14* in multiple datasets, and rs1870138-G associates with higher *TSPAN14* expression in brain and in microglia, but with lower expression in some GTEx tissues.

Missense SNP rs117618017 in exon 1 of *APH1B* (T27I) is the likely single causal variant at its locus, with fine-mapping probability of 90% (Figure 4b). APH1B is a component of the gamma-secretase complex, other members of which (*PSEN1*, *PSEN2*) have rare variants associated with early-onset AD[72]. Interestingly, the AD signal colocalises with an *APH1B* eQTL in monocytes, neutrophils and T-cells, as well as numerous GTEx tissues, and the rs117618017-T allele associates with higher AD risk and higher *APH1B* expression across datasets. rs117618017-T introduces a motif for transcriptional regulator YY1, and is predicted by DeepSEA to increase YY1 binding in multiple 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 colocalises with an eQTL for *CASS4* in Blueprint monocytes and in GTEx whole blood, as well as in fibroblasts. While lead SNP rs6014724 (55% probability), intronic in *CASS4*, shows no evidence of transcription factor (TF) binding in ENCODE data, rs17462136 (7% probability) lies in a region of dense TF binding in the 5' UTR of *CASS4* (Figure 4c). The nucleotide position is highly conserved (GERP score 3.46), overlaps an ATAC peak in microglia, and the rs17462136-C allele introduces a TEAD1 binding motif, making it the strongest functional candidate SNP. 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 candidate genes at each AD locus (Supplementary Table 8), based on a mix of our other evidence sources as well as literature reports, and used these as seed genes

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

similar to the approach used in the priority index for drug discovery[73]. 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 page rank algorithm (see methods). The "networkScore" for a gene thus represent the degree to which the gene is supported by its interaction with top AD candidate genes across all other loci, unbiased by any locus-specific features.

Across AD loci, likely candidate genes were highly enriched for having high network-based gene scores (Wilcoxon rank sum test, $p = 5x10^{-14}$; Supplementary Figure 2). Notably, at our four novel AD loci, the nearest gene (*NCK2*, *TSPAN14*, *SPRED2*, *CCDC6*) was the highest-scoring gene among those within 200 kb, and in each case was one of the top two highest-scoring genes within 500 kb. Many established or recently discovered AD genes were also the top gene within 500 kb by network score, including *ACE*, *CASS4*, *CD2AP*, *PICALM*, *PLCG2*, *PTK2B*, and *TREM2*. 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]. At the *ECHDC3* locus, both *USP6NL* and *CELF2* had high network scores, while at the *EPHA1* locus, *ZYX* was the top scoring gene. Interestingly, AD candidate genes *ABCA7* and *CR1* had only modest network scores, suggesting a need to integrate across independent lines of evidence to prioritize genes.

Genes highly ranked by network propagation also include many outside of genome-wide significant AD loci (Supplementary Table 9). Consistent with their involvement in AD, genes ranked in the top 500 by network score tended to have SNPs with lower p values nearby (within 10 kb) than did remaining genes (Wilcoxon rank sum test, $p = 7x10^{-7}$), suggesting that there remain numerous AD loci to be discovered with larger GWAS sample sizes. Top network-ranked genes include *LILRB2* (nearby SNP $p = 9.8x10^{-6}$), a leukocyte immunoglobulin-like receptor that recognizes multiple HLA alleles, and which may also be involved in amyloid-beta fibril growth[74]; *ABCA1* (SNP $p = 4x10^{-6}$), involved in phospholipid transfer to apolipoproteins and previously associated with AD[75]; *SREBF1* (SNP $p = 2x10^{-6}$), required for lipid homeostasis; *AGRN* (SNP $p = 4x10^{-6}$), involved in synapse formation in mature hippocampal neurons; and *CD19* (SNP $p = 1x10^{-5}$), an antigen coreceptor on B-lymphocytes. 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 immune response, phagocytosis, Ephrin signaling, and complement activation (Supplementary Table 10).

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Integrative gene prioritization from five lines of evidence

We developed a comprehensive gene prioritization score, which incorporates quantitative information from multiple evidence sources, unlike binary indicator scores that have been used by others for GWAS gene prioritization[12,76] (Figure 5, Supplementary Table 11). The five lines of evidence include gene distance to lead SNPs, expression level in either microglia or brain tissue, colocalisation score, network score, and the sum of fine-mapped probability for any coding SNPs within a gene (see methods for details).

To incorporate gene distance, we used a function that decays from 1 to 0 with increasing log-scaled distance up to 500 kb (Supplementary Figure 3). Although it is clear that long-distance gene regulation can occur, recent evidence from both eQTLs[77] and metabolite GWAS[78] suggests that genomic distance from the association peak is a strong predictor of causal target genes. For gene expression, we considered either the expression level of genes in microglia or brain, or the specificity of each gene's expression in these cells / tissues relative to all GTEx tissues. We found that specificity of expression in microglia or brain was at least as informative as the absolute level of expression in these tissues, and that incorporating expression level did not improve a score based solely on expression specificity (Supplementary Figure 4). Hence, we used only expression specificity in our overall prioritization.

The comprehensive prioritization score identified the majority of AD candidate genes previously suggested as causal (Figure 5), suggesting that our evidence sources complement each other. Indeed, although the network score is independent of locus-specific features, gene ranks based on network scores within each locus were highly correlated with gene ranks based on the other four evidence sources (Spearman rho=0.47; $p < 2.2 \times 10^{-16}$). Exemplifying the importance of integrating genetic evidence, *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 colocalisation evidence.

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

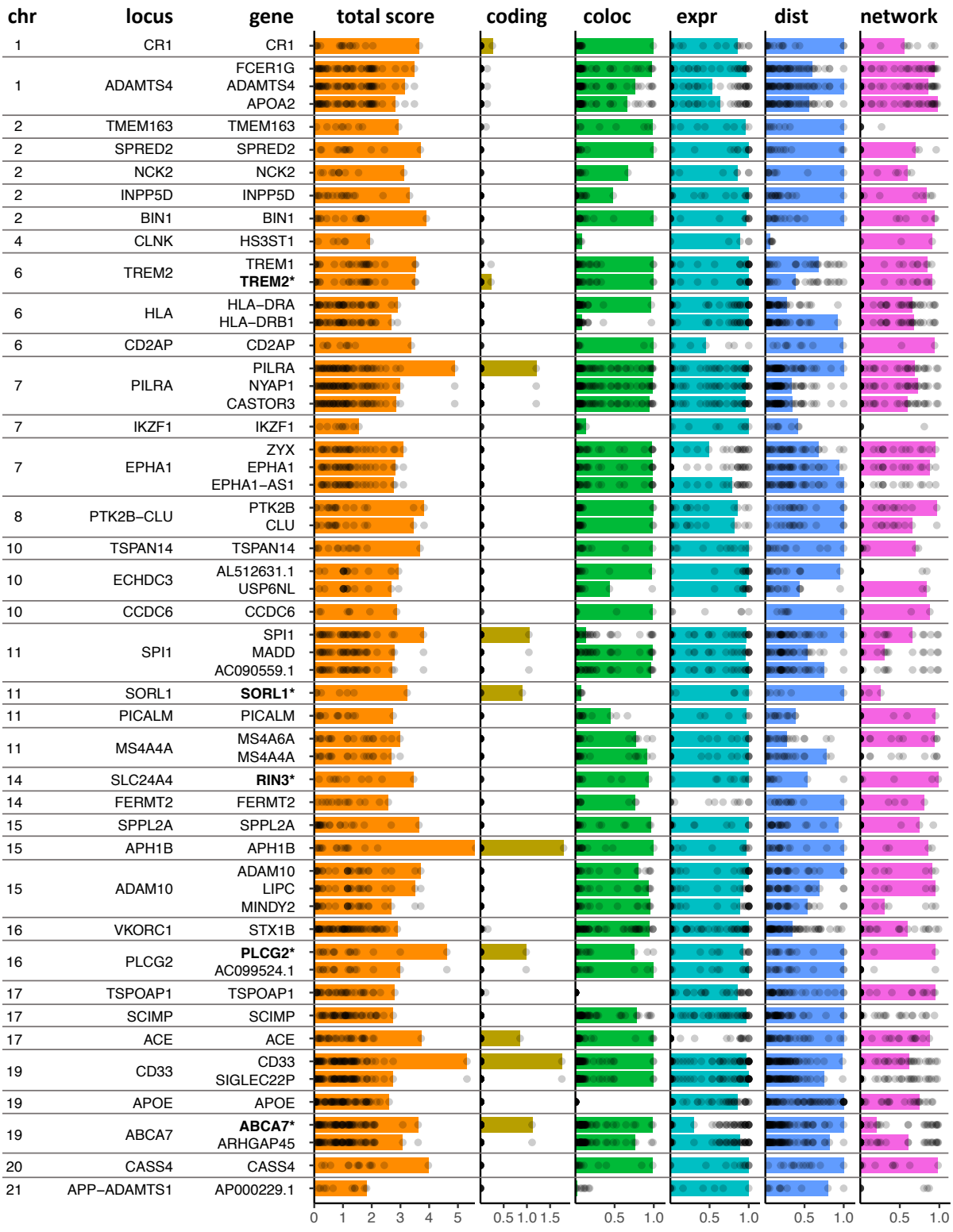


Figure 5: Gene evidence summary. The top gene at each locus is shown, as well as the next 16 top genes by total score. Score components for each gene are indicated by coloured bars, and points show the distribution of scores for all genes within 500 kb at the locus. Bold gene names are those with evidence of causality based on rare variants from other studies.

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

While our prioritisation further supports many established AD candidate genes, it also implicates novel genes. Among these are *FCER1G* at the *ADAMTS4* locus, a gene which negatively regulates immune cell responses[79] and has been reported as a hub gene in microglial gene modules associated with neurodegeneration[80,81]. Another candidate is *ZYX* at the *EPHA1* locus, which receives a top network score, is highly expressed in microglia, and which was recently nominated as an AD risk gene based on chromatin interactions between the *ZYX* promoter and AD risk variants in a *ZYX* enhancer[82]. Finally, at the MS4A locus our prioritisation nominates *MS4A6A*, another gene implicated in AD risk based on chromatin interactions[82], despite this gene being relatively distal (143 kb) from the AD association peak.

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: association peaks span large genomic regions, and non-coding associations can act via regulation of distal genes. We approached this challenge for AD by performing the largest meta-analysis to date, followed by comprehensive fine-mapping, eQTL colocalisation, and quantitative gene prioritization.

Our meta-analysis identified four novel associations with genome-wide significance, near *NCK2*, *SPRED2*, *TSPAN14*, and *CCDC6*. Each of these genes was supported by both eQTL colocalisation and network ranking, and in each case was the nearest gene to the association peak. Indeed, when distance was excluded from the priority score, for 21 of the 37 loci the top prioritized gene was the nearest gene, and for a further 8 loci the top gene was within 100 kb. This is consistent with observations from eQTL studies that the majority of gene regulatory variants lie within 100 kb of their regulated genes[83].

Despite the large number of eQTL datasets that we used, colocalisation of likely AD risk genes was sometimes found in only one or a few datasets; this was the case for *SPRED2* (TwinsUK LCL coloc probability 0.99), *RIN3* (GTEx frontal cortex probability 0.94), and *PILRA* (Fairfax LPS-2hr monocyte coloc probability 0.99). Many factors could account for dataset-specific colocalisations, 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 colocalisation provides only weak evidence for lack of an effect in a given tissue type, whereas positive

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

colocalisation provides strong support for a shared genetic effect. It is therefore useful to look broadly across eQTL studies for colocalisation, which will be facilitated by resources that simplify access to these datasets, such as the eQTL catalogue[11].

Our gene prioritization incorporated multiple lines of genetic evidence to distinguish genes most likely to causally mediate AD risk. This analysis supported roles for many established AD genes, while also pointing to novel candidates. One of our most confidently prioritized genes was *APH1B*, encoding a gamma-secretase complex component involved in APP processing. *APH1B* harbors the likely causal missense variant rs117618017, yet also has strong colocalisation 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[84] and humans[85] that loss of APH1B and gamma-secretase function leads to AD. Although APH1B loss may be associated with non-AD dementia[84,86], effect sizes for rs117618017 on risk were similar both for diagnosed AD (Kunkle et al. OR 95% CI=[1.071, 1.127]) and dementia in UK Biobank (OR 95% CI=[1.064, 1.104]).

Among our novel associations, *TSPAN14* has a role in defining the localisation of *ADAM10*[87], another recently discovered AD gene which is a key component of the gamma-secretase complex, and which 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[88]. Our fine-mapping showed that the risk SNP rs1870138 is also associated with higher risk for inflammatory bowel disease (IBD), an immune-mediated disease, and with higher monocyte count in UK Biobank individuals. 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. *SPRED2* is a negative regulator of ERK/MAPK signalling, and its loss in mice leads to increased macrophage activation and tissue inflammation[89]. The AD-associated SNPs in *SPRED2*, rs268134 and rs268120, are also associated with increased neutrophil percentage and decreased lymphocyte percentage in the UK Biobank.

Recently proposed AD candidate genes supported by our analyses include *RIN3*, *HS3ST1*, and *FCER1G*. As noted above, *FCER1G* is a negative regulator of immune cell responses[79]; *RIN3* interacts with both *BIN1* and *CD2AP* in the early endocytic pathway[90]; *HS3ST1* is involved in cellular uptake of tau[91] and was recently been associated with AD in an independent Norwegian sample[61].

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

In summary, our study reports fine-mapping SNP probabilities for 36 AD-associated regions, including 4 novel loci. By combining evidence from eQTL colocalisations across 109 datasets, functional annotations and gene network analysis, we generate a quantitative prioritization score and provide a comprehensive map of AD candidate genes. Our genetic findings highlight the presence of diverse mechanisms in AD pathogenesis, suggesting different possible entry points for interventions to treat AD or to reduce risk of the disease, and identify candidate targets for therapeutic development.

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Methods

Code for analyses described here can be found at github.com/jeremy37/AD_finemap.

GWAS on family history of AD

Sample QC, variant QC and imputation was performed on all UK Biobank participants as described in Bycroft et al.[92]. After genotype imputation, 93,095,623 variants across 487,409 individuals were available for analysis. To exclude individuals of non-European ancestry, we first extracted the "White British" ancestry subset of participants as described in Bycroft et al. 2018. These individuals self-reported their ethnic background as "British" have similar genetic ancestry based on principal components (PC) analysis. To extract additional individuals of European ancestry, we followed a similar approach to Bycroft et al. and applied Aberrant[93] on PCs 1v2, 3v4 and 5v6 across the individuals who self-reported as "Irish" or "Any other white background". Pairs of first-degree relatives were identified using KING v2.0[94]. We applied KING to 147,522 UK Biobank individuals who had at least one relative identified in Bycroft et al. (UK Biobank 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 < 0.3) and/or those with minor allele frequencies below 0.0005, resulting in 25,647,815 variants available for analysis.

AD cases were extracted from UK Biobank self-report (field 20002), ICD10 diagnoses (fields 41202 and 41204) and ICD10 cause of death (fields 40001 and 40002) data. UK Biobank participants were asked whether they have a biological father, mother or sibling who suffered from Alzheimer's disease/dementia (UK Biobank 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 analyses. All remaining individuals were denoted as controls.

There were 3,046 AD cases, 52,791 AD proxy cases and 355,900 controls in the combined white British and white non-British cohorts. For 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[95].

AD meta-analysis

To enable meta-analysis combining the UK Biobank cohorts with external case-control studies, we first transformed the AD proxy BOLT-LMM summary statistics from the linear scale to a 1/0 log odds ratio:

$$logOR \approx \beta_{LMM} / (f(1-f))$$

with standard error:

$$se \approx se_{LMM} / (f(1-f))$$

where β_{LMM} and se_{LMM} are the SNP effect sizes and standard errors respectively from BOLT-LMM, and f is the fraction of cases in the sample[96]. Since the affected individuals in our analysis include both true and proxy-cases, we then multiplied the transformed logORs and

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

standard errors by 1.897 so that it approximates the logORs obtained from a true case/control study[1].

We combined the transformed UK Biobank white British cohort, the transformed UK Biobank white non-British cohort and the Stage 1 summary statistics from Kunkle et al. using a fixed-effects (inverse variance weighted) meta-analysis across 10,687,126 overlapping variants. For display purposes (Supplementary Table 7), we used CrossMap[97] to convert variant positions from GRCh37 to GRCh38.

Conditional analysis and statistical fine-mapping

To run GCTA, we prepared plink input files with genotypes from 10,000 randomly sampled UK Biobank individuals at variants within +/- 5 Mb from each lead SNP. We excluded variants with INFO < 0.85, or which had a p-value from Cochran's Q test for study heterogeneity < 0.001. We also excluded variants with allele frequency in UK Biobank below 0.1%, as LD estimates are unreliable at low allele counts. We ran GCTA --cojo-slct with a p-value threshold of 10^{-5} to identify secondary signals at each locus, and then retained only loci with a lead p-value below $5x10^{-8}$. For the HLA locus we used a GCTA p-value threshold of $5x10^{-8}$. We also retained the loci *TSPOAP1*, *IKZF1*, and *TMEM163* since they had p < $5x10^{-8}$ 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 at each locus, with --n-causal-snps given as the number of independent SNPs determined by GCTA. For FINEMAP, we excluded variants with allele frequency below 0.2%, since we found that otherwise FINEMAP sometimes selected implausible causal variants, such as pairs of very weakly associated rare variants to explain a common variant signal. 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 compute SNP causal probabilities based on GCTA conditional signals, we converted effect size (beta) and standard error values to approximate Bayes Factors (BF)[98] using a prior of W=0.1 (in Wakefield notation), and used the WTCCC single-causal variant method[48], probability = SNP BF / sum(all SNP BFs).

Colocalisation with eQTLs

For eQTL colocalisation, we downloaded summary statistics for the eQTL datasets mentioned in the main text, as well as the xQTL dataset[19] based on dorsolateral prefrontal cortex brain samples. QTL calling for primary microglia was performed with RASQUAL[99] with the --no-posterior-update option. We determined eQTL genes at FDR 5% for each dataset in a uniform manner, first using Bonferroni correction of lead SNP nominal p values based on the number of variants within 500 kb of the gene, and using the Benjamini-Hochberg method to compute FDR. We matched variants between eQTL and GWAS based on chromosomal position; for datasets in GRCh38 coordinates, we first used CrossMap[97] to convert back to GRCh37 coordinates. We used the coloc package[17] with default priors to perform

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

colocalisation tests between GWAS and eQTL signals where the lead variants were within 500 kb of each other, and passed to coloc all variants within 200 kb of each lead variant. We also performed colocalisation tests using GWAS p-values for each conditionally independent GWAS signal, obtained with GCTA as described above.

Functional annotations

All functional annotations used were in GRCh37 coordinates, as was the AD meta-analysis. We used the Ensembl VEP online Web tool (www.ensembl.org/vep)[54] to predict variant consequences, and to add selected annotations (Supplementary Table 6). We downloaded bed files based on imputed data for Roadmap Epigenomics DNase, histone peaks, 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 & 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). For genome segmentations, we considered 9 states to represent enhancers: TxReg, TxEnh5, TxEnh3, TxEnhW, EnhA1, EnhA2, EnhAF, EnhW1, EnhW2. We used bedtools[100] to determine overlaps, and counted the number of overlaps for each variant with peaks in the above groups. We downloaded FANTOM5[101] permissive enhancer annotations from fantom.gsc.riken.jp/5/data. We downloaded pre-computed SpliceAI scores[57] for variants within genes from github.com/Illumina/SpliceAI. We merged filtered whole-genome and exome scores together to obtain the most comprehensive predictions, and for each AD variant we annotated the maximum predicted score across splice donor gain, donor loss, acceptor gain, acceptor loss. We used the DeepSEA[56] online tool (deepsea.princeton.edu) to annotate variants selected for functional fine-mapping with DeepSEA's "functional significance" score. BigWig files with PhastCons, PhyloP and GERP RS scores in hg19 (GRCh37) coordinates were downloaded from UCSC. We downloaded microglial ATAC-seq based on the study by Gosselin et al.[51], aligned reads to GRCh37 with bwa 0.7.15[102], and called multisample peaks across all 15 datasets using MACS2[103]. We prepared bigWig files from alignments by using bedtools genomecov, followed by bedGraphToBigWig. To visualise microglia ATAC-seq tracks we adapted code from wiggleplotr[104].

Annotation-based fine-mapping

For fine-mapping with PAINTOR, we selected 3,207 variants which had (a) FINEMAP probability >= 0.01% based on the GCTA-identified number of causal variants at the locus, or (b) had FINEMAP probability >= 1% when run with either 1 or 2 causal variants, even if this was not the number identified by GCTA, or (c) 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, which included thresholding certain scores at multiple levels (e.g. CADD >= 5, 10, 20). For Roadmap DNase and enhancer annotations, we included a category based on whether a variant was in a peak or enhancer in >= 10 epigenomes. We next ran PAINTOR v3.1 once for each of the 43 annotations (Figure 2; Supplementary Table 6), allowing 2 causal variants per locus. (In addition to excluding *APOE*, we excluded the *CLNK* locus because PAINTOR failed to run when this locus was input.)

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

To build a multi-annotation model, we performed forward stepwise selection. We selected the best annotation by model log-likelihood (LLK), Blood & immune DNase, and then ran PAINTOR again for each combination of this annotation and the 42 remaining annotations. We continued adding an annotation at each iteration from among those top-ranked by model LLK until the model LLK improvement was less than 1. This occurred at iteration 4, and so we kept the first 3 annotations in the combined model. We computed the mean causal probability for each SNP as the mean of the 3 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 (>1000 interactions) in the IntAct[105] and BioGRID databases[106], as well as edges from the STRING database version 10.5[46] with edge score > 0.75. This combined network included 18,055 genes and 540,421 edges. We identified 36 top candidate genes across AD loci (Supplementary Table 8) to use as seed genes, and assigned weight to these according to the -log10(p value) of the lead SNP at the locus. 33 of the candidate genes were present in the network, while 3 were not (*ECHDC3*, *TMEM163*, *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[107], included in the igraph R package[108]. We found that a gene's resulting PageRank was highly correlated with its node degree, and this made PageRank itself less informative. We therefore compared the PageRank of each gene at the locus 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. To determine gene set enrichment, we used the top 1,000 genes by network rank as input to gProfiler[109] with default settings, with the set of all genes ranked by the network as a background set.

Gene expression

Gene expression values for all tissues were determined in units of transcripts per million (TPM). Both GTEx v8 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 1.5.3 as described in[14], from which we computed median TPM.

Gene prioritization

The **combined score** for each gene is the sum of five scores:

geneScore = codingScore + exprScore + distScore + colocScore + networkScore

The **coding score** is twice the sum of the mean fine-mapping probability for missense or LoF variants in a gene.

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

The **expr(ession) score** is determined by first computing the percentile of expression (measured in TPM) for a gene in microglia relative to all GTEx tissues, and for GTEx brain (mean of brain tissues) relative to all other GTEx tissues. The score is then normalised to be in the range [0, 1], and rewards genes with expression percentile above the 50th:

exprScore = max(microglia pctile - 50, gtex pctile - 50) / 50

The **dist(ance) score** is defined as:

distScore = (log10(maxDist) - log10((abs(*x*)+distBias))) / (log10(maxDist) - log10(distBias))

where *x* is the minimum distance of any portion of the gene footprint to the region defined by independent lead SNPs at a GWAS locus, maxDist is 500,000 and distBias is 6,355, chosen to give reasonable scores over the main range of interest of 0 - 200 kb (Supplementary Figure 3).

The **coloc score** used as part of the combined score is defined as the maximum value of the "H4" hypothesis probability output by the coloc R package. The weighted coloc score this was compared to was designed to accumulate evidence across datasets, prioritizing those most relevant, and is inspired by the Open Targets[47] and StringDB[46] evidence scoring systems. Each QTL dataset was assigned to one of the categories "relevant", "possibly relevant", and "not relevant", which receive weights of 1.0, 0.8, and 0.5. Within each category, the score accumulated by ordering coloc H4 probabilities in descending order, and adding incrementally:

$categoryScore_1 = H4_1$

$categoryScore_2 = categoryScore_1 + (1 - categoryScore_1) * H4_2 / 2$

…

$categoryScore_n = categoryScore_{n-1} + (1 - categoryScore_{n-1}) * H4_n / n$

Within a category, then, a small number of strong colocs will receive a higher score than many weak colocs. The score was then accumulated across categories as:

$colocScore_1 = 1.0 * categoryScore_{relevant} +$

$colocScore_2 = colocScore_1 + 0.8*(1 - colocScore_1) * categoryScore_{possbly\_relevant}$

$colocScore_{final} = colocScore_2 + 0.5*(1 - colocScore_2) * categoryScore_{not\_relevant}$

For example, if there were two "relevant" coloc H4 values of 0.4, and a "not relevant" H4 value of 0.9, the colocScore would be:

1.0 * (0.4 + (1 - 0.4) * 0.4 / 2) + 0.5 * (1 - (0.4 + (1 - 0.4) * 0.4 / 2)) * 0.9 = 0.736

In contrast, with one "relevant" H4 value of 0.9 and two "not relevant" H4 values of 0.4, the colocScore would be:

1.0 * 0.9 + 0.5 * (1 - 0.9) * (0.4 + (1 - 0.4) * 0.4 / 2) = 0.926

The **network score** is determined based on the page rank percentile for a gene relative to permutations:

networkScore = (page_rank_pctile - 50) / 50

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Data availability

Summary statistics from the UK Biobank GWAX for AD and from the meta-analysis will be made available through the NHGRI-EBI GWAS Catalog: www.ebi.ac.uk/gwas/downloads/summary-statistics

Acknowledgements

This work was funded by Open Targets (OTAR037). We thank Jeff Barrett for guidance during initiation of the project; Adam Young for sharing early access to human primary microglia data; and Kaur Alasoo for early access to the eQTL catalogue.

Author contributions

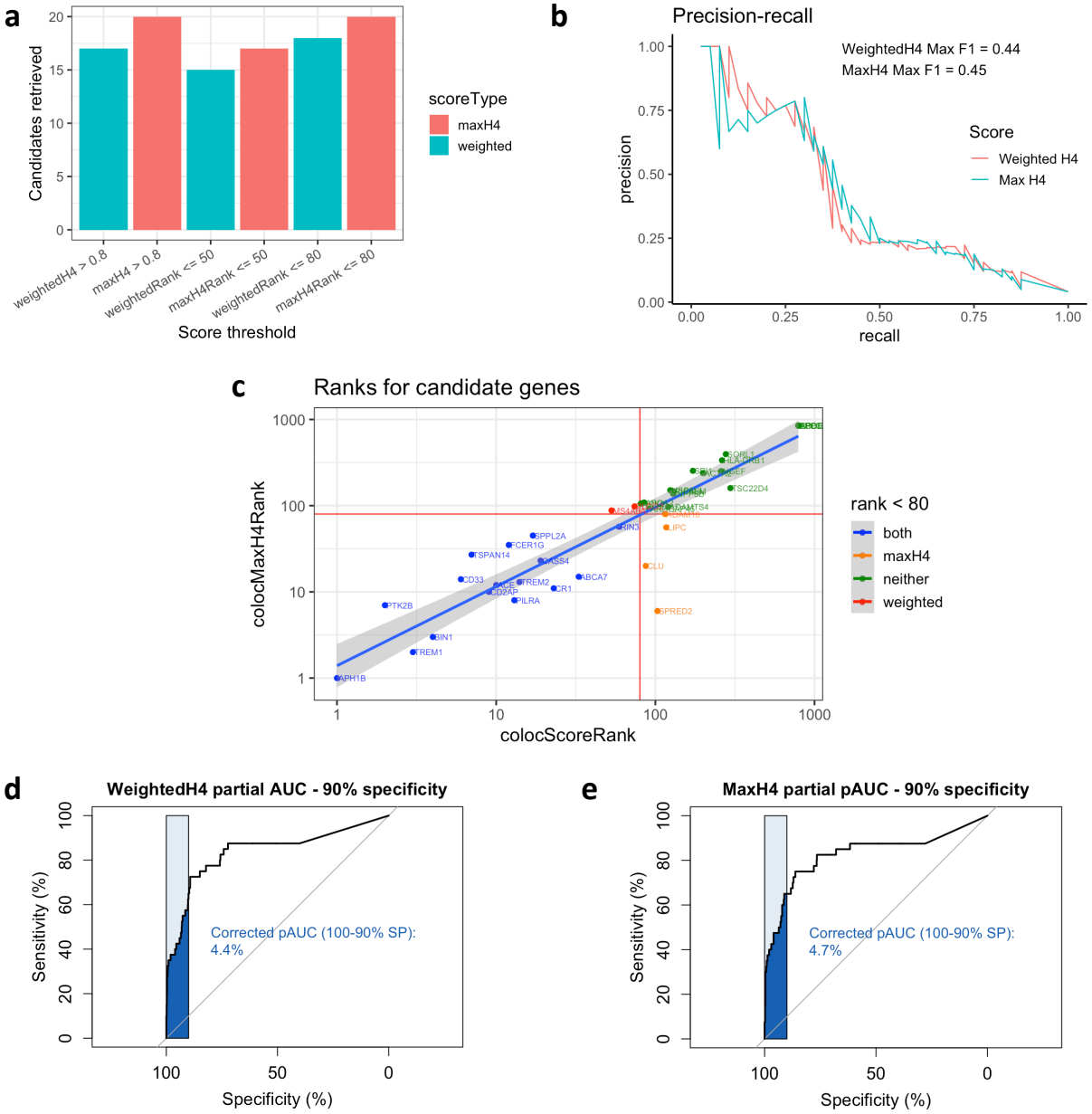
JS planned and conducted the analyses, and wrote the paper. JL performed the GWAX and meta-analysis. SC and EB assisted with fine-mapping, variant and gene prioritization. IB and PB performed and supervised gene network analysis. NK performed microglia eQTL mapping. AB, TJ, DJG, and KE conceived and supervised the study.

Competing interests

JL is an employee of Biogen. DJG is an employee of Genomics Plc. TJ is an employee of GSK. KE is an employee of BioMarin Pharmaceutical.

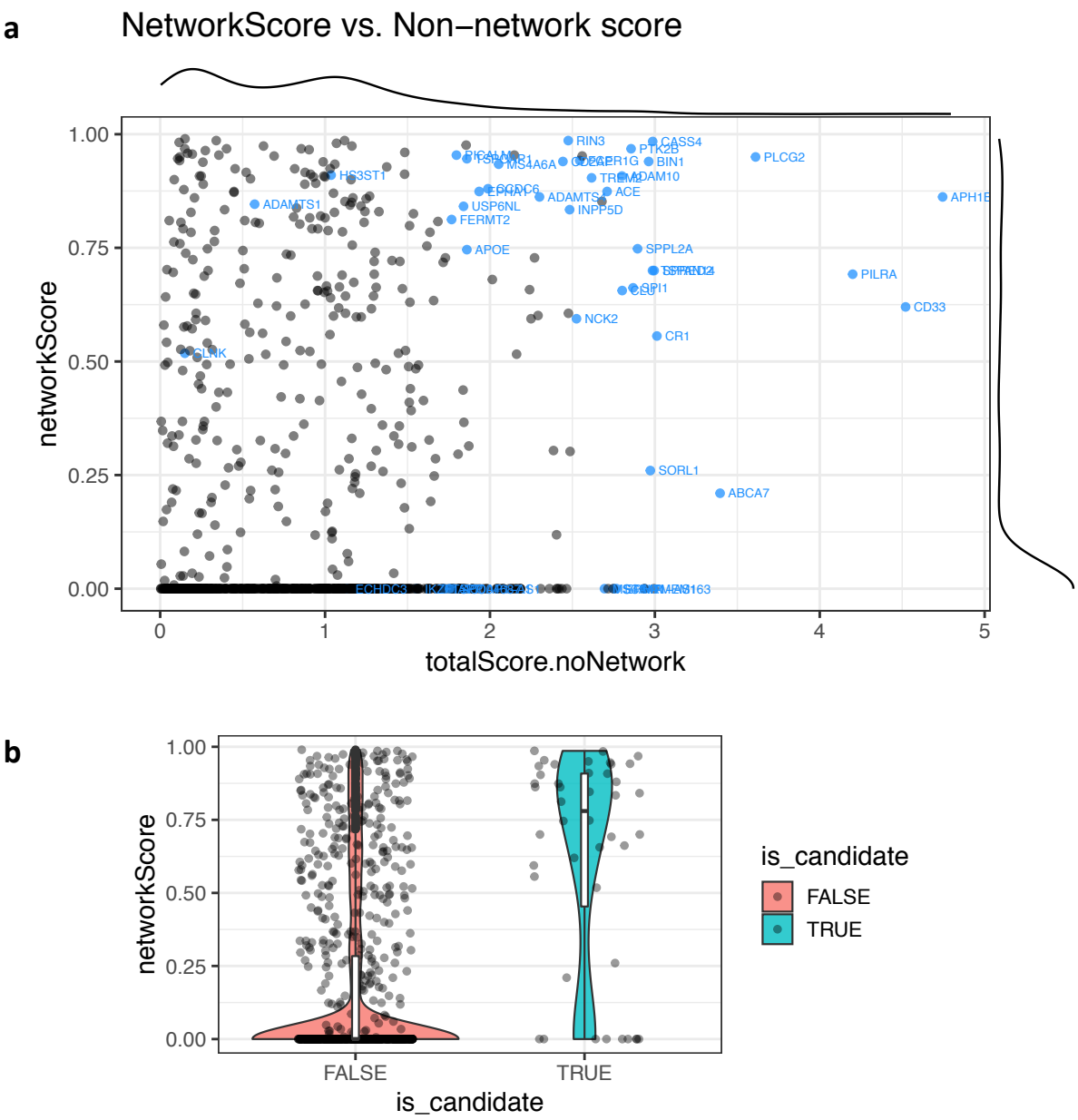
medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Supplementary Figures



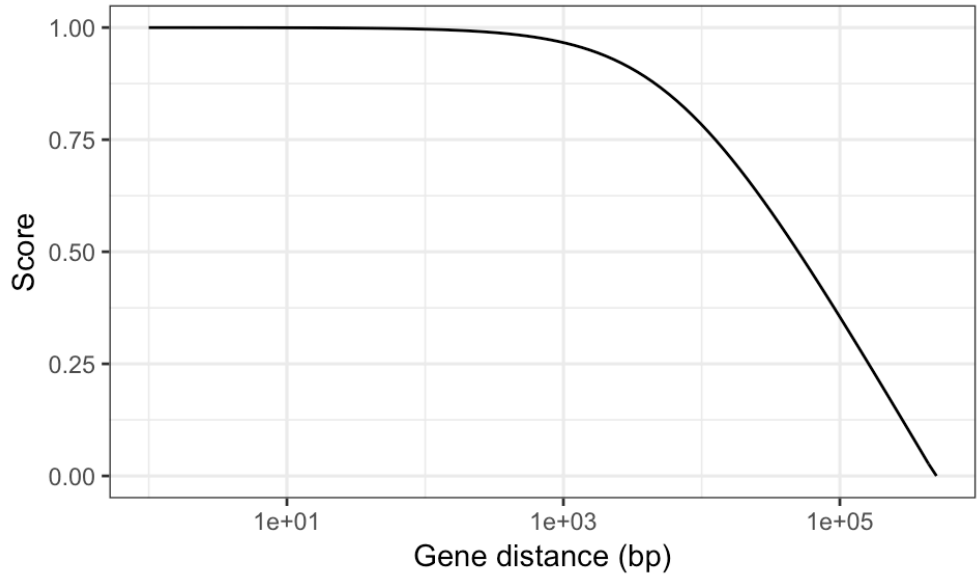
Supplementary Figure 1: Maximum coloc probability ("maxH4") outperformed a weighted coloc score that prioritises relevant cell types. We compared the performance of each score in identifying the top 40 AD candidate genes based on total score without coloc. (a) Barplot showing the number of candidate genes retrieved by either maxH4 or weighted coloc score at three score thresholds. In each case maxH4 retrieved more candidate genes. (b) Precision-recall curves, showing comparable F1 score between the two methods. (c) Scatter plot of gene ranks for top 40 genes relative to all genes, for weighted or maxH4 score. (d,e) Receiver operator characteristic partial area under the curve (AUC) at 90% specificity for (d) weighted coloc score and (e) maxH4 score. The maxH4 score showed slightly higher sensitivity at high specificity.

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .



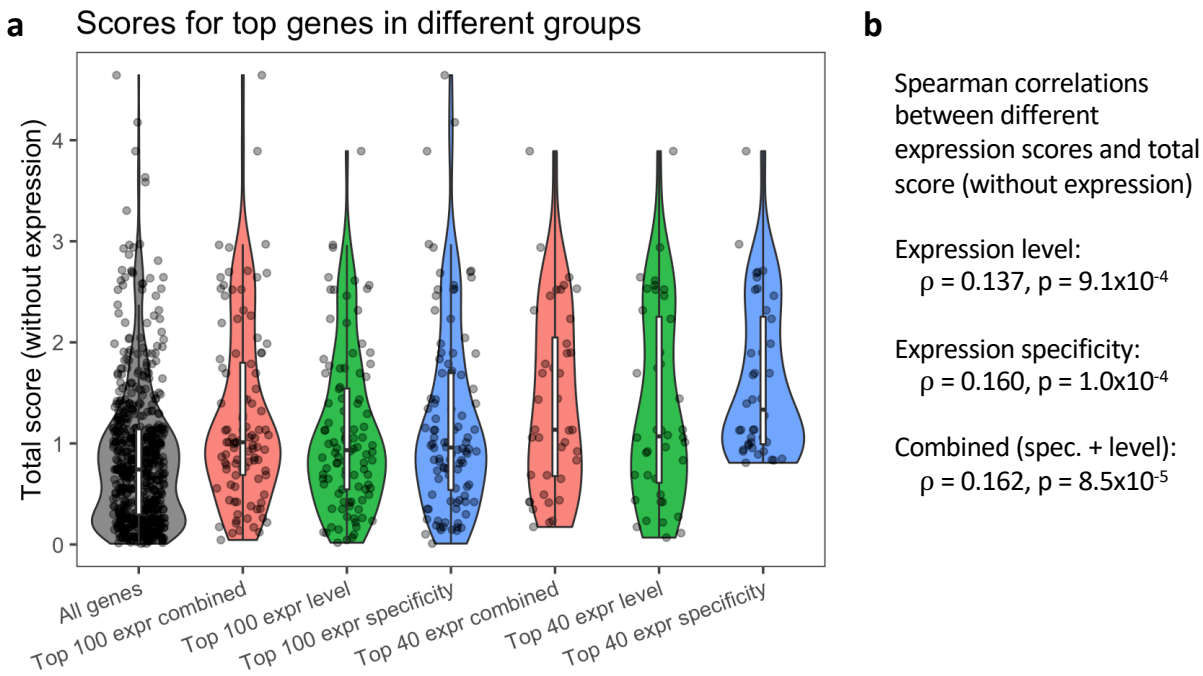
Supplementary Figure 2: Network scores are higher for AD candidate genes.
(a) Scatterplot of network score vs. the total score excluding network information (coloc + fine-mapping + expression + distance) for genes at 36 AD loci included in the main text. (b) Violin plot showing the distribution of scores for AD candidate genes (highlighted in part a) and all others. Note that genes not present in the network (*TMEM163*, *ECHDC3*, *SCIMP*) receive a score of zero by default.

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .



Supplementary Figure 3: The gene distance score decreases with increasing log-scaled distance to the lead SNP.

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .



Supplementary Figure 4: (a) Violin plots showing the distribution of the total gene score, excluding the expression component, in different groups of genes defined by variants of the expression score. Only protein-coding genes were included. For each of the expression scores, the top 100 or top 40 genes were selected. The expression scores compared were: expression level (log10(TPM) / 3); specificity of gene expression in brain or microglia relative to all GTEx tissues (described in methods); combined expression score (0.5 * exprLevelScore + 0.5 * specificityScore). (b) The score based on expression specificity in microglia or brain had slightly higher correlation with the total (non-expression) score than did the expression level score. A score combining both components showed similar correlation to that based on expression specificity alone.

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

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. Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. doi:10.1101/787903.
11. Kerimov, N. *et al.* The eQTL catalogue. *in preparation* (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. *bioRxiv* (2019) doi:10.1101/2019.12.20.874099.
15. Leung, Y. Y. *et al.* Identifying amyloid pathology–related cerebrospinal fluid biomarkers for Alzheimer's disease in a multicohort study. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* vol. 1 339–348 (2015).
16. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–75, S1–3 (2012).
17. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
18. Sieberts, S. K. *et al.* Large eQTL meta-analysis reveals differing patterns between cerebral cortical and cerebellar brain regions. doi:10.1101/638544.
19. 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).
20. Schmiedel, B. J. *et al.* Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell* **175**, 1701–1715.e16 (2018).
21. Jaffe, A. E. *et al.* Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat. Neurosci.* **21**, 1117–1125 (2018).
22. Buil, A. *et al.* Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat. Genet.* **47**, 88–91 (2015).
23. Fairfax, B. P. *et al.* Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles. *Nature Genetics* vol. 44 502–510 (2012).
24. Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).
25. Naranbhai, V. *et al.* Genomic modulators of gene expression in human neutrophils. *Nat. Commun.* **6**, 7545 (2015).
26. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398–1414.e24 (2016).
27. 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).
28. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife* **2**, e00523 (2013).
29. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

30. Kilpinen, H. *et al.* Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**, 370–375 (2017).
31. Nédélec, Y. *et al.* Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. *Cell* **167**, 657–669.e21 (2016).
32. Quach, H. *et al.* Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. *Cell* **167**, 643–656.e17 (2016).
33. Schwartzentruber, J. *et al.* Molecular and functional variation in iPSC-derived sensory neurons. *Nat. Genet.* **50**, 54–61 (2018).
34. 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).
35. Momozawa, Y. *et al.* IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nat. Commun.* **9**, 2427 (2018).
36. 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).
37. Salazar, S. V. *et al.* Alzheimer's Disease Risk Factor Pyk2 Mediates Amyloid-β-Induced Synaptic Dysfunction and Loss. *The Journal of Neuroscience* vol. 39 758–772 (2019).
38. 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).
39. Calafate, S., Flavin, W., Verstreken, P. & Moechars, D. Loss of Bin1 Promotes the Propagation of Tau Pathology. *Cell Rep.* **17**, 931–940 (2016).
40. Nott, A. *et al.* Cell type-specific enhancer-promoter connectivity maps in the human brain and disease risk association. doi:10.1101/778183.
41. Rathore, N. *et al.* Paired Immunoglobulin-like Type 2 Receptor Alpha G78R variant alters ligand binding and confers protection to Alzheimer's disease. *PLOS Genetics* vol. 14 e1007427 (2018).
42. Chan, G. *et al.* CD33 modulates TREM2: convergence of Alzheimer loci. *Nat. Neurosci.* **18**, 1556–1558 (2015).
43. Raj, T. *et al.* CD33: increased inclusion of exon 2 implicates the Ig V-set domain in Alzheimer's disease susceptibility. *Human Molecular Genetics* vol. 23 2729–2736 (2014).
44. Jonsson, T. *et al.* Variant of TREM2 associated with the risk of Alzheimer's disease. *N. Engl. J. Med.* **368**, 107–116 (2013).
45. 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).
46. 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).
47. Koscielny, G. *et al.* Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res.* **45**, D985–D994 (2017).
48. Wellcome Trust Case Control Consortium *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. Gosselin, D. *et al.* An environment-dependent transcriptional network specifies human microglia identity. *Science* **356**, (2017).
52. Alasoo, K. *et al.* Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nature Genetics* vol. 50 424–431 (2018).
53. Roadmap Epigenomics Consortium *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

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

disease. *Acta Neuropathol.* **135**, 827–837 (2018).
60. 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).
61. 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).
62. Jun, G. R. *et al.* Transethnic genome-wide scan identifies novel Alzheimer's disease loci. *Alzheimers. Dement.* **13**, 727–738 (2017).
63. 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).
64. Sassi, C. *et al.* Influence of Coding Variability in APP-Aβ Metabolism Genes in Sporadic Alzheimer's Disease. *PLOS ONE* vol. 11 e0150079 (2016).
65. 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 Genetics* vol. 13 e1006933 (2017).
66. Ghanbari, M. *et al.* A functional variant in the miR-142 promoter modulating its expression and conferring risk of Alzheimer disease. *Human Mutation* vol. 40 2131–2145 (2019).
67. 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).
68. 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).
69. Kamboh, M. I. *et al.* Genome-wide association study of Alzheimer's disease. *Transl. Psychiatry* **2**, e117 (2012).
70. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nat. Genet.* **50**, 1593–1599 (2018).
71. 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).
72. 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 Medicine* vol. 14 e1002270 (2017).
73. Fang, H. *et al.* A genetics-led approach defines the drug target landscape of 30 immune-related traits. *Nat. Genet.* **51**, 1082–1091 (2019).
74. Amin, L. & Harris, D. A. Aβ receptors specifically recognize molecular features displayed by fibril ends and neurotoxic oligomers. doi:10.1101/822361.
75. 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. *Alzheimer's & Dementia* vol. 11 1430–1438 (2015).
76. Chang, D. *et al.* A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat. Genet.* **49**, 1511–1516 (2017).
77. Barbeira, A. N. *et al.* Widespread dose-dependent effects of RNA expression and splicing on complex diseases and traits. doi:10.1101/814350.
78. 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).
79. Sweet, R. A., Nickerson, K. M., Cullen, J. L., Wang, Y. & Shlomchik, M. J. B Cell–Extrinsic Myd88 and Fcer1g Negatively Regulate Autoreactive and Normal B Cell Immune Responses. *The Journal of Immunology* vol. 199 885–893 (2017).
80. 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).
81. Zhang, B. *et al.* Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153**, 707–720 (2013).
82. Novikova, G. *et al.* Integration of Alzheimer's disease genetics and myeloid genomics reveals novel disease risk mechanisms. doi:10.1101/694281.
83. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14–24 (2014).
84. 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).
85. 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).
86. Acx, H. *et al.* Inactivation of γ-secretases leads to accumulation of substrates and non-Alzheimer

medRxiv preprint doi: https://doi.org/10.1101/2020.01.22.20018424; this version posted January 27, 2020. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

neurodegeneration. *EMBO Mol. Med.* **9**, 1088–1099 (2017).
87. Matthews, A. L. *et al.* Regulation of Leukocytes by TspanC8 Tetraspanins and the 'Molecular Scissor' ADAM10. *Front. Immunol.* **9**, 1451 (2018).
88. Schlepckow, K. *et al.* An Alzheimer-associated TREM2 variant occurs at the ADAM cleavage site and affects shedding and phagocytic function. *EMBO Molecular Medicine* vol. 9 1356–1365 (2017).
89. Ohkura, T. *et al.* Spred2 Regulates High Fat Diet-Induced Adipose Tissue Inflammation, and Metabolic Abnormalities in Mice. *Front. Immunol.* **10**, 17 (2019).
90. 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).
91. Zhao, J. *et al.* Rare 3-O-sulfation of Heparan Sulfate Enhances Tau Interaction and Cellular Uptake. *Angew. Chem. Int. Ed Engl.* (2019) doi:10.1002/anie.201913029.
92. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
93. Bellenguez, C. *et al.* A robust clustering algorithm for identifying problematic samples in genome-wide association studies. *Bioinformatics* **28**, 134–135 (2012).
94. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
95. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
96. 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. *The Annals of Applied Statistics* vol. 7 369–390 (2013).
97. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).
98. Wakefield, J. Bayes factors for genome-wide association studies: comparison with P-values. *Genet. Epidemiol.* **33**, 79–86 (2009).
99. Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* **48**, 206–213 (2016).
100. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
101. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
102. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
103. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
104. Alasoo, K. wiggleplotr: Make read coverage plots from BigWig files. *R package version 1.10.1* (2019).
105. Orchard, S. *et al.* The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research* vol. 42 D358–D363 (2014).
106. Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* **45**, D369–D379 (2017).
107. Fogaras, D., Rácz, B., Csalogány, K. & Sarlós, T. Towards Scaling Fully Personalized PageRank: Algorithms, Lower Bounds, and Experiments. *Internet Mathematics* vol. 2 333–358 (2005).
108. Csardi, G., Nepusz, T. & Others. The igraph software package for complex network research. *InterJournal, Complex Systems* **1695**, 1–9 (2006).
109. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research* vol. 47 W191–W198 (2019).